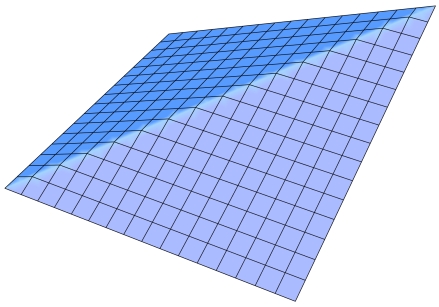


STAT 331: Applied Linear Models

Prof. Javid Ali • Spring 2014 • University of Waterloo
WF 1:00–2:20PM in MC 2065

Transcribed, LaTeXed, and edited by Rob Zimmerman

Note: These are personal lecture notes, not official course materials. They may contain errors (which by default you should attribute to me, Rob Zimmerman, rather than the course instructor). The permanent home of these — and all of my other lecture notes — is https://rob-zimmerman.github.io/coursenotes.

Contents

Lecture 1 — Wednesday, May 07, 2014

1. Simple Linear Regression

Basic Setup

This course is about linear models, and we begin with the simplest case: simple linear regression. The word "simple" here refers to the fact that we have only one explanatory variable (as opposed to multiple regression, which we will study later).

Given a set of observed data points $(x_1, y_1), \ldots, (x_n, y_n)$, our goal is to understand the relationship between x and y. In particular, we want to be able to predict y from x, and we want to quantify the strength of the relationship.

The **simple linear regression model** posits that the relationship between x and y is linear (at least approximately), but that there is random variation around this linear relationship. More precisely, the model is

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

where β_0 is the **intercept**, β_1 is the **slope**, and ε_i is the **error term**.

Throughout simple linear regression, the predictor values $x_1, \ldots, x_n$ are treated as fixed constants. We assume $n \geq 3$, $\sigma^2 > 0$, and at least two distinct predictor values, so

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 > 0.$$

These conditions ensure that the slope is identifiable and that the residual variance has positive degrees of freedom.

The term $\beta_0 + \beta_1 x_i$ is the **structural** (or deterministic) part of the model, while ε_i is the **random** part. We can write

$$\mu_i = \mathbb{E}[y_i] = \beta_0 + \beta_1 x_i,$$

so that $y_i = \mu_i + \varepsilon_i$. The equation $\mu = \beta_0 + \beta_1 x$ is called the **true regression line** or **population regression line**.

Figure 1 illustrates the model. The orange line is the true regression line $\mu = \beta_0 + \beta_1 x$. At a given value x_i, the expected response is $\mu_i = \beta_0 + \beta_1 x_i$, while the observed response is $y_i = \mu_i + \varepsilon_i$. Thus ε_i is the signed vertical deviation of (x_i, y_i) from the regression line.

The simple linear regression model assumes the following about the error terms $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$:

1. $\mathbb{E}[\varepsilon_i] = 0$ for all i.

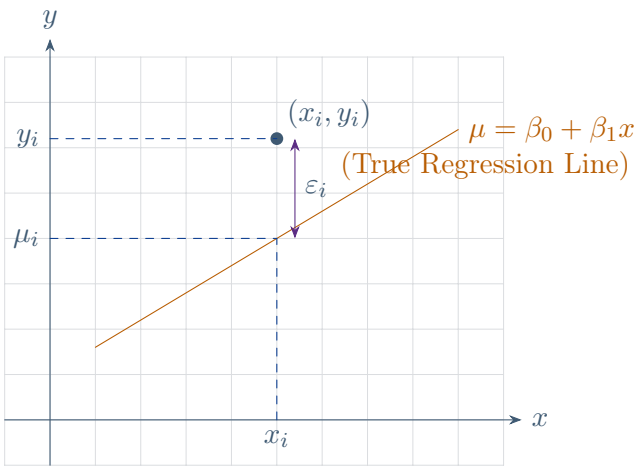


Figure 1

2. $\text{Var}(\varepsilon_i) = \sigma^2$ for all i (constant variance, also called **homoscedasticity**).
3. $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent.
4. The error vector $(\varepsilon_1, \ldots, \varepsilon_n)^\top$ is jointly normal.

Together, these assumptions say that the errors are independent and identically distributed as $\mathcal{N}(0, \sigma^2)$. We list the first three separately because they suffice for many least-squares results; joint normality is needed for the exact small-sample hypothesis tests and confidence intervals developed below.

Since $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, it follows from these assumptions that:

$$\mathbb{E}[y_i] = \beta_0 + \beta_1 x_i, \qquad \text{Var}(y_i) = \sigma^2, \qquad \text{and} \qquad y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2).$$

That is, for each fixed x_i, the response y_i is normally distributed with mean $\beta_0 + \beta_1 x_i$ and variance σ^2.

Figure 2 shows the distribution of y at three different values of x. Each distribution has the same variance σ^2 (i.e., the normal curves have the same spread), and the means lie on the regression line.

The assumption of constant variance is important. If the variance were to increase with x, for example, then the model would not be appropriate. We will discuss diagnostic tools for checking this assumption later in the course.

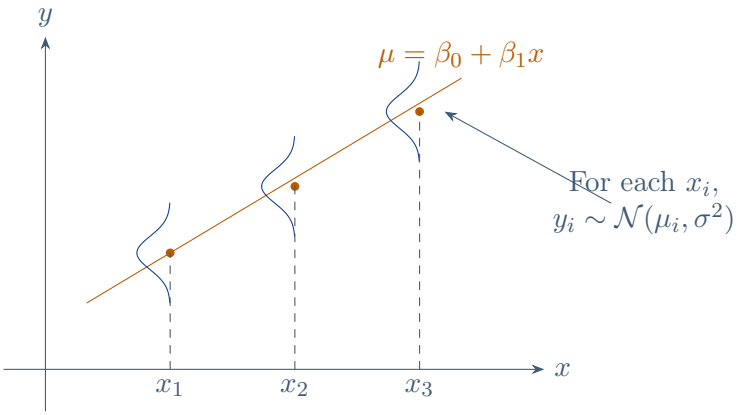


Figure 2

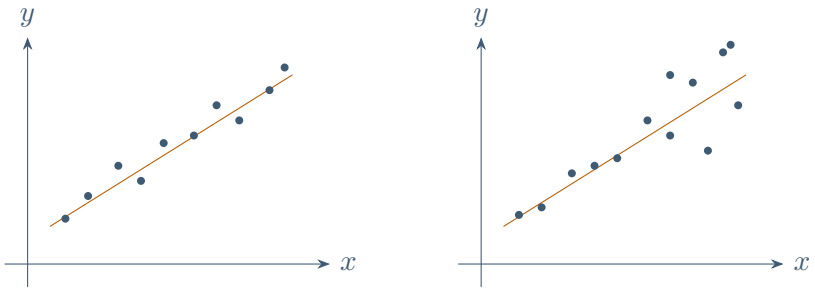


Figure 3

In Figure 3, the left panel shows data consistent with the constant variance assumption: the vertical spread of the points around the regression line is roughly the same for all x. In the right panel, that spread increases with x, violating Assumption (2).

In the next lecture, we will discuss how to estimate the parameters β_0, β_1, and σ^2 from the data.

Lecture 2 — Friday, May 09, 2014

Least Squares Estimation and Interpretation

Recall the simple linear regression model

$$y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{structural}} + \underbrace{\varepsilon_i}_{\text{random (error)}}, \quad i = 1, 2, \ldots, n.$$

We assume:

1. $\mathbb{E}[\varepsilon_i] = 0$,
2. $\text{Var}(\varepsilon_i) = \sigma^2$ (and hence $\text{Var}(y_i) = \sigma^2$),
3. $\varepsilon_1, \ldots, \varepsilon_n$ are independent, and
4. The error vector $(\varepsilon_1, \ldots, \varepsilon_n)^\top$ is jointly normal. Together with Assumptions (1)–(3), this gives independent $\mathcal{N}(0, \sigma^2)$ errors.

In particular, the second assumption states that the spread of the errors is constant across the range of x.

The parameters are β_0, β_1, σ. The intercept β_0 is the mean response when $x = 0$. The slope β_1 is the change in the mean response associated with a one-unit increase in x, since

$$\mathbb{E}[y \mid x = c] = \beta_0 + \beta_1 c \qquad \text{and} \qquad \mathbb{E}[y \mid x = c + 1] = \beta_0 + \beta_1 (c + 1).$$

To derive the least squares estimates of β_0 and β_1, we use the notation θ for a true unknown parameter and $\hat{\theta}$ for an estimator based on sample data. Then $\text{Var}(\theta) = 0$, while $\text{Var}(\hat{\theta})$ is the variance of the sampling distribution of $\hat{\theta}$, typically involving unknown model parameters.

Example 2.1 Let $\theta = \mu$ in the simple response model $y_1, \ldots, y_n \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \sigma^2)$. Then $\hat{\theta} = \bar{y}$ is the usual MLE for the mean, and

$$\text{Var}(\hat{\theta}) = \text{Var}(\bar{y}) = \frac{\sigma^2}{n}.$$

An estimator of this variance is $\widehat{\text{Var}}(\hat{\theta}) = s^2/n$, and the standard error is $\text{se}(\hat{\theta}) = \sqrt{\widehat{\text{Var}}(\hat{\theta})}$.

In simple linear regression, β_0 and β_1 are unknown and $\hat{\beta}_0$ and $\hat{\beta}_1$ are their estimates. Also,

$$\mu_i := \mathbb{E}[y_i] = \beta_0 + \beta_1 x_i$$

is the **true mean**,

$$\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

is the **fitted value** for y_i,

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$$

is the **true (unknown) error**, and

$$e_i := y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

is the **residual (estimated) error**.

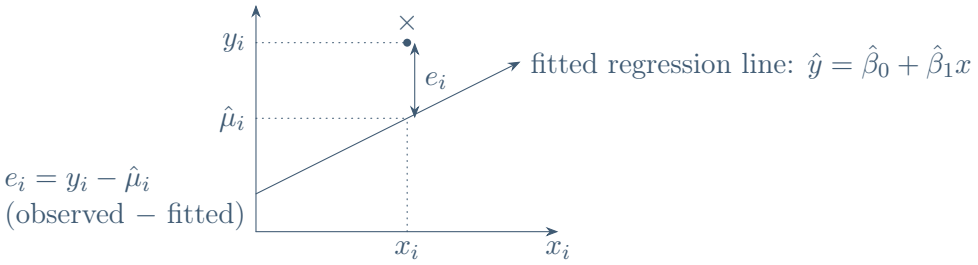


FIGURE 4

Figure 4 marks the residual e_i as the vertical difference between the observed response y_i and fitted value $\hat{\mu}_i$.

The least squares estimates $(\hat{\beta}_0, \hat{\beta}_1)$ minimize the sum of squared deviations

$$S(\beta_0, \beta_1) := \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Proposition 2.2 The least squares estimators satisfy the normal equations

$$\frac{\partial S}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \qquad \text{and} \qquad \frac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0.$$

Consequently,

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \qquad \text{and} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where

$$S_{xy} := \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \qquad \text{and} \qquad S_{xx} := \sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Proof. Setting the partial derivatives of $S(\beta_0, \beta_1)$ equal to zero gives the normal equations. Solving them yields

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \text{and} \qquad \hat{\beta}_1 = \frac{\displaystyle\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\displaystyle\sum_{i=1}^{n} x_i(x_i - \bar{x})}.$$

Also,

$$\sum_{i=1}^{n} x_i(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} y_i(x_i - \bar{x}) = S_{xy}$$

and

$$\sum_{i=1}^{n} x_i(x_i - \bar{x}) = \sum_{i=1}^{n}(x_i - \bar{x})^2 = S_{xx}$$

because

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

Hence $\hat{\beta}_1 = S_{xy}/S_{xx}$, and therefore $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. These stationary values are the unique minimizers: for any (β_0, β_1), the normal equations give the exact decomposition

$$S(\beta_0, \beta_1) = S(\hat{\beta}_0, \hat{\beta}_1) + n\big[(\beta_0 - \hat{\beta}_0) + \bar{x}(\beta_1 - \hat{\beta}_1)\big]^2 + S_{xx}(\beta_1 - \hat{\beta}_1)^2.$$

Because $n > 0$ and $S_{xx} > 0$, equality occurs only at $(\hat{\beta}_0, \hat{\beta}_1)$. □

Proposition 2.3 The least squares estimators satisfy

1. $\mathbb{E}\left[\hat{\beta}_0\right] = \beta_0$ and $\mathbb{E}\left[\hat{\beta}_1\right] = \beta_1$, so the estimators are unbiased.
2. $\text{Var}(\hat{\beta}_1) = \dfrac{\sigma^2}{S_{xx}}$.

Proof. For (1), we have

$$\begin{aligned}\mathbb{E}\left[\hat{\beta}_1\right] &= \mathbb{E}\left[\frac{S_{xy}}{S_{xx}}\right] = \frac{1}{S_{xx}}\mathbb{E}\left[S_{xy}\right] = \frac{1}{S_{xx}}\mathbb{E}\left[\sum_{i=1}^{n}(x_i-\bar{x})y_i\right] \\ &= \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\bar{x})\mathbb{E}\left[y_i\right] = \frac{1}{S_{xx}}\left[\beta_0\sum_{i=1}^{n}(x_i-\bar{x}) + \beta_1\sum_{i=1}^{n}(x_i-\bar{x})x_i\right] \\ &= \frac{1}{S_{xx}}\left[0+\beta_1 S_{xx}\right] = \beta_1.\end{aligned}$$

Also,

$$\mathbb{E}\left[\hat{\beta}_0\right] = \mathbb{E}\left[\bar{y}-\hat{\beta}_1\bar{x}\right] = \mathbb{E}\left[\bar{y}\right] - \bar{x}\mathbb{E}\left[\hat{\beta}_1\right] = \mathbb{E}\left[\beta_0+\beta_1\bar{x}+\bar{\varepsilon}\right] - \bar{x}\beta_1 = \beta_0.$$

For (2),

$$\begin{aligned}\text{Var}(\hat{\beta}_1) &= \text{Var}\left(\frac{S_{xy}}{S_{xx}}\right) = \text{Var}\left(\frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\bar{x})y_i\right) \\ &= \frac{1}{S_{xx}^2}\text{Var}\left(\sum_{i=1}^{n}(x_i-\bar{x})y_i\right) \overset{\text{indep}}{=} \frac{1}{S_{xx}^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\,\text{Var}(y_i) \\ &= \frac{\sigma^2}{S_{xx}^2}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{\sigma^2}{S_{xx}}.\end{aligned}$$

□

We immediately obtain the following corollary:

Corollary 2.4 For any fixed x, the fitted mean $\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x$ is unbiased for $\mu = \beta_0 + \beta_1 x$; that is, $\mathbb{E}\left[\hat{\mu}\right] = \mu$.

Example 2.5 The saved course data contain maintenance costs for 100 aircraft together with the corresponding number of flight hours. The following R code fits the simple linear regression model and constructs the fitted line and its pointwise confidence band.

```
airline <- read.delim("data/airline_slr.txt")
airline_fit <- lm(cost ~ hours, data = airline)

coef(summary(airline_fit))
confint(airline_fit)
```

The fitted regression equation is

$$\widehat{\text{cost}} = 868.49 + 6.4097\,\text{hours}.$$

Thus, an additional flight hour is associated with an estimated increase of 6.41 cost units in mean maintenance cost. The standard error of the slope is 0.4987, giving the 95% confidence interval $(5.4200, 7.3993)$. The residual standard error is 426.40, and $R^2 = 0.6277$, so flight hours explain about 62.8% of the observed variability in cost.

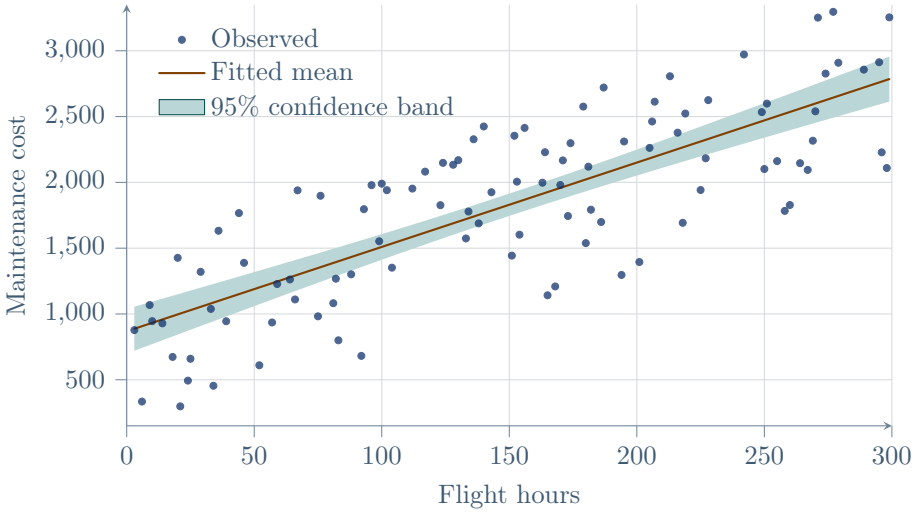


FIGURE 5

Figure 5 shows the airline maintenance data, fitted simple regression line, and pointwise 95% confidence band for the mean response.

Lecture 3 — Wednesday, May 14, 2014

Sampling Properties of Least Squares Estimators

Recall the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1, 2, \ldots, n$, with independent errors $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, and the least squares estimators $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ and $\hat{\beta}_1 = S_{xy}/S_{xx}$.

We also introduced some standard notation. θ denotes the true unknown population value, and $\hat{\theta}$ denotes its estimator from the sample. Also, $\text{Var}(\hat{\theta})$ is the theoretical sampling variance of $\hat{\theta}$, while $\widehat{\text{Var}}(\hat{\theta})$ is an estimator of $\text{Var}(\hat{\theta})$, typically obtained by replacing σ with $\hat{\sigma}$.

We found that the least squares estimators are unbiased (i.e., $\mathbb{E}\left[\hat{\beta}_0\right] = \beta_0$ and $\mathbb{E}\left[\hat{\beta}_1\right] = \beta_1$) and $\text{Var}(\hat{\beta}_1) = \sigma^2/S_{xx}$. To complete the story, we have the following:

Proposition 3.1

$$\text{Var}(\hat{\beta}_0) = \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]\sigma^2.$$

Proof. The estimator can be written as

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{\bar{x}\sum_{i=1}^{n}(x_i - \bar{x})y_i}{S_{xx}} = \sum_{i=1}^{n}\left[\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{S_{xx}}\right] y_i.$$

Because the observations are independent and $\text{Var}(y_i) = \sigma^2$, the variance of this linear combination

is

$$\begin{aligned}
\text{Var}(\hat{\beta}_0) &= \sum_{i=1}^{n} \left[\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}}\right]^2 \text{Var}(y_i) \\
&= \sigma^2 \sum_{i=1}^{n} \left[\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}}\right]^2 \\
&= \sigma^2 \sum_{i=1}^{n} \left(\frac{1}{n^2} - 2\frac{(x_i - \bar{x})\bar{x}}{nS_{xx}} + \frac{(x_i - \bar{x})^2\bar{x}^2}{S_{xx}^2}\right) \\
&= \sigma^2 \left[\sum_{i=1}^{n} \frac{1}{n^2} - 2\frac{\bar{x}}{nS_{xx}} \sum_{i=1}^{n}(x_i - \bar{x}) + \frac{\bar{x}^2}{S_{xx}^2} \sum_{i=1}^{n}(x_i - \bar{x})^2\right] \\
&= \sigma^2 \left[\frac{1}{n} - 2\frac{\bar{x}}{nS_{xx}} \cdot 0 + \frac{\bar{x}^2}{S_{xx}^2} \cdot S_{xx}\right] \\
&= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right).
\end{aligned}$$

Therefore,

$$\text{Var}(\hat{\beta}_0) = \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right] \sigma^2.$$

□

For simple regression, Theorem 10.1 also has the following intercept-specific form.

Theorem 3.2 (BLUE property of the intercept) Among all linear unbiased estimators of β_0, the least squares estimator $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$ has the smallest variance.

Proof. Write

$$\hat{\beta}_0 = \sum_{i=1}^{n} b_i y_i, \qquad \text{and} \qquad b_i = \frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}}.$$

Consider any other linear estimator $\tilde{\beta}_0 = \sum_{i=1}^{n} a_i y_i$. Since

$$\mathbb{E}[\tilde{\beta}_0] = \beta_0 \sum_{i=1}^{n} a_i + \beta_1 \sum_{i=1}^{n} a_i x_i,$$

the estimator is unbiased for β_0 only if

$$\sum_{i=1}^{n} a_i = 1 \qquad \text{and} \qquad \sum_{i=1}^{n} a_i x_i = 0.$$

Let $a_i = b_i + c_i$. The same two identities hold for the coefficients b_i, so

$$\sum_{i=1}^{n} c_i = 0 \qquad \text{and} \qquad \sum_{i=1}^{n} c_i x_i = 0.$$

Consequently,

$$\sum_{i=1}^{n} b_i c_i = \frac{1}{n}\sum_{i=1}^{n} c_i - \frac{\bar{x}}{S_{xx}}\sum_{i=1}^{n}(x_i - \bar{x})c_i = 0.$$

Using independence and the common error variance,

$$\begin{aligned}
\text{Var}(\tilde{\beta}_0) &= \sigma^2 \sum_{i=1}^{n} a_i^2 = \sigma^2 \sum_{i=1}^{n} (b_i + c_i)^2 \\
&= \sigma^2 \sum_{i=1}^{n} b_i^2 + \sigma^2 \sum_{i=1}^{n} c_i^2 + 2\sigma^2 \sum_{i=1}^{n} b_i c_i \\
&= \text{Var}(\hat{\beta}_0) + \sigma^2 \sum_{i=1}^{n} c_i^2 \geq \text{Var}(\hat{\beta}_0).
\end{aligned}$$

Thus $\hat{\beta}_0$ is the best linear unbiased estimator of β_0. □

As a consequence, we have an expression for the variance of the fitted mean at a given input $x = x_0$.

Corollary 3.3

$$\text{Var}(\hat{\mu}_0) = \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]\sigma^2,$$

where $\hat{\mu}_0$ is the fitted value at $x = x_0$; that is, $\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

Recall that $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, so

$$\hat{\mu}_0 = \bar{y} + \hat{\beta}_1(x_0 - \bar{x}) = \sum_{i=1}^{n}\left[\frac{1}{n} + \frac{(x_i - \bar{x})(x_0 - \bar{x})}{S_{xx}}\right] y_i,$$

and the stated variance formula follows by the same linear combination argument. To make $\text{Var}(\hat{\mu}_0)$ small, it is helpful to have a wide range of x-values, so that S_{xx} is large.

The least squares fit also satisfies several useful identities.

Proposition 3.4 The least squares fit satisfies

1. $$\sum_{i=1}^{n} e_i = 0,$$
   and hence $\bar{e} = 0$.
2. $$\sum_{i=1}^{n} e_i x_i = 0,$$
   so the residual–predictor cross-product is zero. Consequently, the sample correlation between e and x is zero whenever the residual variance is positive.
3. $$\sum_{i=1}^{n} \hat{\mu}_i e_i = 0.$$
4. The point $(\bar{x}, \bar{y})$ lies on the fitted regression line.

Proof. For (1) and (2), the normal equations give

$$S = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^{n} e_i^2.$$

Setting $\dfrac{\partial S}{\partial \beta_0} = 0$ gives

$$\sum_{i=1}^{n} e_i = 0,$$

and setting $\dfrac{\partial S}{\partial \beta_1} = 0$ gives

$$\sum_{i=1}^{n} e_i x_i = 0.$$

Since $\bar{e} = 0$, we also have

$$\sum_{i=1}^{n} (e_i - \bar{e})(x_i - \bar{x}) = \sum_{i=1}^{n} e_i x_i = 0.$$

For (3),

$$\sum_{i=1}^{n} \hat{\mu}_i e_i = \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i) e_i = \hat{\beta}_0 \sum_{i=1}^{n} e_i + \hat{\beta}_1 \sum_{i=1}^{n} e_i x_i = 0.$$

For (4),

$$\hat{\mu}(\bar{x}) = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} = (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 \bar{x} = \bar{y}.$$

□

Proposition 3.5 The least squares estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2},$$

and it is unbiased for σ^2.

Proof. Let

$$\text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

In simple linear regression with intercept, we have

$$\text{SSE} = S_{yy} - \frac{S_{xy}^2}{S_{xx}},$$

where $S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$ and $S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x}) y_i$.

First, compute $\mathbb{E}[S_{yy}]$. Since

$$y_i - \bar{y} = \beta_1 (x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon}),$$

we get

$$S_{yy} = \beta_1^2 S_{xx} + 2\beta_1 \sum_{i=1}^{n} (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon}) + \sum_{i=1}^{n} (\varepsilon_i - \bar{\varepsilon})^2.$$

The middle term has expectation zero, and

$$\mathbb{E}\left[\sum_{i=1}^{n} (\varepsilon_i - \bar{\varepsilon})^2\right] = \mathbb{E}\left[\sum_{i=1}^{n} \varepsilon_i^2 - n\bar{\varepsilon}^2\right] = n\sigma^2 - n\frac{\sigma^2}{n} = (n-1)\sigma^2.$$

Hence

$$\mathbb{E}[S_{yy}] = \beta_1^2 S_{xx} + (n-1)\sigma^2.$$

Next, compute $\mathbb{E}[S_{xy}^2]$. Because

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})y_i = \beta_1 S_{xx} + \sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i,$$

we have

$$\mathbb{E}[S_{xy}] = \beta_1 S_{xx}, \qquad \text{and} \qquad \text{Var}(S_{xy}) = \sigma^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sigma^2 S_{xx}.$$

Therefore,

$$\mathbb{E}[S_{xy}^2] = \text{Var}(S_{xy}) + \big(\mathbb{E}[S_{xy}]\big)^2 = \sigma^2 S_{xx} + \beta_1^2 S_{xx}^2.$$

Putting these together,

$$\begin{aligned}
\mathbb{E}[\text{SSE}] &= \mathbb{E}[S_{yy}] - \frac{\mathbb{E}[S_{xy}^2]}{S_{xx}} \\
&= \big(\beta_1^2 S_{xx} + (n-1)\sigma^2\big) - \frac{\sigma^2 S_{xx} + \beta_1^2 S_{xx}^2}{S_{xx}} \\
&= (n-2)\sigma^2.
\end{aligned}$$

Thus,

$$\mathbb{E}[\hat{\sigma}^2] = \mathbb{E}\left[\frac{\text{SSE}}{n-2}\right] = \frac{1}{n-2}\mathbb{E}[\text{SSE}] = \sigma^2,$$

so $\hat{\sigma}^2$ is unbiased for σ^2. □

The denominator $n-2$ appears because n is the number of observations and two parameters (β_0, β_1) are estimated. The key identity is

$$\mathbb{E}\left[\sum_{i=1}^{n} e_i^2\right] = (n-2)\sigma^2,$$

so

$$\mathbb{E}\left[\frac{1}{n-2}\sum_{i=1}^{n} e_i^2\right] = \sigma^2,$$

which implies that $\hat{\sigma}^2$ is unbiased.

Hypothesis Testing and Confidence Intervals

The slope β_1 is typically the main parameter of interest. To ask whether y has a nonzero linear association with x, we test

$$H_0 : \beta_1 = 0 \qquad \text{versus} \qquad H_A : \beta_1 \neq 0.$$

This is a two-sided test. For a one-sided question about whether the average change in y exceeds a specified value β, we use

$$H_0 : \beta_1 \leq \beta \qquad \text{versus} \qquad H_A : \beta_1 > \beta,$$

which asks whether a one-unit increase in x is associated with an average increase of more than β units in y.

As usual, the conclusion is stated as "reject H_0" or "fail to reject H_0." We *never* "accept H_0".

Lecture 4 — Friday, May 16, 2014

Recall that $\hat{\beta}_1 = S_{xy}/S_{xx}$, $\text{Var}(\hat{\beta}_1) = \sigma^2/S_{xx}$, $\widehat{\text{Var}}(\hat{\beta}_1) = \hat{\sigma}^2/S_{xx}$, and

$$\text{se}(\hat{\beta}_1) = \sqrt{\widehat{\text{Var}}(\hat{\beta}_1)} = \frac{\hat{\sigma}}{\sqrt{S_{xx}}},$$

where

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n-2}.$$

For inference on β_1, the key standardized quantity (or **discrepancy measure**) is

$$\frac{\hat{\beta}_1 - \beta_1}{\text{se}(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{S_{xx}}},$$

which measures the distance between estimate and target value in standard error units. We now determine its sampling distribution.

We use the following standard sampling distributions.

Lemma 4.1

1. If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $(X - \mu)/\sigma \sim \mathcal{N}(0, 1)$.
2. If $Z_1, \ldots, Z_n$ are independent and each has distribution $\mathcal{N}(0, 1)$, then $Z_i^2 \sim \chi^2_{(1)}$ for $i = 1, \ldots, n$, and $Z_1^2 + \cdots + Z_n^2 \sim \chi^2_{(n)}$.
3. If $Z \sim \mathcal{N}(0, 1)$ and $U \sim \chi^2_{(n)}$ where Z and U are independent, then $Z/\sqrt{U/n} \sim t_{(n)}$.
4. $(\chi^2_{(m)}/m)/(\chi^2_{(n)}/n) \sim F(m, n)$ when numerator and denominator are independent.

Here and below, $t_\alpha(\nu)$ and $F_\alpha(r, s)$ denote upper-tail critical values: $\mathbb{P}(T > t_\alpha(\nu)) = \alpha$ for $T \sim t(\nu)$, and similarly for the $F(r, s)$ distribution.

Theorem 4.2 Under the normal simple linear regression model,

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{S_{xx}}} \sim t_{(n-2)}.$$

Proof. We proceed in three steps.

Step 1. Since $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed as $\mathcal{N}(0, \sigma^2)$, we have

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2).$$

Writing $c_i = (x_i - \bar{x})/S_{xx}$, we obtain

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n} c_i y_i,$$

so $\hat{\beta}_1$ is normally distributed. Since $\mathbb{E}\left[\hat{\beta}_1\right] = \beta_1$ and $\mathrm{Var}(\hat{\beta}_1) = \sigma^2/S_{xx}$, it follows that

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right) \qquad \text{and} \qquad \frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim \mathcal{N}(0, 1).$$

Step 2. Let $\mathbf{X}$ be the $n \times 2$ design matrix with columns $\mathbf{1}$ and $(x_1, \ldots, x_n)^\top$, and set

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top.$$

Then $\boldsymbol{e} = (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}$. The matrix $\mathbf{H}$ is the orthogonal projection onto the two-dimensional column space of $\mathbf{X}$. Consequently, $\mathbf{I} - \mathbf{H}$ is symmetric and idempotent with rank $n-2$, so an orthogonal matrix $\mathbf{P}$ satisfies

$$\mathbf{I} - \mathbf{H} = \mathbf{P}\,\mathrm{diag}(\mathbf{I}_{n-2}, \mathbf{O}_2)\mathbf{P}^\top.$$

Because $\boldsymbol{z} = \mathbf{P}^\top\boldsymbol{\varepsilon}/\sigma \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, its components are independent standard normal variables. Hence

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} = \frac{\boldsymbol{e}^\top\boldsymbol{e}}{\sigma^2} = \sum_{i=1}^{n-2} z_i^2 \sim \chi^2_{(n-2)}.$$

Step 3. The vectors $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y}$ and $\boldsymbol{e} = (\mathbf{I} - \mathbf{H})\boldsymbol{y}$ are jointly normal, and

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}, \boldsymbol{e}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{I} - \mathbf{H}) = \mathbf{O}.$$

They are therefore independent. In particular, $\hat{\beta}_1$ is independent of $\hat{\sigma}^2 = \boldsymbol{e}^\top\boldsymbol{e}/(n-2)$. For (3), Lemma 4.1 now gives

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \Big/ \sqrt{\frac{(n-2)\hat{\sigma}^2/\sigma^2}{n-2}} \sim t_{(n-2)}.$$

Simplifying yields

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{S_{xx}}} \sim t_{(n-2)}.$$

□

A $100(1-\alpha)\%$ confidence interval for β_1 is obtained as usual. For example, $\alpha = 0.05$ gives a 95% interval.

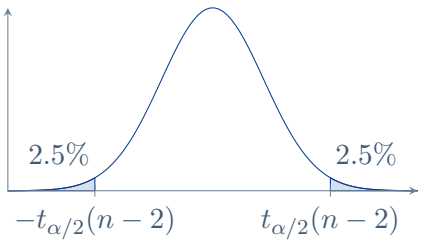


FIGURE 6

Figure 6 shows the two tails used for slope inference. With cutoff $t_{0.025}(n-2)$, we have $\alpha = 0.05$.

We use

$$\mathbb{P}\left(\left|\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{S_{xx}}}\right| < t_{\alpha/2}(n-2)\right) = 1 - \alpha.$$

Therefore, the interval has the form

$$\hat{\beta}_1 \pm t_{\alpha/2}(n-2) \cdot \text{se}(\hat{\beta}_1),$$

which follows the standard template estimate $\pm$ (critical value) $\cdot$ (standard error).

Corollary 4.3 A $100(1-\alpha)\%$ confidence interval for β_1 is

$$\hat{\beta}_1 \pm t_{\alpha/2}(n-2) \cdot \text{se}(\hat{\beta}_1).$$

The repeated-sampling interpretation is that if this construction were applied to many independent samples, then $100(1-\alpha)\%$ of the resulting intervals would contain the fixed value β_1. For the observed sample, the interval may be described informally as the range of parameter values that remain plausible at the corresponding confidence level.

These distributional results lead directly to the usual hypothesis tests. For a two-tailed test, we use $H_0 : \beta_1 = b$ versus $H_A : \beta_1 \neq b$. Under H_0,

$$\frac{\hat{\beta}_1 - b}{\hat{\sigma}/\sqrt{S_{xx}}} \sim t_{(n-2)}.$$

The sample test statistic is therefore

$$t^* = \frac{\hat{\beta}_1 - b}{\hat{\sigma}/\sqrt{S_{xx}}}.$$

At significance level α, we reject H_0 if $|t^*| > t_{\alpha/2}(n-2)$.

For an upper-tailed test, we use $H_0 : \beta_1 \leq b$ versus $H_A : \beta_1 > b$. The test statistic is again $t^* = (\hat{\beta}_1 - b)/(\hat{\sigma}/\sqrt{S_{xx}})$. We reject H_0 if $t^* > t_\alpha(n-2)$; otherwise, we fail to reject H_0.

Suppose we construct a 95% confidence interval for β_1. If we test $H_0 : \beta_1 = b$ versus $H_A : \beta_1 \neq b$ at the 5% significance level, then we reject H_0 if and only if b does not lie in that interval.

The following worked example illustrates these calculations in a concrete data set.

$\bar{x} = 148.81$, $\bar{y} = 1822.3112$, $S_{xx} = 731093.39$, $S_{yy} = 47854282.3674$, $S_{xy} = 4686071.3864$.

Coefficient	Estimate	Standard error	Test statistic
Intercept	$\hat{\beta}_0$	$\text{se}(\hat{\beta}_0) = \hat{\sigma}\sqrt{1/n + \bar{x}^2/S_{xx}}$	$\hat{\beta}_0/\text{se}(\hat{\beta}_0)$
Hours	$\hat{\beta}_1$	$\text{se}(\hat{\beta}_1) = \hat{\sigma}/\sqrt{S_{xx}}$	$\hat{\beta}_1/\text{se}(\hat{\beta}_1)$

R also reports the corresponding two-sided p-values for the tests $H_0 : \beta_i = 0$.

Residual standard error: $\hat{\sigma}$ on $n - 2$ degrees of freedom.

1. $$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = 6.409675 \qquad \text{and} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 868.487437.$$

2. $$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n-2}} = \sqrt{\frac{17818085.4344}{98}} = 426.400279.$$

3. R^2: the closer this value is to 1, the better the model fits the data. Here 0.628 indicates a reasonably good fit.

4. In R, `cor(x, y)` computes the sample correlation.

5. Method 1 gives
$$t^* = \frac{\hat{\beta}_1 - 0}{\text{se}(\hat{\beta}_1)} = \frac{6.410}{0.499} = 12.853$$
under $H_0 : \beta_1 = 0$. Since $|t^*| > t_{0.025}(98) = 1.9845$, we reject H_0 in favor of H_A.
Method 2: the p-value from the table is $< 2 \cdot 10^{-16} < 5\%$, so we again reject H_0.

6. Here the test is $H_0 : \beta_1 \leq 6$ versus $H_A : \beta_1 > 6$. The largest rejection probability under the null occurs at its boundary $\beta_1 = 6$, where
$$t^* = \frac{\hat{\beta}_1 - 6}{\text{se}(\hat{\beta}_1)} = \frac{6.410 - 6}{0.499} = 0.8215.$$

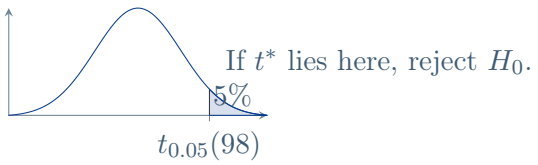


FIGURE 7

Figure 7 shows the upper-tail rejection region for this test.

$0.8215 < t_{0.05}(98) = 1.6606$, so there is no evidence against $H_0 : \beta_1 \leq 6$ at the 5% significance level.

7. Now we use $t_{0.05}(98) = 1.6606$. Our estimate is $\hat{\beta}_1 \pm t_{0.05}(98) \cdot \text{se}(\hat{\beta}_1)$, so

$$6.410 \pm 1.6606 \cdot 0.499 = (5.5816, 7.2378).$$

Thus we are 90% confident that one additional flight hour is associated with an average increase between 5.5816 and 7.2378 cost units in maintenance cost. This interval contains values both < 6 and ≥ 6, which supports the conclusion in the previous part.

8. The predicted value is

$$\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 \cdot 120 = 868.487437 + 6.409675(120) = 1637.648497.$$

A 95% prediction interval is

$$\hat{y}_p \pm t_{\alpha/2}(98) \cdot \sqrt{\widehat{\text{Var}}(\hat{y}_p - y_p)} = \hat{y}_p \pm t_{0.025}(98) \cdot \sqrt{\hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(120 - \bar{x})^2}{S_{xx}}\right]}$$

and therefore

$$1637.648497 \pm 1.984467 \cdot 428.767756 = (786.772839, 2488.524155).$$

Lecture 5 — Friday, May 23, 2014

Prediction Intervals

Recall from the previous lecture that

$$\text{Test statistic} = \frac{\text{estimate} - \text{assumed value}}{\text{standard error}},$$

and

$$\text{Confidence interval} = \text{estimate} \pm \text{critical value} \cdot \text{standard error}.$$

We now turn to prediction at a new covariate value $x = x_p$. The fitted predictor

$$\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p$$

is computed from the sample and estimates the mean response at x_p. A future response

$$y_p = \beta_0 + \beta_1 x_p + \varepsilon_p$$

contains a new error ε_p that is independent of the fitted sample. The main quantity in a prediction interval is therefore the **prediction error** $\hat{y}_p - y_p$.

Proposition 5.1 For a future response at $x = x_p$, the prediction error satisfies

$$\mathbb{E}\left[\hat{y}_p - y_p\right] = 0$$

and

$$\mathrm{Var}(\hat{y}_p - y_p) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}\right].$$

Proof. Because $\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p$ and $y_p = \beta_0 + \beta_1 x_p + \varepsilon_p$,

$$\mathbb{E}\left[\hat{y}_p - y_p\right] = \mathbb{E}\left[\hat{y}_p\right] - \mathbb{E}\left[y_p\right] = \mathbb{E}\left[\hat{\beta}_0 + \hat{\beta}_1 x_p\right] - \mathbb{E}\left[\beta_0 + \beta_1 x_p + \varepsilon_p\right] = \beta_0 + \beta_1 x_p - (\beta_0 + \beta_1 x_p + 0) = 0.$$

Also,

$$\begin{aligned}\mathrm{Var}(\hat{y}_p - y_p) &= \mathrm{Var}(\hat{y}_p) + \mathrm{Var}(y_p) = \mathrm{Var}(\hat{\mu}_p) + \sigma^2 \\ &= \sigma^2 \left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}\right] + \sigma^2 = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}\right].\end{aligned}$$

□

Corollary 5.2 The prediction error has estimated standard deviation

$$\mathrm{se}(\hat{y}_p - y_p) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}},$$

so a $100(1-\alpha)\%$ prediction interval for y_p is

$$\hat{y}_p \pm t_{\alpha/2}(n-2) \cdot \mathrm{se}(\hat{y}_p - y_p).$$

Analysis of Variance (ANOVA)

We can also use **ANOVA** in the simple regression case to test $H_0 : \beta_1 = 0$.

Our model is $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, for $i = 1, \ldots, n$. Also, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, $\hat{\beta}_1 = S_{xy}/S_{xx}$, and $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. If $\beta_1 = 0$, then $y_i = \beta_0 + \varepsilon_i$ and $\hat{\beta}_0 = \bar{y}$. The idea of ANOVA is to separate the total variability, written SST, into two components: variability due to regression, SSR, and variability due to error, SSE.

Proposition 5.3 The total sum of squares decomposes as

$$\text{SST} = \text{SSR} + \text{SSE},$$

where

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 = \sum_{i=1}^{n} e_i^2$$

and

$$\text{SSR} = \sum_{i=1}^{n} (\hat{\mu}_i - \bar{y})^2 = \hat{\beta}_1^2 S_{xx}.$$

Proof. Write

$$y_i - \bar{y} = (y_i - \hat{\mu}_i) + (\hat{\mu}_i - \bar{y}).$$

Squaring and summing over i gives

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 + \sum_{i=1}^{n} (\hat{\mu}_i - \bar{y})^2 + 2 \sum_{i=1}^{n} (y_i - \hat{\mu}_i)(\hat{\mu}_i - \bar{y}).$$

The first term is SSE. For the second term,

$$\text{SSR} = \sum_{i=1}^{n} (\hat{\mu}_i - \bar{y})^2 = \sum_{i=1}^{n} (\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i - \bar{y})^2 = \sum_{i=1}^{n} \hat{\beta}_1^2 (x_i - \bar{x})^2 = \hat{\beta}_1^2 S_{xx}.$$

For the cross term,

$$2 \sum_{i=1}^{n} (y_i - \hat{\mu}_i)(\hat{\mu}_i - \bar{y}) = 2 \sum_{i=1}^{n} e_i \big(\hat{\beta}_1 (x_i - \bar{x})\big) = 2\hat{\beta}_1 \sum_{i=1}^{n} e_i (x_i - \bar{x}) = 2\hat{\beta}_1 \left[\sum_{i=1}^{n} e_i x_i - \bar{x} \sum_{i=1}^{n} e_i \right] = 0.$$

Hence SST = SSR + SSE. □

Theorem 5.4 Under $H_0 : \beta_1 = 0$,

$$F^* = \frac{\text{SSR}/1}{\text{SSE}/(n-2)} \sim F(1, n-2).$$

Proof. Since

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2,$$

we have $\mathbb{E}\left[\hat{\sigma}^2\right] = \sigma^2$. Also,

$$\mathbb{E}\left[\text{SSR}\right] = \mathbb{E}\left[\hat{\beta}_1^2 S_{xx}\right] = S_{xx}\mathbb{E}\left[\hat{\beta}_1^2\right] = S_{xx}\left[\text{Var}(\hat{\beta}_1) + \mathbb{E}\left[\hat{\beta}_1\right]^2\right] = S_{xx}\left[\frac{\sigma^2}{S_{xx}} + \beta_1^2\right] = \sigma^2 + \beta_1^2 S_{xx}.$$

Under $H_0 : \beta_1 = 0$,

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\text{Var}(\hat{\beta}_1)}} = \frac{\hat{\beta}_1}{\sqrt{\text{Var}(\hat{\beta}_1)}} \sim \mathcal{N}(0,1),$$

so

$$\frac{\hat{\beta}_1^2}{\text{Var}(\hat{\beta}_1)} \sim \chi^2_{(1)}.$$

Because $\text{Var}(\hat{\beta}_1) = \sigma^2/S_{xx}$, this becomes

$$\frac{\hat{\beta}_1^2 S_{xx}}{\sigma^2} = \frac{\text{SSR}}{\sigma^2} \sim \chi^2_{(1)}.$$

Also,

$$\frac{\text{SSE}}{\sigma^2} = \sum_{i=1}^{n} \frac{e_i^2}{\sigma^2} \sim \chi^2_{(n-2)}.$$

The projection argument used in the slope t-test shows that $\hat{\boldsymbol{\beta}}$ is independent of the residual vector $\boldsymbol{e}$. Since $\text{SSR} = \hat{\beta}_1^2 S_{xx}$ is a function of $\hat{\boldsymbol{\beta}}$ and $\text{SSE} = \boldsymbol{e}^\top \boldsymbol{e}$ is a function of $\boldsymbol{e}$, the two sums of squares are independent. It follows that

$$\frac{\text{SSR}/\sigma^2}{1} \Big/ \frac{\text{SSE}/\sigma^2}{n-2} = \frac{\text{SSR}/1}{\text{SSE}/(n-2)} \sim F(1, n-2).$$

□

Remark 5.5 We often write MSR = SSR/1 and MSE = SSE/$(n-2)$ so that F = MSR/MSE $\sim F(1, n-2)$.

So the general rule is to reject H_0 when

$$F^* = \frac{\text{SSR}/1}{\text{SSE}/(n-2)} > F_\alpha(1, n-2).$$

Lecture 6 — Wednesday, May 28, 2014

The Coefficient of Determination

Recall that the ANOVA F-statistic is

$$\frac{\text{MSR}}{\text{MSE}} = \frac{\text{SSR}/1}{\text{SSE}/(n-2)} = \frac{\hat{\beta}_1^2 S_{xx}}{\hat{\sigma}^2} \sim F(1, n-2),$$

for testing $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$.

The **coefficient of determination** (or **R-squared statistic**) is the quantity

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = \frac{\text{SSR}}{\text{SST}},$$

which measures goodness of fit. Note that SSE represents unexplained variability and SST represents total variability.

Proposition 6.1 Assume SST $= S_{yy} > 0$, so that R^2 and the sample correlation are defined. Then the coefficient of determination satisfies

1. $0 \leq R^2 \leq 1$.
2. $R^2 = 1$ if SSE $= 0$.
3. $R^2 = 0$ if SSR $= 0$.

4.
$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\hat{\beta}_1^2 S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \left(\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}\right)^2 = r^2,$$
where r is the sample correlation coefficient.

Proof. For (1), note that
$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}}.$$
Because $\text{SSE} = \sum_{i=1}^{n} e_i^2 \geq 0$ and $\text{SST} = \text{SSR} + \text{SSE}$ with $\text{SSR} \geq 0$, we have
$$0 \leq \frac{\text{SSE}}{\text{SST}} \leq 1.$$
Hence $0 \leq R^2 \leq 1$.

For (2), if $\text{SSE} = 0$, then
$$R^2 = 1 - \frac{0}{\text{SST}} = 1.$$

For (3), if $\text{SSR} = 0$, then
$$R^2 = \frac{\text{SSR}}{\text{SST}} = 0.$$

For (4), use $\hat{\beta}_1 = S_{xy}/S_{xx}$ and
$$\text{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 = \hat{\beta}_1^2 S_{xx}.$$
Therefore,
$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\hat{\beta}_1^2 S_{xx}}{S_{yy}}.$$
Substituting $\hat{\beta}_1 = S_{xy}/S_{xx}$ gives
$$R^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \left(\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}\right)^2 = r^2,$$
where r is the sample correlation coefficient. □

2. Matrix Methods and Multiple Linear Regression

Before moving further into multiple regression, we review the matrix and random vector facts used throughout the course.

Write $\mathbf{A}_{m\times n} = (a_{ij})_{m\times n}$ for a matrix of constants with m rows and n columns; we omit the subscripts when these values are clear. Likewise,

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix},$$

denote vectors of constants.

Definition 6.2

1. $\mathbf{A}_{n\times n}$ is **symmetric** when $\mathbf{A}^\top = \mathbf{A}$.
2. Two vectors are **orthogonal** when $\boldsymbol{x}^\top \boldsymbol{y} = 0$. The matrix $\mathbf{A}$ is **orthogonal** when
$$\mathbf{A}^\top \mathbf{A} = \mathbf{A}\mathbf{A}^\top = \mathbf{I}.$$
3. Vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k$ are **linearly independent** if and only if $c_1\boldsymbol{x}_1 + \cdots + c_k\boldsymbol{x}_k = 0$ only when
$$c_1 = \cdots = c_k = 0.$$
4. The **rank** of $\mathbf{A}_{m\times n}$ is the maximum number of linearly independent columns.
5. The **trace** of $\mathbf{A}_{n\times n}$ is
$$\operatorname{tr}(\mathbf{A}_{n\times n}) = \sum_{i=1}^{n} a_{ii}$$
6. A vector $\boldsymbol{v}_j \neq \mathbf{0}$ is an **eigenvector** of $\mathbf{A}$ when there exists $\lambda \in \mathbb{R}$ such that $\mathbf{A}\boldsymbol{v}_j = \lambda \boldsymbol{v}_j$, where λ is the corresponding **eigenvalue**.
7. $\mathbf{A}_{n\times n}$ is **idempotent** when $\mathbf{A}^2 = \mathbf{A}$.

The preceding matrix notions satisfy the following standard facts.

Lemma 6.3

1. If $\mathbf{A}_{m\times n}$ and $\mathbf{B}_{n\times m}$ are conformable, then
$$\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA}).$$
2. For a symmetric matrix $\mathbf{A}_{n\times n}$, the eigenvalues $\lambda_1, \ldots, \lambda_n$ are real-valued. Furthermore, there exists an orthogonal matrix $\mathbf{P}$ such that $\mathbf{A} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^\top$, where $\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}$ and $\mathbf{P} = \begin{bmatrix} \boldsymbol{v}_1 & \ldots & \boldsymbol{v}_n \end{bmatrix}$.
3. A symmetric idempotent matrix has the following properties:
   (a) All eigenvalues of $\mathbf{A}$ are either 0 or 1.
   (b) There exists an orthogonal matrix $\mathbf{P}$ such that $\mathbf{A} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^\top$, where $\mathbf{\Lambda} = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 0 \end{bmatrix}$.
   (c)
   $$\text{tr}(\mathbf{A}) = \text{rank}(\mathbf{A}) = \sum_{i=1}^{n} \text{eigenvalue}_i,$$
   which is the number of ones in $\mathbf{\Lambda}$.

Proof of the idempotent claims. For the idempotent claims, if $\mathbf{A}\boldsymbol{v} = \lambda\boldsymbol{v}$, then

$$\mathbf{A}^2\boldsymbol{v} = \lambda^2\boldsymbol{v}.$$

Since $\mathbf{A}^2 = \mathbf{A}$, we obtain

$$\lambda\boldsymbol{v} = \lambda^2\boldsymbol{v},$$

so $\lambda \in \{0, 1\}$ because $\boldsymbol{v} \neq \mathbf{0}$. The spectral form follows from symmetry, and then

$$\text{tr}(\mathbf{A}) = \text{tr}\Big(\mathbf{P}\mathbf{\Lambda}\mathbf{P}^\top\Big) = \text{tr}\Big(\mathbf{P}^\top\mathbf{P}\mathbf{\Lambda}\Big) = \text{tr}(\mathbf{\Lambda}),$$

which equals both the rank and the number of ones on the diagonal of $\mathbf{\Lambda}$. □

We begin with the notion of a random vector.

Definition 6.4 Suppose $y_1, y_2, \ldots, y_n$ are random variables such that $\mathbb{E}[y_i] = \mu_i$, $\text{Var}(y_i) = \sigma_i^2$, and $\text{Cov}(y_i, y_j) = \sigma_{ij}$. Then $\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ is called a **random vector**.

The expected value of $\boldsymbol{y}$ is

$$\mathbb{E}[\boldsymbol{y}] = \begin{bmatrix} \mathbb{E}[y_1] \\ \vdots \\ \mathbb{E}[y_n] \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix}.$$

The covariance matrix is $\text{Var}(\boldsymbol{y}) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \ldots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \ldots & \ldots & \sigma_n^2 \end{bmatrix}$, which is symmetric.

If $y_1, \ldots, y_n$ are uncorrelated, then $\sigma_{ij} = 0$ for $i \neq j$, and so $\text{Var}(\boldsymbol{y}) = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{bmatrix}$. If $\sigma_i^2 = \sigma^2$, then $\text{Var}(\boldsymbol{y}) = \sigma^2 \mathbf{I}_{n \times n}$.

Lecture 7 — Friday, May 30, 2014

Random Vectors and Covariance Matrices

Recall that a random vector $\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ is a vector of random variables. If $\mathbb{E}[y_i] = \mu_i$, then

$$\mathbb{E}[\boldsymbol{y}] = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix} = \boldsymbol{\mu}.$$

Its covariance matrix is $\{\text{Cov}(y_i, y_j)\}_{i,j=1}^n$. If $y_1, \ldots, y_n$ are independent and $\sigma_i^2 = \sigma^2$ for all i, then $\text{Var}(\boldsymbol{y}) = \sigma^2\mathbf{I}$. We can also write $\text{Var}(\boldsymbol{y}) = \mathbb{E}\left[(\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})^\top\right]$.

Proposition 7.1 Let $\boldsymbol{y}$ be a random vector and let $\mathbf{A}$, $\boldsymbol{b}$, and $\boldsymbol{c}$ be conformable constants. Then

1. $\mathbb{E}\left[\mathbf{A}_{p\times n}\boldsymbol{y}_{n\times 1} + \boldsymbol{b}_{p\times 1}\right] = \mathbf{A}\mathbb{E}\left[\boldsymbol{y}\right] + \boldsymbol{b}$.
2. $\text{Var}(\boldsymbol{y} + \boldsymbol{c}) = \text{Var}(\boldsymbol{y})$.
3. $\text{Var}(\mathbf{A}\boldsymbol{y}) = \mathbf{A}\,\text{Var}(\boldsymbol{y})\mathbf{A}^\top$.

Proof. For (3),

$$\begin{aligned}\text{Var}(\mathbf{A}\boldsymbol{y}) &= \mathbb{E}\left[(\mathbf{A}\boldsymbol{y} - \mathbf{A}\boldsymbol{\mu})(\mathbf{A}\boldsymbol{y} - \mathbf{A}\boldsymbol{\mu})^\top\right] \\ &= \mathbb{E}\left[(\mathbf{A}(\boldsymbol{y} - \boldsymbol{\mu}))(\mathbf{A}(\boldsymbol{y} - \boldsymbol{\mu}))^\top\right] = \mathbb{E}\left[\mathbf{A}(\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})^\top\mathbf{A}^\top\right] \\ &= \mathbf{A}\mathbb{E}\left[(\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})^\top\right]\mathbf{A}^\top = \mathbf{A}\,\text{Var}(\boldsymbol{y})\mathbf{A}^\top.\end{aligned}$$

□

We also need the multivariate normal distribution. A random vector $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ is **multivariate normal** when every linear combination $\boldsymbol{a}^\top\boldsymbol{y}$ is univariate normal. We write $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{y}]$ and $\boldsymbol{\Sigma} = \text{Var}(\boldsymbol{y})$. When $\boldsymbol{\Sigma}$ is positive definite, this distribution has density

$$f(\boldsymbol{y}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})\right]$$

for $-\infty < y_i < \infty$. The linear-combination definition also permits positive-semidefinite $\boldsymbol{\Sigma}$, including degenerate normal vectors that do not have a density on all of $\mathbb{R}^n$.

The following properties are useful.

Proposition 7.2 If $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then

1. If $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{v} = \mathbf{A}\boldsymbol{y} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$.
2. If $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and all off-diagonal covariances are zero, then $y_1, \ldots, y_n$ are indepen-

dent.

3. If $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{u} = \mathbf{A}\boldsymbol{y}$, and $\boldsymbol{w} = \mathbf{B}\boldsymbol{y}$, then $\boldsymbol{u}$ and $\boldsymbol{w}$ are independent if $\mathbf{A}\operatorname{Var}(\boldsymbol{y})\mathbf{B}^\top = \mathbf{O}$.
4. If $\boldsymbol{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, then $y_1, \ldots, y_n$ are independent and each follows $\mathcal{N}(0,1)$. Consequently,

$$\boldsymbol{y}^\top \boldsymbol{y} = \sum_{i=1}^{n} y_i^2 \sim \chi^2(n).$$

If $\boldsymbol{z} = \mathbf{P}\boldsymbol{y}$ where $\mathbf{P}$ is orthogonal, then $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. If $\boldsymbol{y} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^\top$ is positive definite, then $\boldsymbol{\Lambda}^{-1/2}\mathbf{P}^\top \boldsymbol{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Proof. For (1), linear transformations preserve multivariate normality. Hence

$$\boldsymbol{v} = \mathbf{A}\boldsymbol{y} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu},\ \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top),$$

where the mean and covariance are from Proposition 7.1.

For (2), the random vector $(y_1, \ldots, y_n)^\top$ is jointly normal. In a jointly normal vector, pairwise zero covariances imply mutual independence. Therefore, when all off-diagonal covariances are zero, $y_1, \ldots, y_n$ are independent.

For (3), define

$$\boldsymbol{u} = \mathbf{A}\boldsymbol{y} \qquad \text{and} \qquad \boldsymbol{w} = \mathbf{B}\boldsymbol{y}.$$

By (1), $(\boldsymbol{u}^\top, \boldsymbol{w}^\top)^\top$ is jointly normal. Also,

$$\operatorname{Cov}(\boldsymbol{u}, \boldsymbol{w}) = \mathbb{E}\left[(\boldsymbol{u} - \mathbf{A}\boldsymbol{\mu})(\boldsymbol{w} - \mathbf{B}\boldsymbol{\mu})^\top\right] = \mathbf{A}\operatorname{Var}(\boldsymbol{y})\mathbf{B}^\top.$$

If $\mathbf{A}\operatorname{Var}(\boldsymbol{y})\mathbf{B}^\top = \mathbf{O}$, then $\operatorname{Cov}(\boldsymbol{u}, \boldsymbol{w}) = \mathbf{O}$, so $\boldsymbol{u}$ and $\boldsymbol{w}$ are independent by (2).

For (4), if $\boldsymbol{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, each component has mean 0 and variance 1, so each $y_i \sim \mathcal{N}(0,1)$. Since $\mathbf{I}$ has zero off-diagonal entries, the components are uncorrelated, and by (2) they are independent. Therefore,

$$\boldsymbol{y}^\top \boldsymbol{y} = \sum_{i=1}^{n} y_i^2 \sim \chi^2(n).$$

If $\boldsymbol{z} = \mathbf{P}\boldsymbol{y}$ with $\mathbf{P}$ orthogonal, then by (1)

$$\boldsymbol{z} \sim \mathcal{N}(\mathbf{P}\mathbf{0}, \mathbf{P}\mathbf{I}\mathbf{P}^\top) = \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

Finally, if $\boldsymbol{y} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^\top$, let

$$\boldsymbol{q} = \boldsymbol{\Lambda}^{-1/2}\mathbf{P}^\top \boldsymbol{y}.$$

Again by (1), $\boldsymbol{q}$ is multivariate normal with mean $\mathbf{0}$ and covariance

$$\boldsymbol{\Lambda}^{-1/2}\mathbf{P}^\top \boldsymbol{\Sigma}\mathbf{P}\boldsymbol{\Lambda}^{-1/2} = \boldsymbol{\Lambda}^{-1/2}\mathbf{P}^\top(\mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^\top)\mathbf{P}\boldsymbol{\Lambda}^{-1/2} = \mathbf{I}.$$

Hence $\boldsymbol{\Lambda}^{-1/2}\mathbf{P}^\top \boldsymbol{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. □

The following differentiation rules are used repeatedly in the matrix derivations to come.

Lemma 7.3

1. Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable with $f(\boldsymbol{y}) = f(y_1, \ldots, y_n)$. Then

$$\frac{\partial f(\boldsymbol{y})}{\partial \boldsymbol{y}} = \begin{bmatrix} \frac{\partial f(\boldsymbol{y})}{\partial y_1} \\ \vdots \\ \frac{\partial f(\boldsymbol{y})}{\partial y_n} \end{bmatrix}.$$

2. If $\boldsymbol{c} = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix}$ and

$$f(\boldsymbol{y}) = \boldsymbol{c}^\top \boldsymbol{y} = \sum_{i=1}^n c_i y_i,$$

then

$$\frac{\partial f(\boldsymbol{y})}{\partial \boldsymbol{y}} = \boldsymbol{c}.$$

3. If $\mathbf{A} = (a_{ij})_{n\times n}$ and $f(\boldsymbol{y}) = \boldsymbol{y}^\top \mathbf{A}\boldsymbol{y}$, then

$$f(\boldsymbol{y}) = \sum_{i=1}^n \sum_{j=1}^n a_{ij} y_i y_j$$

and

$$\frac{\partial f(\boldsymbol{y})}{\partial \boldsymbol{y}} = (\mathbf{A} + \mathbf{A}^\top)\boldsymbol{y}.$$

In particular, the derivative is $2\mathbf{A}\boldsymbol{y}$ when $\mathbf{A}$ is symmetric.

We now turn to the multiple linear regression model.

Multiple linear regression assumes that y is linearly related to a collection of predictors. Here y is the **response variable** and $x_1, \ldots, x_p$ are the **predictor** or **explanatory variables**. The data consist of $\{(y_i, x_{i1}, \ldots, x_{ip}) : i = 1, \ldots, n\}$, where n is the sample size. The model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i.$$

As in simple regression, the predictor values are treated as fixed. We assume $n > p + 1$, $\sigma^2 > 0$, and that the design matrix $\mathbf{X}$, including its intercept column, has full column rank $p + 1$.

We assume the following:

1. $\mathbb{E}[\varepsilon_i] = 0$.
2. $\mathrm{Var}(\varepsilon_i) = \sigma^2$, which implies that $\mathrm{Var}(y_i) = \sigma^2$.
3. $\varepsilon_1, \ldots, \varepsilon_n$ are all independent, which implies that $y_1, \ldots, y_n$ are independent.
4. $\varepsilon_1, \ldots, \varepsilon_n$ are independent and each follows $\mathcal{N}(0, \sigma^2)$, which implies that $y_1, \ldots, y_n$ are independent with distributions $\mathcal{N}(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \sigma^2)$. In analogy to simple linear regression, this assumption implies Assumptions (1), (2), and (3).

The parameters are interpreted through

$$\mathbb{E}[y_i \mid x_{i1}, \ldots, x_{ip}] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

If x_{i1} is increased by c while the other predictors are held fixed, then

$$\mathbb{E}[y_i \mid x_{i1} + c, x_{i2}, \ldots, x_{ip}] = \beta_0 + \beta_1(x_{i1} + c) + \cdots + \beta_p x_{ip}.$$

Subtracting gives

$$\mathbb{E}[y_i \mid x_{i1} + c, x_{i2}, \ldots, x_{ip}] - \mathbb{E}[y_i \mid x_{i1}, \ldots, x_{ip}] = \beta_1 c.$$

Setting $c = 1$ shows that β_1 is the change in the conditional mean response when x_1 increases by one unit and all other predictors are held fixed. The same interpretation applies to each β_j.

Lecture 8 — Wednesday, June 04, 2014

Multiple Regression Least Squares Estimation

In multiple linear regression, $\beta_0, \ldots, \beta_p$ are unknown parameters, and $\hat{\beta}_0, \ldots, \hat{\beta}_p$ are their estimates from the sample. The least squares estimator of $\boldsymbol{\beta}$ is obtained by choosing $\hat{\beta}_0, \ldots, \hat{\beta}_p$ to minimize

$$S(\beta_0, \ldots, \beta_p) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2.$$

In matrix form,

$$\begin{aligned} S(\boldsymbol{\beta}) &= (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) = (\boldsymbol{y}^\top - \boldsymbol{\beta}^\top \mathbf{X}^\top)(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \boldsymbol{y}^\top \boldsymbol{y} - \boldsymbol{y}^\top \mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^\top \mathbf{X}^\top \boldsymbol{y} + \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{y}^\top \boldsymbol{y} - \boldsymbol{y}^\top \mathbf{X}\boldsymbol{\beta} - (\boldsymbol{\beta}^\top \mathbf{X}^\top \boldsymbol{y})^\top + \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{y}^\top \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}. \end{aligned}$$

Differentiating gives

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0 - 2(\boldsymbol{y}^\top \mathbf{X})^\top + 2(\mathbf{X}^\top \mathbf{X})\boldsymbol{\beta}.$$

Setting this equal to zero yields

$$-2\mathbf{X}^\top \boldsymbol{y} + 2\mathbf{X}^\top \mathbf{X}\hat{\boldsymbol{\beta}} = 0 \implies \mathbf{X}^\top \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^\top \boldsymbol{y}.$$

Assuming $\mathbf{X}$ has full column rank (that is, $\operatorname{rank}(\mathbf{X}) = p + 1$), $\mathbf{X}^\top \mathbf{X}$ is invertible, so

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{y}.$$

Moreover, the Hessian is $2\mathbf{X}^\top \mathbf{X}$, which is positive definite under full column rank. Thus this stationary point is the unique global minimizer. To obtain $\hat{\beta}_i$, take the ith entry of $\hat{\boldsymbol{\beta}}$ (for $i = 0, \ldots, p$), equivalently the ith row of $(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ multiplied by $\boldsymbol{y}$.

Proposition 8.1

1) $\mathbb{E}\left[\hat{\boldsymbol{\beta}}\right] = \boldsymbol{\beta}$, so $\hat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$.
2) $\operatorname{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$.

Proof. For (1)),

$$\mathbb{E}\left[\hat{\boldsymbol{\beta}}\right] = \mathbb{E}\left[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y}\right] = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbb{E}\left[\boldsymbol{y}\right] = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

For (2)),

$$\begin{aligned}\operatorname{Var}(\hat{\boldsymbol{\beta}}) &= \operatorname{Var}((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y}) = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\operatorname{Var}(\boldsymbol{y})[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top]^\top \\ &= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}.\end{aligned}$$

□

An estimator for σ^2 is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-(p+1)} = \frac{\boldsymbol{e}^\top\boldsymbol{e}}{n-p-1},$$

where $\boldsymbol{e} = \boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. Accordingly,

$$\widehat{\operatorname{Var}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\mathbf{X}^\top\mathbf{X})^{-1}.$$

The estimated covariance matrix can be written explicitly as

$$\widehat{\operatorname{Var}}(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} \widehat{\operatorname{Var}}(\hat{\beta}_0) & \widehat{\operatorname{Cov}}(\hat{\beta}_0,\hat{\beta}_1) & \cdots & \widehat{\operatorname{Cov}}(\hat{\beta}_0,\hat{\beta}_p) \\ \widehat{\operatorname{Cov}}(\hat{\beta}_1,\hat{\beta}_0) & \widehat{\operatorname{Var}}(\hat{\beta}_1) & \cdots & \widehat{\operatorname{Cov}}(\hat{\beta}_1,\hat{\beta}_p) \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{\operatorname{Cov}}(\hat{\beta}_p,\hat{\beta}_0) & \widehat{\operatorname{Cov}}(\hat{\beta}_p,\hat{\beta}_1) & \cdots & \widehat{\operatorname{Var}}(\hat{\beta}_p) \end{bmatrix}.$$

Lecture 9 — Friday, June 06, 2014

The Hat Matrix and Residual Structure

Recall that

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y}, \qquad \mathbb{E}\left[\hat{\boldsymbol{\beta}}\right] = \boldsymbol{\beta}, \qquad \text{and} \qquad \operatorname{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^\top\mathbf{X})^{-1}\sigma^2,$$

and

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n-p-1} = \frac{\boldsymbol{e}^\top \boldsymbol{e}}{n-p-1},$$

where $\boldsymbol{e} = \boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$.

The matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ is called the **hat matrix**.

Proposition 9.1 Let $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ be the hat matrix.

1. The fitted values satisfy $\hat{\boldsymbol{\mu}} = \mathbf{H}\boldsymbol{y}$.
2. The residual vector satisfies $\boldsymbol{e} = (\mathbf{I} - \mathbf{H})\boldsymbol{y}$.
3. $\mathbf{H}$ is symmetric and idempotent.
4. $\mathbf{I} - \mathbf{H}$ is idempotent.

Proof. For (1) and (2), the definitions give

$$\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y} = \mathbf{H}\boldsymbol{y}, \qquad \text{and} \qquad \boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{\mu}} = (\mathbf{I} - \mathbf{H})\boldsymbol{y}.$$

For (3),

$$\mathbf{HH} = [\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top][\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top] = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{I}\mathbf{X}^\top = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{H},$$

so $\mathbf{H}$ is idempotent. Moreover,

$$\mathbf{H}^\top = [\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top]^\top = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{H},$$

so $\mathbf{H}$ is symmetric.

For (4),

$$(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H}) = \mathbf{I}^2 - 2\mathbf{I}\mathbf{H} + \mathbf{H}^2 = \mathbf{I} - \mathbf{H},$$

so $\mathbf{I} - \mathbf{H}$ is idempotent. □

We have several further results and consequences of the least squares formulation:

Proposition 9.2

1. The fitted values satisfy $\hat{\boldsymbol{\mu}} = \mathbf{H}\boldsymbol{y}$. In particular, $\mathbb{E}[\hat{\boldsymbol{\mu}}] = \boldsymbol{\mu}$ and $\mathrm{Var}(\hat{\boldsymbol{\mu}}) = \sigma^2\mathbf{H}$.
2. The residual vector satisfies $\boldsymbol{e} = (\mathbf{I} - \mathbf{H})\boldsymbol{y}$. In particular, $\mathbb{E}[\boldsymbol{e}] = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$.
3. $\mathbf{X}^\top\boldsymbol{e} = \mathbf{0}$ and $\hat{\boldsymbol{\mu}}^\top\boldsymbol{e} = 0$. Equivalently, the residual vector is orthogonal to the fitted subspace.
4. Under the fourth model assumption, $\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$.
5. Under the fourth model assumption, $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{e}$ are independent of each other.
6. $$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-p-1}\boldsymbol{e}^\top\boldsymbol{e}$$
   is unbiased for σ^2. Under the fourth model assumption, it is also independent of $\hat{\boldsymbol{\beta}}$.
7. Under the fourth model assumption, the variance estimator satisfies
   $$\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p-1).$$

Proof. For (1),

$$\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y} = \mathbf{H}\boldsymbol{y}.$$

Therefore,

$$\mathbb{E}[\hat{\boldsymbol{\mu}}] = \mathbf{X}\mathbb{E}\left[\hat{\boldsymbol{\beta}}\right] = \mathbf{X}\boldsymbol{\beta} = \boldsymbol{\mu} \qquad \text{and} \qquad \mathrm{Var}(\hat{\boldsymbol{\mu}}) = \mathrm{Var}(\mathbf{H}\boldsymbol{y}) = \mathbf{H}\,\mathrm{Var}(\boldsymbol{y})\mathbf{H}^\top = \sigma^2\mathbf{H}.$$

For (2),

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{\mu}} = \boldsymbol{y} - \mathbf{H}\boldsymbol{y} = (\mathbf{I} - \mathbf{H})\boldsymbol{y}.$$

Hence,

$$\begin{aligned}\mathbb{E}[\boldsymbol{e}] &= (\mathbf{I} - \mathbf{H})\mathbb{E}[\boldsymbol{y}] = (\mathbf{I} - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = \mathbf{0},\\ \mathrm{Var}(\boldsymbol{e}) &= \mathrm{Var}((\mathbf{I} - \mathbf{H})\boldsymbol{y}) = (\mathbf{I} - \mathbf{H})\,\mathrm{Var}(\boldsymbol{y})(\mathbf{I} - \mathbf{H})^\top = \sigma^2(\mathbf{I} - \mathbf{H}).\end{aligned}$$

For (3),

$$\mathbf{X}^\top\boldsymbol{e} = \mathbf{X}^\top(\mathbf{I} - \mathbf{H})\boldsymbol{y} = (\mathbf{X}^\top - \mathbf{X}^\top\mathbf{H})\boldsymbol{y} = (\mathbf{X}^\top - \mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\boldsymbol{y} = (\mathbf{X}^\top - \mathbf{X}^\top)\boldsymbol{y} = \mathbf{0}.$$

For the second part,

$$\hat{\boldsymbol{\mu}}^\top \boldsymbol{e} = (\mathbf{X}\hat{\boldsymbol{\beta}})^\top \boldsymbol{e} = \hat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \boldsymbol{e} = \hat{\boldsymbol{\beta}}^\top \mathbf{0} = 0.$$

For (4), $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y}$ is a linear transformation of $\boldsymbol{y}$. Since $\boldsymbol{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, (1) of Proposition 7.2 yields

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}).$$

For (5), both $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{e}$ are linear transformations of $\boldsymbol{y}$, so (3) of Proposition 7.2 shows that it is enough to verify $\mathrm{Cov}(\hat{\boldsymbol{\beta}}, \boldsymbol{e}) = \mathbf{0}$. We compute

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}, \boldsymbol{e}) = \mathrm{Cov}((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y}, (\mathbf{I}-\mathbf{H})\boldsymbol{y}) = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \mathrm{Var}(\boldsymbol{y})(\mathbf{I}-\mathbf{H})^\top$$

$$= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\sigma^2(\mathbf{I}-\mathbf{H}) = \sigma^2\left[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top - (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{H}\right] = \mathbf{0}.$$

Therefore $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{e}$ are independent.

For (6), by (2) of Proposition 9.2,

$$\mathrm{Var}(\boldsymbol{e}) = \sigma^2(\mathbf{I}-\mathbf{H}) = \mathbb{E}\left[(\boldsymbol{e} - \mathbb{E}[\boldsymbol{e}])(\boldsymbol{e} - \mathbb{E}[\boldsymbol{e}])^\top\right] = \mathbb{E}\left[\boldsymbol{e}\boldsymbol{e}^\top\right].$$

Therefore,

$$\mathbb{E}\left[\boldsymbol{e}^\top\boldsymbol{e}\right] = \mathbb{E}\left[\mathrm{tr}\left(\boldsymbol{e}^\top\boldsymbol{e}\right)\right] = \mathbb{E}\left[\mathrm{tr}\left(\boldsymbol{e}\boldsymbol{e}^\top\right)\right] = \mathrm{tr}\left(\mathbb{E}\left[\boldsymbol{e}\boldsymbol{e}^\top\right]\right)$$
$$= \mathrm{tr}(\mathrm{Var}(\boldsymbol{e})) = \mathrm{tr}\big((\mathbf{I}-\mathbf{H})\sigma^2\big) = \sigma^2[n-(p+1)] = \sigma^2(n-p-1).$$

Hence

$$\mathbb{E}\left[\hat{\sigma}^2\right] = \mathbb{E}\left[\frac{1}{n-p-1}\boldsymbol{e}^\top\boldsymbol{e}\right] = \sigma^2,$$

so $\hat{\sigma}^2$ is unbiased. Since $\hat{\sigma}^2$ is a function of $\boldsymbol{e}$, (5) implies that $\hat{\sigma}^2$ is independent of $\hat{\boldsymbol{\beta}}$.

For (7), by (2) of Proposition 9.2, $\mathrm{Var}(\boldsymbol{e}) = \sigma^2(\mathbf{I}-\mathbf{H})$. Also, $\mathbf{I}-\mathbf{H}$ is symmetric, so by spectral decomposition $\mathbf{I}-\mathbf{H} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^\top$ where $\mathbf{P}$ is orthogonal. Because $\mathbf{I}-\mathbf{H}$ is idempotent, its eigenvalues are 0 or 1, and therefore $\boldsymbol{\Lambda} = \mathrm{diag}(1,\ldots,1,0,\ldots,0)$ with $n-p-1$ ones. Let

$$\boldsymbol{u} = \mathbf{P}^\top\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{\Lambda}).$$

Then $u_1,\ldots,u_{n-p-1}$ are independent and each follows $\mathcal{N}(0,\sigma^2)$, while $u_{n-p},\ldots,u_n = 0$. There-

fore,

$$\sum_{i=1}^{n-p-1} \left(\frac{u_i}{\sigma}\right)^2 = \sum_{i=1}^{n} \left(\frac{u_i}{\sigma}\right)^2 \sim \chi^2(n-p-1).$$

Now,

$$\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} = \frac{\boldsymbol{e}^\top \boldsymbol{e}}{\sigma^2} = \frac{(\mathbf{P}\boldsymbol{u})^\top(\mathbf{P}\boldsymbol{u})}{\sigma^2} = \frac{1}{\sigma^2}\boldsymbol{u}^\top\mathbf{P}^\top\mathbf{P}\boldsymbol{u} = \frac{\boldsymbol{u}^\top\boldsymbol{u}}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{u_i}{\sigma}\right)^2 \sim \chi^2(n-p-1).$$

□

In simple linear regression, this expansion can be written explicitly in matrix form. Let

$$\hat{\mu}_* = \hat{\beta}_0 + \hat{\beta}_1 x_* = \begin{bmatrix} 1 & x_* \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}, \qquad \text{and} \qquad \operatorname{Var}(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} V_{00} & V_{01} \\ V_{10} & V_{11} \end{bmatrix},$$

where $V_{00} = \operatorname{Var}(\hat{\beta}_0)$, $V_{11} = \operatorname{Var}(\hat{\beta}_1)$, and $V_{01} = V_{10} = \operatorname{Cov}(\hat{\beta}_0, \hat{\beta}_1)$. Then

$$\begin{aligned} \operatorname{Var}(\hat{\mu}_*) &= \begin{bmatrix} 1 & x_* \end{bmatrix} \begin{bmatrix} V_{00} & V_{01} \\ V_{10} & V_{11} \end{bmatrix} \begin{bmatrix} 1 \\ x_* \end{bmatrix} \\ &= V_{00} + V_{10}x_* + V_{01}x_* + V_{11}x_*^2 \\ &= \operatorname{Var}(\hat{\beta}_0) + 2x_* \operatorname{Cov}(\hat{\beta}_0, \hat{\beta}_1) + x_*^2 \operatorname{Var}(\hat{\beta}_1). \end{aligned}$$

Lecture 10 — Wednesday, June 11, 2014

The Gauss–Markov Theorem

We continue with another important property of the least squares estimator.

Theorem 10.1 (Gauss–Markov theorem) Under the fixed-design assumptions $\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$, $\operatorname{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$, and $\operatorname{rank}(\mathbf{X}) = p+1$, the least squares estimator

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y}$$

is the **best linear unbiased estimator** (**BLUE**) of $\boldsymbol{\beta}$. More precisely, its covariance matrix is minimal in the positive-semidefinite order among all linear unbiased estimators. Equivalently, every linear combination $\boldsymbol{x}_*^\top \hat{\boldsymbol{\beta}}$ has no greater variance than the same linear combination of any other linear unbiased estimator.

Proof. Consider another linear estimator $\tilde{\boldsymbol{\theta}} = \mathbf{M}\boldsymbol{y}$, where $\mathbf{M} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{A}$. Then

$$\mathbb{E}\left[\tilde{\boldsymbol{\theta}}\right] = \mathbb{E}\left[\mathbf{M}\boldsymbol{y}\right] = \mathbb{E}\left[((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{A})\boldsymbol{y}\right] = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} + \mathbf{A}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta} + \mathbf{A}\mathbf{X}\boldsymbol{\beta}.$$

Thus $\tilde{\boldsymbol{\theta}}$ is unbiased for every $\boldsymbol{\beta}$ if and only if $\mathbf{A}\mathbf{X} = \mathbf{O}$. Also,

$$\begin{aligned}
\operatorname{Var}(\tilde{\boldsymbol{\theta}}) &= \operatorname{Var}(\mathbf{M}\boldsymbol{y}) = \mathbf{M}\operatorname{Var}(\boldsymbol{y})\mathbf{M}^\top = \mathbf{M}\cdot\sigma^2\mathbf{I}\cdot\mathbf{M}^\top = \sigma^2\mathbf{M}\mathbf{M}^\top \\
&= \sigma^2[((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{A})((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{A})^\top] \\
&= \sigma^2[(\mathbf{X}^\top\mathbf{X})^{-1} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{A}^\top + \mathbf{A}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} + \mathbf{A}\mathbf{A}^\top] \\
&= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} + \sigma^2\mathbf{A}\mathbf{A}^\top = \operatorname{Var}(\hat{\boldsymbol{\beta}}) + \sigma^2\mathbf{A}\mathbf{A}^\top.
\end{aligned}$$

For any linear predictor $\boldsymbol{x}_*^\top\tilde{\boldsymbol{\theta}}$,

$$\begin{aligned}
\operatorname{Var}(\boldsymbol{x}_*^\top\tilde{\boldsymbol{\theta}}) &= \boldsymbol{x}_*^\top\operatorname{Var}(\tilde{\boldsymbol{\theta}})\boldsymbol{x}_* = \boldsymbol{x}_*^\top\operatorname{Var}(\hat{\boldsymbol{\beta}})\boldsymbol{x}_* + \boldsymbol{x}_*^\top\sigma^2\mathbf{A}\mathbf{A}^\top\boldsymbol{x}_* \\
&= \operatorname{Var}(\boldsymbol{x}_*^\top\hat{\boldsymbol{\beta}}) + (\mathbf{A}^\top\boldsymbol{x}_*)^\top(\mathbf{A}^\top\boldsymbol{x}_*)\sigma^2 \geq \operatorname{Var}(\boldsymbol{x}_*^\top\hat{\boldsymbol{\beta}}).
\end{aligned}$$

Thus $\hat{\boldsymbol{\beta}}$ has the smallest variance among linear unbiased estimators. □

Corollary 10.2 Under the normal multiple linear regression model, write

$$(\mathbf{X}^\top\mathbf{X})^{-1} = (j_{k\ell})_{k,\ell=0}^{p}.$$

Then

$$T = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{j_{jj}}} \sim t(n-p-1),$$

where $\operatorname{se}(\hat{\beta}_j) = \hat{\sigma}\sqrt{j_{jj}}$. Also,

$$\operatorname{Var}(y_* - \hat{y}_*) = \sigma^2\left(1 + \boldsymbol{x}_*^\top(\mathbf{X}^\top\mathbf{X})^{-1}\boldsymbol{x}_*\right).$$

In the simple two-parameter case,

$$\operatorname{Var}(\hat{\mu}_*) = \operatorname{Var}(\hat{\beta}_0) + 2x_* \operatorname{Cov}(\hat{\beta}_0, \hat{\beta}_1) + x_*^2 \operatorname{Var}(\hat{\beta}_1).$$

Proof. For (4), (5), and (7) of Proposition 9.2,

$$\frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{j_{jj}}} \sim \mathcal{N}(0,1), \qquad \text{and} \qquad \frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p-1),$$

and these two quantities are independent. Therefore,

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{j_{jj}}} = \frac{\frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{j_{jj}}}}{\sqrt{\frac{(n-p-1)\hat{\sigma}^2/\sigma^2}{n-p-1}}} \sim t(n-p-1).$$

Also,

$$\operatorname{Var}(\hat{y}_*) = \operatorname{Var}(\boldsymbol{x}_*^\top \hat{\boldsymbol{\beta}}) = \boldsymbol{x}_*^\top \operatorname{Var}(\hat{\boldsymbol{\beta}}) \boldsymbol{x}_* = \sigma^2 \boldsymbol{x}_*^\top (\mathbf{X}^\top \mathbf{X})^{-1} \boldsymbol{x}_*.$$

Since $y_* = \boldsymbol{x}_*^\top \boldsymbol{\beta} + \varepsilon_*$ with $\operatorname{Var}(\varepsilon_*) = \sigma^2$ and ε_* independent of $\hat{y}_*$, it follows that

$$\operatorname{Var}(y_* - \hat{y}_*) = \sigma^2 \left(1 + \boldsymbol{x}_*^\top (\mathbf{X}^\top \mathbf{X})^{-1} \boldsymbol{x}_*\right).$$

The simple two-parameter formula is the expansion of the quadratic form $\boldsymbol{x}_*^\top \operatorname{Var}(\hat{\boldsymbol{\beta}}) \boldsymbol{x}_*$. □

As an aside, the simple two-parameter variance expansion can be derived directly in matrix form. Let

$$\hat{\mu}_* = \hat{\beta}_0 + \hat{\beta}_1 x_* = \begin{bmatrix} 1 & x_* \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \qquad \text{and} \qquad \operatorname{Var}(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} V_{00} & V_{01} \\ V_{10} & V_{11} \end{bmatrix},$$

where $V_{00} = \operatorname{Var}(\hat{\beta}_0)$, $V_{11} = \operatorname{Var}(\hat{\beta}_1)$, and $V_{01} = V_{10} = \operatorname{Cov}(\hat{\beta}_0, \hat{\beta}_1)$. Then

$$\begin{aligned} \operatorname{Var}(\hat{\mu}_*) &= \begin{bmatrix} 1 & x_* \end{bmatrix} \begin{bmatrix} V_{00} & V_{01} \\ V_{10} & V_{11} \end{bmatrix} \begin{bmatrix} 1 \\ x_* \end{bmatrix} \\ &= V_{00} + V_{10} x_* + V_{01} x_* + V_{11} x_*^2 \\ &= \operatorname{Var}(\hat{\beta}_0) + 2x_* \operatorname{Cov}(\hat{\beta}_0, \hat{\beta}_1) + x_*^2 \operatorname{Var}(\hat{\beta}_1). \end{aligned}$$

Multiple Regression Inference Summary

The preceding distributional results give the following working forms. For $j = 0, 1, \ldots, p$, a $100(1-\alpha)\%$ confidence interval for β_j is

$$\hat{\beta}_j \pm t_{\alpha/2}(n-p-1)\hat{\sigma}\sqrt{j_{jj}}.$$

For the two-sided test $H_0 : \beta_j = b_0$ against $H_A : \beta_j \neq b_0$, use

$$t^* = \frac{\hat{\beta}_j - b_0}{\hat{\sigma}\sqrt{j_{jj}}}$$

and reject when $|t^*| > t_{\alpha/2}(n-p-1)$. For the upper-tailed test $H_0 : \beta_j \leq b_0$ against $H_A : \beta_j > b_0$, reject when $t^* > t_\alpha(n-p-1)$.

For a new unit with predictor vector $\boldsymbol{x}_* = (1, x_{*1}, \ldots, x_{*p})^\top$, the predicted response is $\hat{y}_* = \boldsymbol{x}_*^\top \hat{\boldsymbol{\beta}}$. Since the new response is independent of the fitted sample,

$$\mathbb{E}[y_* - \hat{y}_*] = 0, \qquad \text{and} \qquad \mathrm{Var}(y_* - \hat{y}_*) = \sigma^2 \left[1 + \boldsymbol{x}_*^\top (\mathbf{X}^\top \mathbf{X})^{-1} \boldsymbol{x}_*\right].$$

Thus a $100(1-\alpha)\%$ prediction interval is

$$\hat{y}_* \pm t_{\alpha/2}(n-p-1)\hat{\sigma}\sqrt{1 + \boldsymbol{x}_*^\top (\mathbf{X}^\top \mathbf{X})^{-1} \boldsymbol{x}_*}.$$

The corresponding overall analysis-of-variance test is as follows.

Theorem 10.3 (Overall multiple regression F-test) Under the normal multiple regression model and the null hypothesis

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0,$$

the statistic

$$F^* = \frac{\mathrm{SSR}/p}{\mathrm{SSE}/(n-p-1)} \sim F(p, n-p-1),$$

where

$$\mathrm{SSR} = \sum_{i=1}^{n} (\hat{\mu}_i - \bar{y})^2, \qquad \mathrm{SSE} = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2, \qquad \text{and} \qquad \mathrm{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

and $\text{SST} = \text{SSR} + \text{SSE}$. The alternative is that at least one slope coefficient is nonzero, and the test rejects for large values of F^*.

Proof. This is the special case of Theorem 20.2 obtained by taking $\mathbf{A} = [\mathbf{0}\ \mathbf{I}_p]$, $\boldsymbol{c} = \mathbf{0}$, and $r = p$. Under these restrictions the reduced model contains only an intercept, so $\text{SSE}_{\text{red}} = \text{SST}$ and $\text{SSE}_{\text{red}} - \text{SSE} = \text{SSR}$. □

Lecture 11 — Friday, June 13, 2014

3. Specification and Categorical Effects

Specification Issues and Multicollinearity

Remark 11.1 For the regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon,$$

the F-test is

$$H_0 : \beta_1 = \beta_2 = 0$$

versus H_A : at least one $\beta_j \neq 0$. This is a single joint test of the simultaneous null; it is not equivalent to conducting two separate coefficient tests.

Recall the general setup $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y}$ and $\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$.

A central specification issue in multiple linear regression is multicollinearity.

We assumed the columns of $\mathbf{X}$ were linearly independent, implying that $(\mathbf{X}^\top\mathbf{X})^{-1}$ exists; then $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y}$, where

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}_{n\times(p+1)} = \begin{bmatrix} \boldsymbol{x}_0 & \dots & \boldsymbol{x}_p \end{bmatrix}.$$

There are two cases of multicollinearity:

1) **Exact linear dependence.** Suppose that at least one of the $\boldsymbol{x}_j$ is a linear combination of the others. Then $|\mathbf{X}^\top\mathbf{X}| = 0$, or $\text{Rank}(\mathbf{X}) < p+1$. Then $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y}$ does not exist. For instance, suppose $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ and $x_1 = 5 + 3x_2$, so the predictors are perfectly linearly dependent. Then

$$y = \beta_0 + \beta_1(5 + 3x_2) + \beta_2 x_2 + \varepsilon \implies y = (\beta_0 + 5\beta_1) + (3\beta_1 + \beta_2)x_2 + \varepsilon.$$

   Renaming those two terms gives $y = \beta_0^* + \beta_1^* x_2 + \varepsilon$. In that case, drop x_1 if x_2 is included, since x_1 is redundant.

2) **Near linear dependence.** Suppose there exists a near (but not perfect) linear relationship between an $\boldsymbol{x}_j$ and the other $\boldsymbol{x}$. Then $(\mathbf{X}^\top\mathbf{X})^{-1}$ still exists, but $\mathbf{X}^\top\mathbf{X}$ has a very small eigenvalue and is ill-conditioned. Its inverse consequently has a very large eigenvalue in the corresponding direction.

   This has several consequences:

   (a) The computation of $\hat{\boldsymbol{\beta}}$ can become numerically unstable.

   (b) The signs of the estimated coefficients $\hat{\beta}_j$ can be misleading.

   (c) The estimated coefficients can be highly sensitive to small changes in the data.

   (d) Implausible coefficient values or magnitudes can appear.

   (e) Since $\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$, variance estimates tend to be inflated; important predictors may appear insignificant, and confidence intervals can become too wide to be useful for interpretation.

Common remedies for multicollinearity include the following.

1) Check pairwise correlation among predictors. In R, `cov(X)` gives the covariance matrix, `cor(X)` gives the correlation matrix, and `plot(X)` gives a pairwise scatterplot. If two x_j are highly correlated, we may remove one. Low pairwise correlation does not rule out a multivariable linear dependence.

2) A better measure is the **variance inflation factor (VIF)**.

   a) Treat x_j as the response.

   b) Regress x_j on the other predictors. Model is $x_j = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_{j-1}x_{j-1} + \alpha_{j+1}x_{j+1} + \cdots + \alpha_p x_p + \varepsilon^*$.

c) Denote the model coefficient of determination as R_j^2. Compute $\text{VIF}_j = 1/(1 - R_j^2)$, the variance inflation factor.

d) Calculate VIF_j for $j = 1, 2, \ldots, p$.

As a guideline,

$$\max_{1 \le j \le p} \{\text{VIF}_j\} > 10$$

is often taken as evidence of multicollinearity, although some references use 5 instead.

The variance estimate $\widehat{\text{Var}}(\hat{\beta}_j)$ is proportional to $1/(1 - R_j^2) = \text{VIF}_j$, which explains the name. If $R_j^2 = 0$, then the centered x_j column is orthogonal to the span of the other centered predictors and $\text{VIF}_j = 1$. If $R_j^2 > 0$, then $\text{VIF}_j > 1$, so the variance estimate of $\hat{\beta}_j$ is inflated.

Example 11.2 The following simulated example contrasts perfect and near-perfect linear dependence. In the first fit, $x_2 = 5 + 3x_1$, so R reports the coefficient of x_2 as undefined. The second fit adds only a tiny amount of noise to that relationship.

```
set.seed(1000000)

# Perfect linear dependence: R cannot estimate both slopes.
x1 <- 1:10
x2 <- 5 + 3 * x1
y <- 4 + x1 + rnorm(10, 0, 0.5)
perfect_fit <- lm(y ~ x1 + x2)
coef(summary(perfect_fit))

# Near linear dependence: the fit is strong but the slopes are unstable.
x1 <- 1:10
x2 <- 5 + 3 * x1 + rnorm(10, 0, 0.01)
y <- 4 + x1 + rnorm(10, 0, 0.5)
near_fit <- lm(y ~ x1 + x2)
reduced_fit <- lm(y ~ x2)
vif_x1 <- 1 / (1 - summary(lm(x1 ~ x2))$r.squared)

cor(x1, x2)
vif_x1
coef(summary(near_fit))
coef(summary(reduced_fit))
```

In the near-dependence fit, the predictor correlation is 0.9999995 and the variance inflation factor is approximately 967,749. Although the overall fit has adjusted $R^2 = 0.983$, the two fitted slopes are -15.87 and 5.63, with standard errors 44.91 and 14.96; neither is individually significant. After dropping x_1, the coefficient of x_2 is 0.3451 with standard error 0.0144. The example shows why a high R^2 does not prevent individual coefficients from becoming unstable under multicollinearity.

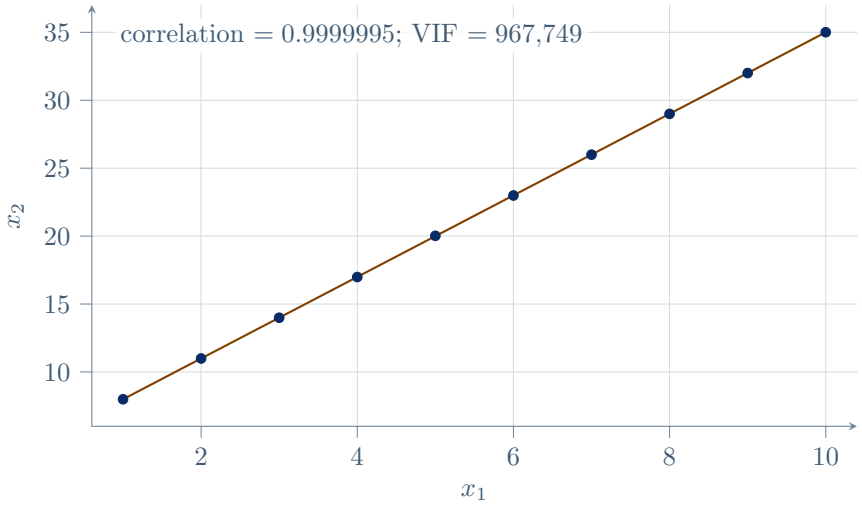


Figure 8

Figure 8 shows the near-perfect linear relationship from Example 11.2. The points are visually indistinguishable from a line even though the dependence is not exact.

Lecture 12 — Friday, June 20, 2014

Dummy Variables for Categorical Predictors

We can extend the linear regression model to include **categorical variables** (or **factors**) through **indicator variables** (or **dummy variables**).

Example 12.1 Data:

$$\text{make} \leftarrow \text{categorical } x_2,$$
$$\text{engine size} \leftarrow \text{predictor } x_1, \qquad \text{and} \qquad \text{fuel consumption} \leftarrow \text{response } y.$$

Suppose $n = 10$. BMW observations are $y_1, \ldots, y_4$, and AUDI observations are $y_5, \ldots, y_{10}$. Let

$$x_{i2} = \begin{cases} 0, & i = 1, 2, 3, 4 \\ 1, & i = 5, \ldots, 10. \end{cases}$$

First assume make has an effect on y and that the effect is the same regardless of engine size.

$$\mathbb{E}[y_i] = \begin{cases} \beta_0 + \beta_1 x_{i1}, & i = 1, \ldots, 4 \\ \beta_0 + \beta_1 x_{i1} + \beta_2, & i = 5, \ldots, 10 \end{cases} \implies \mathbb{E}[y_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

Model $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$. In matrix form, $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{X} = \left[\begin{array}{ccc} 1 & x_{11} & 0 \\ \vdots & \vdots & \vdots \\ 1 & x_{41} & 0 \\ \hline 1 & x_{51} & 1 \\ \vdots & \vdots & \vdots \\ 1 & x_{10,1} & 1 \end{array}\right]$.

For instance, we may test $H_0 : \beta_2 = 0$ to determine whether make has an effect on y.

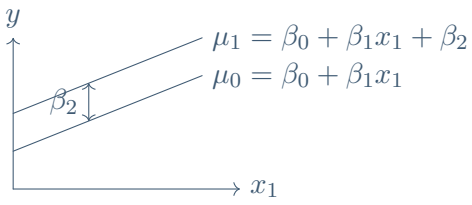


FIGURE 9

Figure 9 shows the parallel fitted lines produced by a make indicator effect with no interaction.

Second, suppose make again has an effect on y, but the effect changes with engine size; this

is called an **interaction**.

$$\mathbb{E}[y_i] = \begin{cases} \beta_0 + \beta_1 x_{i1}, & i = 1, \ldots, 4 \\ \beta_0 + (\beta_1 + \beta_3) x_{i1} + \beta_2, & i = 5, \ldots, 10 \end{cases} \implies \mathbb{E}[y_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2}$$

Here $\mathbf{X} = \left[\begin{array}{cccc} 1 & x_{11} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{41} & 0 & 0 \\ \hline 1 & x_{51} & 1 & x_{51} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{10,1} & 1 & x_{10,1} \end{array}\right]$.

For instance, we may test $H_0 : \beta_3 = 0$ to determine whether the effect of engine size on y depends on make.

Example 12.2 Data:

$$\text{diet} \leftarrow \text{categorical predictor } x \qquad \text{and} \qquad \text{weight} \leftarrow \text{response } y.$$

Suppose $n = 10$. Diet 1 observations are $y_1, \ldots, y_3$, Diet 2 observations are $y_4, \ldots, y_6$, and Diet 3 observations are $y_7, \ldots, y_{10}$.

Does diet affect weight gained? Let μ_j denote the mean response under Diet j, for $j = 1, 2, 3$. We can test

$$H_0 : \mu_1 = \mu_2 = \mu_3.$$

One formulation lets $x_{ij} = \begin{cases} 1, & i \in \text{Diet } j \\ 0, & \text{otherwise} \end{cases}$, etc. We remove the intercept β_0 because $\mathbf{X}$ does not have linearly independent columns when all group indicators and an intercept are included. This avoids multicollinearity. Then

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i,$$

with

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ \hline 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ \hline 0 & 0 & 1 \\ \vdots & \vdots & \vdots \end{bmatrix}.$$

If an intercept were included, $\mathbf{X}$ would contain a column of ones that is the sum of the current indicator columns, so it would be redundant.

A more common formulation is $y = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ \hline 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ \hline 1 & 0 & 0 \\ \vdots & \vdots & \vdots \end{bmatrix}.$$

Then

$$\mu_1 = \beta_0 + \beta_1, \qquad \mu_2 = \beta_0 + \beta_2, \qquad \text{and} \qquad \mu_3 = \beta_0.$$

Diet 3 is the base case in this formulation. β_0 is the average weight gained under Diet 3, $\beta_1 = \mu_1 - \mu_3$ is the excess average weight gained under Diet 1 relative to Diet 3, and β_2 is interpreted similarly. Testing

$$H_0 : \mu_1 = \mu_2 = \mu_3 \qquad \text{is equivalent to} \qquad H_0 : \beta_1 = \beta_2 = 0$$

in the ANOVA F-test reported by R.

Example 12.3 Consider forced expiratory volume (FEV), a measure of pulmonary function, for 654 children and young adults. The response is FEV in litres; age and height are quantitative predictors; sex and smoking status are binary predictors. The data come from the Journal of Statistics Education data archive.

The following code compares parallel age trends for smokers and nonsmokers with a model that permits the age slope to depend on smoking status.

```r
fev_data <- read.table("data/fev.dat.txt")
names(fev_data) <- c("age", "fev", "height", "sex", "smoke")
fev_data$smoke <- factor(fev_data$smoke,
                         levels = c(0, 1),
                         labels = c("Nonsmoker", "Smoker"))

parallel_fit <- lm(fev ~ smoke + age, data = fev_data)
interaction_fit <- lm(fev ~ smoke * age, data = fev_data)
coef(summary(parallel_fit))
coef(summary(interaction_fit))
```

Without the interaction, the fitted smoking coefficient is -0.2090, with $p = 0.0099$. Once the interaction is included, the fitted lines are

$$\widehat{\text{FEV}} = \begin{cases} 0.2534 + 0.2426\,\text{age}, & \text{nonsmoker}, \\ 2.1970 + 0.0799\,\text{age}, & \text{smoker}. \end{cases}$$

The interaction estimate is -0.1627, with $p = 1.65 \times 10^{-7}$, and adjusted R^2 increases from 0.5753 to 0.5922. The fitted association between age and FEV therefore differs by smoking status in these observational data; it should not be given a causal interpretation.

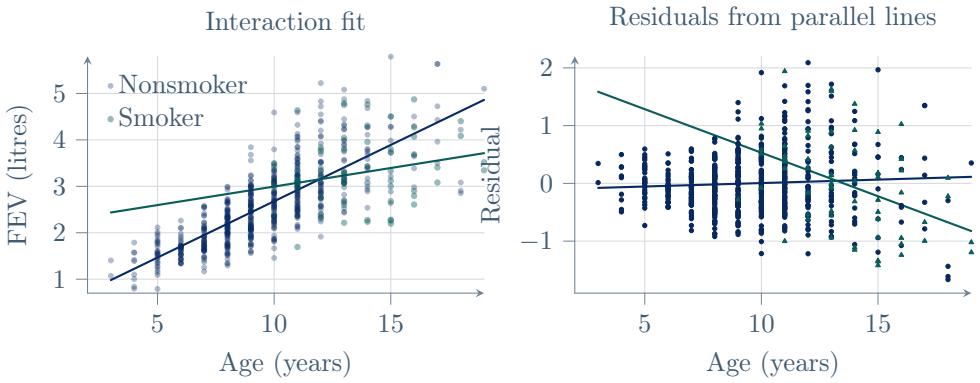


Figure 10

Figure 10 shows FEV against age with fitted lines from the interaction model between age and smoking, together with residuals from the parallel lines model. The opposing residual trends reveal the missing interaction; the smokers in this observational sample are also concentrated at older ages, so the two groups have very different age distributions.

Lecture 13 — Wednesday, June 25, 2014

4. Residual Analysis

Residual Diagnostics and Studentized Residuals

We now turn to residual analysis for the multiple regression model. The true error is $\varepsilon_i = y_i - \mu_i$, where $\mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$. The model assumption is that $\varepsilon_1, \ldots, \varepsilon_n$ are independent and each follows $\mathcal{N}(0, \sigma^2)$. The estimated errors (i.e., residuals) are $e_i = y_i - \hat{\mu}_i$. Some key properties follow:

Proposition 13.1

1) $$\sum_{i=1}^{n} e_i = 0,$$
and therefore $\bar{e} = 0$.

2) $$\sum_{i=1}^{n} e_i x_{ik} = 0$$
for $k = 1, \ldots, p$. Thus the residual–predictor sample cross-product is zero; the corresponding sample correlation is zero whenever both sample variances are positive.

3) $$\sum_{i=1}^{n} e_i \hat{\mu}_i = 0,$$
so the residual and fitted-value vectors are orthogonal. Their sample correlation is zero whenever both sample variances are positive.

4) $$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \mathbf{H}\boldsymbol{y} = (\mathbf{I} - \mathbf{H})\boldsymbol{y},$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$. Therefore

$$\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}, (\mathbf{I} - \mathbf{H})\sigma^2),$$

so $e_i \sim \mathcal{N}(0, (1 - h_{ii})\sigma^2)$ and $\text{Cov}(e_i, e_j) = -h_{ij}\sigma^2$ for $i \neq j$.

5) The studentized residuals are $d_i = e_i/(\hat{\sigma}\sqrt{1 - h_{ii}})$ for $i = 1, \ldots, n$. These do not follow a t-distribution because the numerator and denominator are not independent. In practice, the d_i are often treated as approximately $\mathcal{N}(0, 1)$.

Proof. For (1)), the normal equations in multiple regression give

$$\mathbf{X}^\top\boldsymbol{e} = \mathbf{0}.$$

Because the first column of $\mathbf{X}$ is $\mathbf{1}$, this implies

$$\mathbf{1}^\top\boldsymbol{e} = \sum_{i=1}^{n} e_i = 0,$$

and hence $\bar{e} = 0$.

For (2)), the kth predictor column of $\mathbf{X}$ is $(x_{1k}, \ldots, x_{nk})^\top$, so the kth component of $\mathbf{X}^\top\boldsymbol{e} = \mathbf{0}$ gives

$$\sum_{i=1}^{n} e_i x_{ik} = 0, \quad k = 1, \ldots, p.$$

For (3)), since $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, we have

$$\hat{\boldsymbol{\mu}}^\top\boldsymbol{e} = (\mathbf{X}\hat{\boldsymbol{\beta}})^\top\boldsymbol{e} = \hat{\boldsymbol{\beta}}^\top\mathbf{X}^\top\boldsymbol{e} = 0.$$

Thus $\sum_{i=1}^{n} e_i\hat{\mu}_i = 0$.

For (4)),

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \mathbf{H}\boldsymbol{y} = (\mathbf{I} - \mathbf{H})\boldsymbol{y}.$$

Under the normal model, $\boldsymbol{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, so

$$\boldsymbol{e} \sim \mathcal{N}\Big((\mathbf{I} - \mathbf{H})\mathbf{X}\boldsymbol{\beta},\ (\mathbf{I} - \mathbf{H})\sigma^2\mathbf{I}(\mathbf{I} - \mathbf{H})^\top\Big) = \mathcal{N}(\mathbf{0}, (\mathbf{I} - \mathbf{H})\sigma^2),$$

since $(\mathbf{I}-\mathbf{H})\mathbf{X}=\mathbf{0}$ and $\mathbf{I}-\mathbf{H}$ is symmetric idempotent. Therefore,

$$\mathrm{Var}(e_i)=\sigma^2(1-h_{ii}) \qquad \text{and} \qquad \mathrm{Cov}(e_i,e_j)=-\sigma^2 h_{ij}, \quad i\neq j.$$

For (5)),

$$d_i=\frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

does not have an exact t-distribution because e_i and $\hat{\sigma}$ are not independent ($\hat{\sigma}$ is computed from the residual vector containing e_i). In practice, d_i is often treated as approximately $\mathcal{N}(0,1)$. □

Useful diagnostics include plotting (i) the d_i, (ii) the e_i versus each predictor, and (iii) the e_i versus the $\hat{\mu}_i$. If the assumptions hold, these plots should show random scatter.

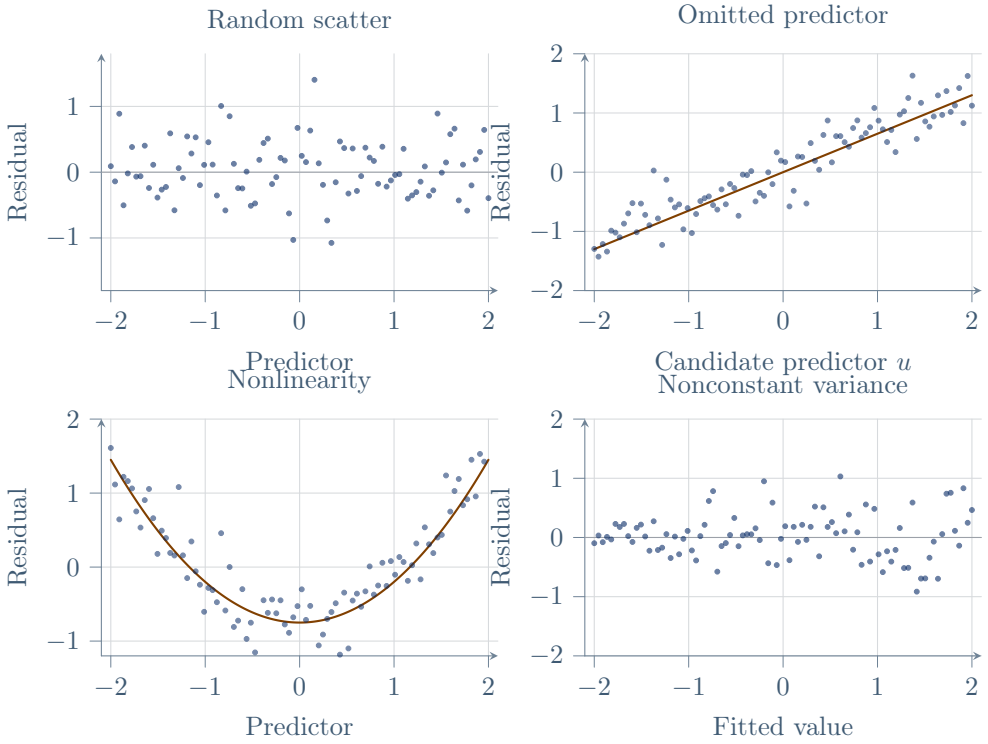


Figure 11

Figure 11 collects canonical residual patterns. Random scatter is compatible with the model assumptions, while systematic association, curvature, and changing spread point to an omitted predictor, a nonlinear mean, and heteroscedasticity, respectively. The coordinates are generated in R and rendered natively with PGFPlots.

To assess normality, sort the studentized residuals as $d_{(1)} \le \cdots \le d_{(n)}$ and set

$$p_i = \frac{i - 1/2}{n}, \quad i = 1, \ldots, n.$$

If the residual distribution is approximately normal, then

$$\mathbb{E}[d_{(i)}] \approx \Phi^{-1}(p_i),$$

so a plot of $d_{(i)}$ against $\Phi^{-1}(p_i)$ should be approximately linear. Systematic departures from the reference line reveal tail or skewness behaviour.

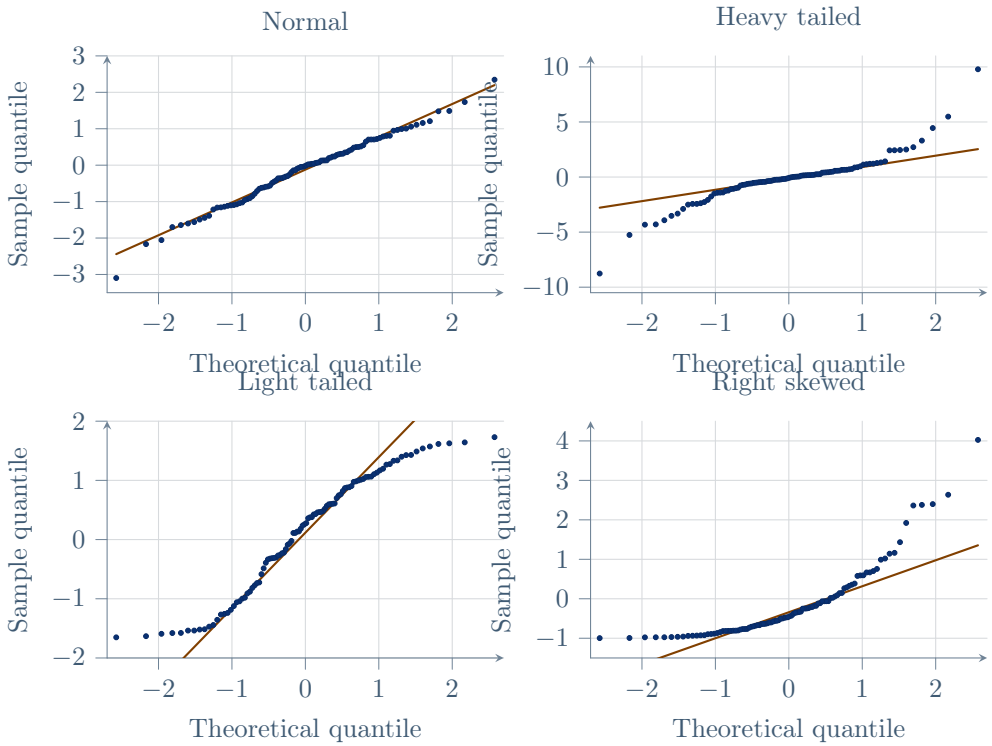


FIGURE 12

Figure 12 compares normal Q–Q patterns for normal, heavy-tailed, light-tailed, and right-skewed samples. The samples are generated in R, and the Q–Q coordinates are rendered natively with PGFPlots.

An **added-variable plot** checks a candidate predictor u after accounting for the predictors already in the model. Let $\boldsymbol{e}$ be the residual vector from regressing $\boldsymbol{y}$ on $x_1, \ldots, x_p$, and let $\boldsymbol{e}_u$ be the residual vector from regressing u on those same predictors. A plot of $\boldsymbol{e}$ against $\boldsymbol{e}_u$ shows the part of the response–candidate relationship not explained by the current model. Random scatter gives little

reason to add u, whereas systematic association supports its inclusion.

Curvature may be handled through **polynomial regression**. For one predictor, a degree-r model has mean

$$\mu = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_r x^r,$$

which is an ordinary multiple linear regression with predictors $x, x^2, \ldots, x^r$. The degree should be kept as small as the diagnostics permit, and hierarchical polynomial models retain every lower-order term whenever a higher-order term is included. With several predictors, inspect residuals against each predictor and against the fitted values to decide which variables may require polynomial terms.

Example 13.2 The FEV data from Example 12.3 also provide a useful residual diagnostics example. We first fit an additive model with age, height, sex, and smoking status. The curvature in the residuals against height suggests adding a quadratic height term.

```
fev_data <- read.table("data/fev.dat.txt")
names(fev_data) <- c("age", "fev", "height", "sex", "smoke")
fev_data$sex <- factor(fev_data$sex)
fev_data$smoke <- factor(fev_data$smoke)

linear_fit <- lm(fev ~ age + height + sex + smoke, data = fev_data)
quadratic_fit <- lm(fev ~ age + height + sex + smoke + I(height^2),
                    data = fev_data)
cubic_fit <- lm(fev ~ age + height + sex + smoke + I(height^2) + I(height^3),
                data = fev_data)

c(linear_adjusted_r2 = summary(linear_fit)$adj.r.squared,
  quadratic_adjusted_r2 = summary(quadratic_fit)$adj.r.squared,
  cubic_adjusted_r2 = summary(cubic_fit)$adj.r.squared)
coef(summary(quadratic_fit))["I(height^2)", ]
coef(summary(cubic_fit))["I(height^3)", ]
```

The coefficient of `I(height^2)` is 0.003125, with $p = 7.35 \times 10^{-14}$, and adjusted R^2 increases from 0.7740 to 0.7924. Adding a cubic height term provides a useful hierarchy check: its coefficient has $p = 0.7372$, while adjusted R^2 falls slightly to 0.7921, so the quadratic model is retained. It removes much of the systematic curvature, but the residual spread still increases with the fitted mean and the Q–Q plot still shows heavy tails. Adding a term can repair one aspect of model specification without repairing every model assumption.

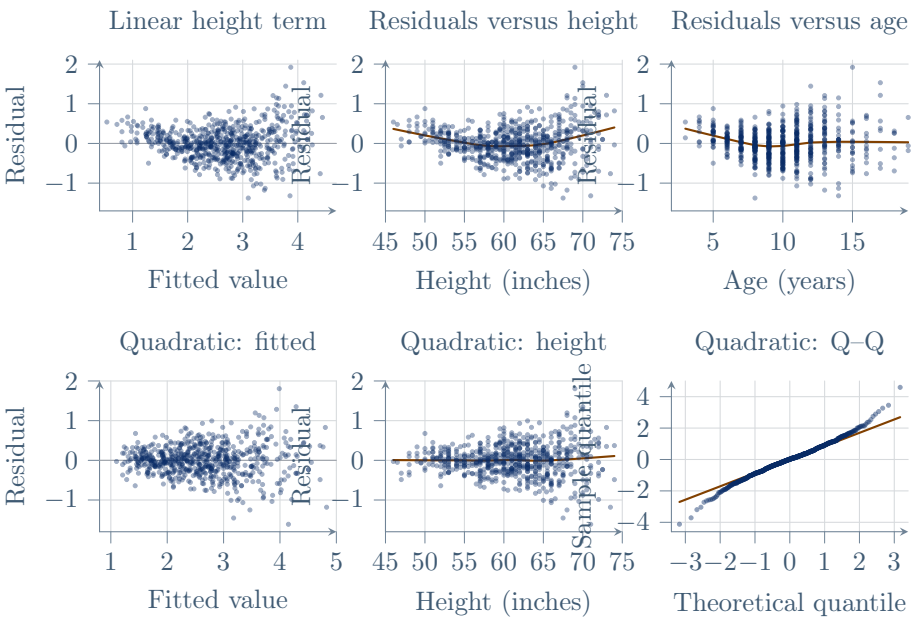


Figure 13

Figure 13 shows the FEV residual diagnostics before and after adding a quadratic height term, including residuals against age and, after the quadratic term is added, against height. The diagnostic values are computed by the R code in Example 13.2 and rendered natively with PGFPlots.

Lecture 14 — Friday, June 27, 2014

Variance Stabilizing Transformations in Regression

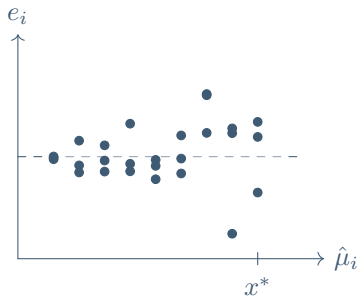


Figure 14

Suppose that after fitting the model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, the residual plot in Figure 14 is produced. If we want to predict the value of y at $x = x^*$, would the width of the prediction interval be overstated or understated?

It would be *understated*. The constant-variance fit pools information across the full predictor range, so its estimate need not equal the larger local error variance near x^*. More generally, a homoscedastic prediction interval is too narrow in regions where the true variance exceeds the fitted common variance and too wide where it is smaller.

More generally, consider the model $y_i = \mu_i + \varepsilon_i$ where $\mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$, and assume more realistically that $\mathrm{Var}(y_i) = \sigma^2[h(\mu_i)]^2$ — that is, the variance of y_i is a function of μ_i. Instead of using y_i, we use a transformation $g(y_i)$ so that $\mathrm{Var}(g(y_i))$ is reasonably stable. A Taylor expansion gives $g(y) \approx g(\mu) + g'(\mu)(y - \mu)$. Then $\mathrm{Var}(g(y)) \approx [g'(\mu)]^2 \sigma^2 [h(\mu)]^2$. We want $g'(\mu) \cdot h(\mu) = c \in \mathbb{R}$, which implies $g'(\mu) = c/(h(\mu))$ and therefore

$$g(\mu) = \int \frac{c}{h(\mu)} \, \mathrm{d}\mu.$$

Such a function g is called a **variance stabilizing transformation**. For example:

1. $$h(\mu) = \mu \implies g(\mu) = \int \frac{c}{\mu} \, \mathrm{d}\mu = c \log(\mu).$$

   Here we could try $g(y) = \log(y)$.

2. $$h(\mu) = \sqrt{\mu} \implies g(\mu) = \int \frac{c}{\sqrt{\mu}} \, \mathrm{d}\mu = 2c\sqrt{\mu}.$$

   Here we could try $g(y) = \sqrt{y}$.

A widely used family of such transformations is given by the Box–Cox family, which requires $y_i > 0$. Here the model is $y_i = \mu_i + \varepsilon_i$. The **Box–Cox transformations** are a family of power transformations

$$g(y_i) = (y_i^\lambda - 1)/\lambda \quad \text{for} \lambda \neq 0.$$

Here λ is chosen such that $\mathrm{Var}(g(y_i))$ is approximately constant. Observe the following:

1. $\lambda = 1$ gives $g(y) = y - 1$, which is equivalent to no transformation in a model with an intercept.
2. $\lambda = 1/2$ gives $g(y) = 2(\sqrt{y} - 1)$, which is equivalent up to shift and scale to a square root

transformation.

3. $\lambda = 0$ implies $\lim_{\lambda \to 0} g(y_i) = \log(y_i)$ by L'Hôpital's rule.

We would like to estimate the new parameter λ by maximum likelihood. Assuming $g(y_i) \sim \mathcal{N}(\mu_{i,\lambda}, \sigma^2_\lambda)$, the log-likelihood is

$$\ell(\lambda) = -\frac{1}{2\sigma^2_\lambda} \sum_{i=1}^{n} (g(y_i) - \mu_{i,\lambda})^2 - \frac{n}{2} \log(\sigma^2_\lambda) + (\lambda - 1) \sum_{i=1}^{n} \log(y_i).$$

We want to maximize this with respect to $\lambda, \boldsymbol{\beta}, \sigma^2_\lambda$. Unfortunately this is not very easy to do directly in practice, so instead we use the **profile likelihood** as follows:

1. Consider a sequence of λ's, for example $\{-2, -1.9, \ldots, 1.9, 2\}$.
2. For each λ, using $g(y_i)$ as the response, find the least squares estimates of $\boldsymbol{\beta}_\lambda$ and σ_λ. Also compute $\ell(\lambda)$.
3. Select the value of λ which gives the largest $\ell(\lambda)$; denote it by $\hat{\lambda}$.

In R, we can load `MASS` and run `boxcox(mod)`.

Example 14.1 A Box–Cox transformation is most useful when it is illustrated with reproducible data. Applying `boxcox()` to the positive FEV response gives the following example:

```
fev_data <- read.table("data/fev.dat.txt")
names(fev_data) <- c("age", "fev", "height", "sex", "smoke")
fev_data$sex <- factor(fev_data$sex)
fev_data$smoke <- factor(fev_data$smoke)

fev_fit <- lm(fev ~ age + height + sex + smoke + I(height^2),
              data = fev_data)
lambda_grid <- seq(-1, 1, by = 0.01)
boxcox_values <- MASS::boxcox(fev_fit, lambda = lambda_grid, plotit = FALSE)
lambda_hat <- boxcox_values$x[which.max(boxcox_values$y)]
cutoff <- max(boxcox_values$y) - qchisq(0.95, df = 1) / 2
lambda_interval <- range(boxcox_values$x[boxcox_values$y >= cutoff])
c(lambda_hat = lambda_hat,
  lower_95 = lambda_interval[1], upper_95 = lambda_interval[2])
```

The profile likelihood is maximized at $\hat{\lambda} = 0.13$, with an approximate 95% profile likelihood interval of $(-0.02, 0.28)$. Since $\lambda = 0$ lies in this interval, a log transformation of FEV is a reasonable simple choice. The purpose of the plot is to identify a small collection of plausible transformations; it does not make the final model checking step unnecessary.

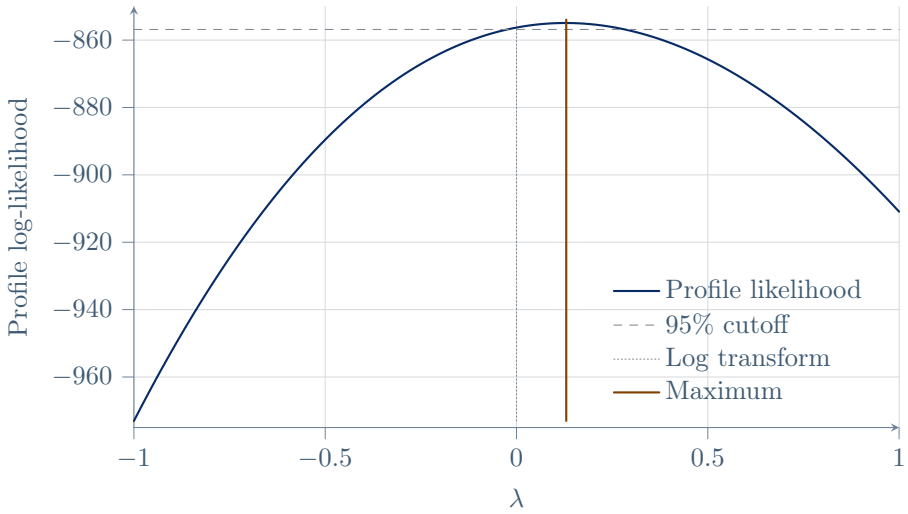


Figure 15

Figure 15 shows the Box–Cox profile log-likelihood for the FEV model. The dotted line marks the log transformation, and the solid vertical line marks the maximizing value.

Lecture 15 — Wednesday, July 02, 2014

Weighted Least Squares

Another remedy for heteroscedasticity is **weighted least squares**. Again consider the model $y_i = \mu_i + \varepsilon_i$ where the ε_i are independent with distributions $\mathcal{N}(0, \sigma^2 v_i^2)$, the known relative standard deviations satisfy $v_i > 0$, and $\mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$. Since $\text{Var}(\varepsilon_i) = \sigma^2 v_i^2$, we have $\text{Var}(\varepsilon_i / v_i) = \sigma^2$, which is constant. Thus we rewrite the model as

$$\frac{y_i}{v_i} = \frac{\beta_0}{v_i} + \beta_1 \left(\frac{x_{i1}}{v_i}\right) + \cdots + \beta_p \left(\frac{x_{ip}}{v_i}\right) + \frac{\varepsilon_i}{v_i}.$$

Let $x^*_{i0} = 1/v_i$ and $x^*_{ik} = x_{ik}/v_i$ for $k = 1, \ldots, p$. Then

$$y^*_i = \frac{y_i}{v_i} = \beta_0 x^*_{i0} + \beta_1 x^*_{i1} + \cdots + \beta_p x^*_{ip} + \varepsilon^*_i,$$

where the transformed errors are independent and each follows $\mathcal{N}(0, \sigma^2)$.

Weighted least squares estimates are obtained by minimizing the function

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y^*_i - \beta_0 x^*_{i0} - \cdots - \beta_p x^*_{ip})^2 = \sum_{i=1}^{n} \frac{1}{v_i^2} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2.$$

In matrix form, let $\mathbf{W} = \operatorname{diag}(w_1, \ldots, w_n)$ where $w_i = 1/v_i^2$. Then $S(\boldsymbol{\beta}) = (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{W} (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})$, which is another quadratic form.

Key results are as follows:

Proposition 15.1

1. The weighted least squares estimator of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}_w = (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \boldsymbol{y}$.
2. $\mathbb{E}\left[\hat{\boldsymbol{\beta}}_w\right] = \boldsymbol{\beta}$, so it is unbiased.
3. $\operatorname{Var}(\hat{\boldsymbol{\beta}}_w) = \sigma^2 (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1}$.
4. $$\hat{\sigma}^2 = \frac{1}{n - p - 1} \sum_{i=1}^{n} w_i e_i^2$$
   where $e_i = y_i - \hat{\beta}_{0w} - \hat{\beta}_{1w} x_{i1} - \cdots - \hat{\beta}_{pw} x_{ip}$.

Proof. For (1), write

$$S(\boldsymbol{\beta}) = (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{W} (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}).$$

Differentiating with respect to $\boldsymbol{\beta}$ gives

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^\top \mathbf{W} \boldsymbol{y} + 2\mathbf{X}^\top \mathbf{W} \mathbf{X} \boldsymbol{\beta}.$$

Setting this equal to $\mathbf{0}$ yields the weighted normal equations

$$\mathbf{X}^\top \mathbf{W}\mathbf{X}\,\hat{\boldsymbol{\beta}}_w = \mathbf{X}^\top \mathbf{W}\boldsymbol{y},$$

so (assuming $\mathbf{X}^\top\mathbf{W}\mathbf{X}$ is invertible)

$$\hat{\boldsymbol{\beta}}_w = (\mathbf{X}^\top \mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{W}\boldsymbol{y}.$$

For (2),

$$\mathbb{E}[\hat{\boldsymbol{\beta}}_w] = (\mathbf{X}^\top \mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{W}\,\mathbb{E}[\boldsymbol{y}] = (\mathbf{X}^\top \mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{W}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

Hence $\hat{\boldsymbol{\beta}}_w$ is unbiased.

For (3), since

$$\mathrm{Var}(\boldsymbol{y}) = \sigma^2\,\mathrm{diag}(v_1^2, \ldots, v_n^2) = \sigma^2\mathbf{W}^{-1},$$

we have

$$\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\beta}}_w) &= \mathrm{Var}\Big((\mathbf{X}^\top \mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{W}\boldsymbol{y}\Big) \\
&= (\mathbf{X}^\top \mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{W}\,\mathrm{Var}(\boldsymbol{y})\mathbf{W}\mathbf{X}(\mathbf{X}^\top \mathbf{W}\mathbf{X})^{-1} \\
&= (\mathbf{X}^\top \mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{W}(\sigma^2\mathbf{W}^{-1})\mathbf{W}\mathbf{X}(\mathbf{X}^\top \mathbf{W}\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}^\top \mathbf{W}\mathbf{X})^{-1}.
\end{aligned}$$

For (4), define transformed variables

$$\boldsymbol{y}^* = \mathbf{W}^{1/2}\boldsymbol{y} \qquad \text{and} \qquad \mathbf{X}^* = \mathbf{W}^{1/2}\mathbf{X}.$$

Then weighted least squares is ordinary least squares on

$$\boldsymbol{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^* \qquad \text{and} \qquad \mathrm{Var}(\boldsymbol{\varepsilon}^*) = \sigma^2\mathbf{I}.$$

If $\boldsymbol{e}^* = \boldsymbol{y}^* - \mathbf{X}^*\hat{\boldsymbol{\beta}}_w$, then

$$\boldsymbol{e}^{*\top}\boldsymbol{e}^* = \boldsymbol{e}^\top \mathbf{W}\boldsymbol{e} = \sum_{i=1}^n w_i e_i^2.$$

By the usual OLS result in the transformed model,

$$\mathbb{E}[\boldsymbol{e}^{*\top}\boldsymbol{e}^*] = (n - p - 1)\sigma^2,$$

so

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^n w_i e_i^2$$

is an unbiased estimator of σ^2.

□

For diagnostics, plot $e_i/v_i = e_i^*$ versus x_{ik} or $\hat{\mu}_i$ to check whether the variance has stabilized.

Estimating v_i is difficult in practice. We can start with a plot of e_i versus x_{ik}, try a form such as $v_i = x_{ik}^{\gamma}$ for some $k = 1, \ldots, p$ and $\gamma \in \mathbb{R}$, and then recheck plots of e_i/v_i versus $\hat{\mu}_i$ or x_{ik} until the variance pattern is approximately constant.

Why use weights? The homoscedastic linear model assigns every error the same variance σ^2, whereas weighted least squares allows $\text{Var}(\varepsilon_i) = \sigma^2 v_i^2$ and accounts for the known relative variances through the weights.

To make $\text{Var}(\varepsilon_i)$ constant, use $y_i/v_i = \mu_i/v_i + \varepsilon_i/v_i$, where $\varepsilon_i/v_i \sim \mathcal{N}(0, \sigma^2)$. This transformed model is equivalent to the weighted formulation.

Proposition 15.2 The weighted least squares criterion can be written as

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(\frac{y_i}{v_i} - \frac{\mu_i}{v_i} \right)^2,$$

and the corresponding variance estimator is

$$\hat{\sigma}^2 = \frac{S(\hat{\boldsymbol{\beta}}_w)}{n-p-1} = \frac{\sum_{i=1}^{n} (e_i/v_i)^2}{n-p-1}.$$

These weighted residuals should have constant variance. That is why we plot e_i/v_i versus $\hat{\mu}_i$ or x_{ik}, rather than plotting e_i alone.

Outliers (Extreme Observations)

We also need to assess the effect of extreme observations. There are essentially two cases:

1. If an apparent outlier is a recording error, we should correct it from the original source when possible; otherwise, we should treat it as missing rather than inventing a replacement value.
2. If an outlier is a valid observation, it may reveal a missing predictor or a poorly specified

model. We should compare the fit with and without the point. If the conclusions are essentially unchanged, we keep it. If they change substantially, the point is **influential**; any decision to exclude it needs a substantive justification, since removal can redefine the target population.

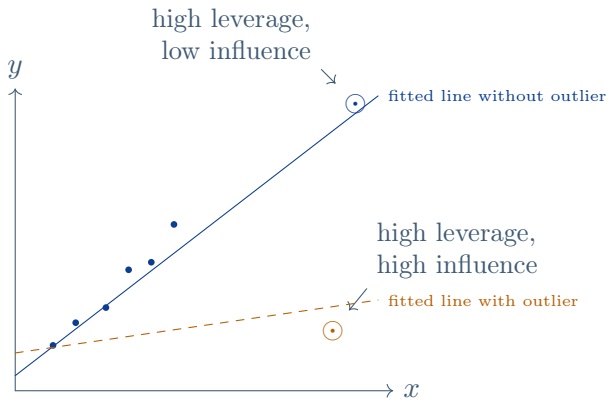


FIGURE 16

Figure 16 contrasts a high-leverage but noninfluential point with an influential point that substantially changes the fitted line.

The effect can be severe because least squares minimizes squared errors, so large residuals can dominate the fit.

In R, `car::outlierTest(mod)` implements a Bonferroni-adjusted test based on deleted studentized residuals.

Lecture 16 — Friday, July 04, 2014

Recall that

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 = \sum_{i=1}^{n} e_i^2$$

is minimized in ordinary least squares. Because residuals are squared, outliers can have a large effect on the fit. An alternative criterion is minimizing

$$\sum_{i=1}^{n} |e_i|,$$

which is less sensitive to extreme observations.

To detect outliers, start by constructing a residual plot. Using the studentized residuals $d_i = e_i/(\hat{\sigma}\sqrt{1-h_{ii}})$, check whether $d_1, \ldots, d_n$ look approximately $\mathcal{N}(0,1)$. Typically, most values should lie within 3 standard deviations, and roughly 95% should lie in $(-2, 2)$.

A standard formal procedure is based on the following result:

Proposition 16.1 Fix an observation i, assume $n > p+2$, $h_{ii} < 1$, and suppose the design matrix remains full rank after observation i is deleted. Refit the model with the remaining $n-1$ observations. Then

$$\hat{\sigma}^2_{(-i)} = \frac{1}{(n-1)-(p+1)} \sum_{\substack{j=1 \\ j\neq i}}^{n} e^2_{j(-i)} = \frac{1}{n-p-2} \sum_{\substack{j=1 \\ j\neq i}}^{n} e^2_{j(-i)},$$

and

$$(n-p-2)\frac{\hat{\sigma}^2_{(-i)}}{\sigma^2} \sim \chi^2(n-p-2).$$

The deleted residual test statistic is

$$t_i = \frac{e_i}{\hat{\sigma}_{(-i)}\sqrt{1-h_{ii}}} \sim t(n-p-2).$$

Proof. Delete observation i, and write the reduced model as

$$\boldsymbol{y}_{(-i)} = \mathbf{X}_{(-i)}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{(-i)}, \qquad \text{and} \qquad \boldsymbol{\varepsilon}_{(-i)} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_{n-1}).$$

By the usual normal theory result for multiple regression applied to this reduced sample, the residual variance estimator is

$$\hat{\sigma}^2_{(-i)} = \frac{\sum_{j\neq i} e^2_{j(-i)}}{(n-1)-(p+1)} = \frac{1}{n-p-2}\sum_{j\neq i} e^2_{j(-i)},$$

and

$$\frac{(n-p-2)\hat{\sigma}^2_{(-i)}}{\sigma^2} \sim \chi^2(n-p-2).$$

Now define the deleted residual

$$r_i \coloneqq y_i - \hat{y}_{i(-i)}, \qquad \text{and} \qquad \hat{y}_{i(-i)} = \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_{(-i)}.$$

Because y_i is independent of the reduced-sample data and $\hat{\boldsymbol{\beta}}_{(-i)}$ is a linear function of the reduced-sample response vector, r_i is normal with mean 0 and variance

$$\mathrm{Var}(r_i) = \frac{\sigma^2}{1 - h_{ii}}.$$

Using the standard leave-one-out identity

$$r_i = \frac{e_i}{1 - h_{ii}},$$

we obtain

$$\frac{e_i}{\sigma\sqrt{1 - h_{ii}}} \sim \mathcal{N}(0, 1).$$

Also, $\hat{\sigma}^2_{(-i)}$ is computed from the reduced sample and is independent of r_i, hence independent of $e_i/\sqrt{1 - h_{ii}}$. Therefore

$$t_i = \frac{e_i}{\hat{\sigma}_{(-i)}\sqrt{1 - h_{ii}}} = \frac{\dfrac{e_i}{\sigma\sqrt{1 - h_{ii}}}}{\sqrt{\dfrac{(n - p - 2)\hat{\sigma}^2_{(-i)}/\sigma^2}{n - p - 2}}} \sim t(n - p - 2).$$

□

To conduct the **outlier test** for observation i, we delete it, refit the model to the remaining $n - 1$ observations, and compute t_i as in Proposition 16.1. If $|t_i| > t_{\alpha/2}(n-p-2)$, we flag the observation as an outlier.

That pointwise rule does not control the chance of making at least one false discovery across all n observations. We can instead test each observation at level α/n; this is a **Bonferroni correction**.

1) For a single test ($n = 1$) with $\alpha = 0.05$, the false positive probability is

$$\mathbb{P}(\text{reject } H_0 \mid H_0 \text{ is true}) = \frac{\alpha}{1} = 5\%.$$

2) For $n > 1$ independent tests, each at nominal level $\alpha = 5\%$,

$$\mathbb{P}(\geq 1 \text{ false positive}) = 1 - [1 - \alpha]^n = 1 - 0.95^n \xrightarrow{n\to\infty} 1.$$

The Bonferroni correction does not require independence: by the union bound,

$$\mathbb{P}(\geq 1 \text{ false positive}) \leq \sum_{i=1}^{n} \frac{\alpha}{n} = \alpha.$$

Influential Cases

The main question here is whether an outlier is **influential**, meaning that removing it changes conclusions substantially.

Leverage measures how unusual an observation is in the predictor space. A large residual measures outlyingness in the response direction, while **influence** measures how much the fitted model changes when we remove the observation. A point is most likely to be influential when it combines substantial leverage with a large residual.

Lecture 17 — Wednesday, July 09, 2014

Leverage and Influential Observation Patterns

We next examine leverage more closely. A point has high leverage when it lies far from the bulk of the data in the x-direction. High-leverage points can strongly affect fitted coefficients and predictions.

Cases:

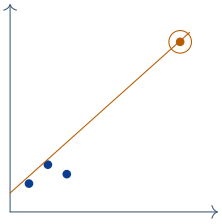
1. 

   Figure 17

   Figure 17 shows a point far from the other predictors that substantially changes the fitted line, so it has high leverage and high influence.

2.

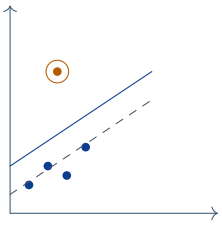


FIGURE 18

In Figure 18, leverage is low because the point lies near $\bar{x}$. Its influence is also low because fitted coefficients and predictions change only slightly.

3.

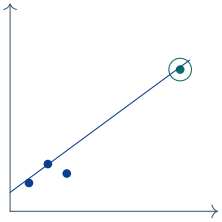


FIGURE 19

Figure 19 shows high leverage without high influence.

Overall, high leverage is usually a prerequisite for strong influence, but not every high-leverage point is influential.

To formalize leverage, recall the hat matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top = [h_{ij}]_{n\times n},$$

and $e_i \sim \mathcal{N}(0, \sigma^2(1 - h_{ii}))$.

If $h_{ii} \approx 1$, then $e_i \approx 0$, so $y_i - \hat{\mu}_i \approx 0$, and hence $y_i \approx \hat{\mu}_i$. In this case, the fitted line is pulled toward the ith observation. The quantity h_{ii} is the **leverage** of point i.

Proposition 17.1

1) $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top$ is a function of the x alone, so h_{ii} is also determined only by the x.

2) We have the bounds

$$\frac{1}{n} \leq h_{ii} \leq 1.$$

3)

$$\sum_{i=1}^{n} h_{ii} = p + 1.$$

Therefore, the average leverage is $\bar{h} = (p+1)/n$. A common guideline is that $h_{ii} > 2\bar{h}$ indicates a high-leverage point.

4) h_{ii} is smallest when x_i is near $\bar{x}$. In the SLR case,

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}},$$

which is minimized at $x_i = \bar{x}$.

Proof. For (1)),

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top$$

is determined entirely by $\mathbf{X}$, so each diagonal entry h_{ii} depends only on the x-values.

For (2)), from

$$\mathrm{Var}(\boldsymbol{e}) = \sigma^2(\mathbf{I} - \mathbf{H}),$$

we get

$$\mathrm{Var}(e_i) = \sigma^2(1 - h_{ii}) \geq 0,$$

so $h_{ii} \leq 1$. To show the lower bound when an intercept is present, write

$$h_{ii} = \frac{1}{n} + (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top (\mathbf{X}_c^\top \mathbf{X}_c)^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}),$$

where $\mathbf{X}_c$ is the centered predictor matrix (excluding the intercept). The quadratic term is non-negative, and hence $h_{ii} \geq 1/n$.

For (3)),

$$\sum_{i=1}^{n} h_{ii} = \mathrm{tr}(\mathbf{H}) = \mathrm{tr}\Big(\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top\Big) = \mathrm{tr}\Big((\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{X}\Big) = \mathrm{tr}\big(I_{(p+1)\times(p+1)}\big) = p + 1.$$

For (4)), in simple linear regression,

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}},$$

so h_{ii} is minimized when $(x_i - \bar{x})^2$ is minimized, namely at $x_i = \bar{x}$. □

To identify highly influential cases, we use the following diagnostics.

Proposition 17.2 The coefficient perturbation quadratic form satisfies

$$\frac{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top(\mathbf{X}^\top\mathbf{X})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})/(p+1)}{\hat{\sigma}^2} \sim F(p+1, n-p-1).$$

Proof. Under the normal multiple regression model,

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}\Big(\boldsymbol{\beta},\, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\Big).$$

Therefore,

$$\boldsymbol{Z} \coloneqq \frac{1}{\sigma}(\mathbf{X}^\top\mathbf{X})^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{p+1}),$$

which implies

$$\boldsymbol{Z}^\top\boldsymbol{Z} = \frac{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top(\mathbf{X}^\top\mathbf{X})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{\sigma^2} \sim \chi^2_{(p+1)}.$$

Also,

$$\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-p-1)}.$$

Under the normal model, $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are independent, so these two chi-square variables are independent. Hence

$$\frac{\dfrac{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top(\mathbf{X}^\top\mathbf{X})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{(p+1)\sigma^2}}{\dfrac{(n-p-1)\hat{\sigma}^2}{(n-p-1)\sigma^2}} \sim F(p+1, n-p-1),$$

which is exactly

$$\frac{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top(\mathbf{X}^\top\mathbf{X})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})/(p+1)}{\hat{\sigma}^2} \sim F(p+1, n-p-1).$$

□

Cook's distance for the ith observation is defined by

$$D_i = \frac{1}{(p+1)\hat{\sigma}^2}(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(-i)})^\top(\mathbf{X}^\top\mathbf{X})(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(-i)}),$$

where $\hat{\boldsymbol{\beta}}_{(-i)}$ is the coefficient vector obtained after deleting observation i. It is a measure of influence.

Several practical remarks are useful:

1. D_i does not have an exact F-distribution, but it is often compared with an $F(p+1, n-p-1)$ reference.
2. A common rule of thumb is that $D_i > 1$ (and sometimes $D_i > 0.5$) indicates that observation i may be influential.
3. An equivalent fitted value form is

$$D_i = \frac{(\hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}_{(-i)})^\top (\hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}_{(-i)})}{(p+1)\hat{\sigma}^2},$$

   so Cook's distance measures the impact of an observation on fitted values (and therefore on estimated coefficients).
4. Another equivalent expression is

$$D_i = \frac{h_{ii}}{1 - h_{ii}} \cdot \frac{d_i^2}{p+1}.$$

Lecture 18 — Friday, July 11, 2014

Recall Cook's distance:

$$\begin{aligned} D_i &= \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)})^\top (\mathbf{X}^\top \mathbf{X})(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)})}{(p+1)\hat{\sigma}^2} \\ &= \frac{(\hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}_{(-i)})^\top (\hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}_{(-i)})}{(p+1)\hat{\sigma}^2} = \frac{h_{ii}}{1 - h_{ii}} \frac{d_i^2}{p+1}, \end{aligned}$$

where h_{ii} is leverage and $d_i = e_i / (\hat{\sigma}\sqrt{1 - h_{ii}})$.

5. Model Selection

Model Selection Criteria

We now turn to model selection. For a model with p predictors, the prediction error variance is

$$\mathrm{Var}(y_{\text{new}} - \hat{y}_{\text{new}}) = \mathrm{Var}(y_{\text{new}} - \hat{\mu}_{\text{new}}) = \mathrm{Var}(y_{\text{new}}) + \mathrm{Var}(\hat{\mu}_{\text{new}}) = \sigma^2 + \mathrm{Var}(\hat{\mu}_{\text{new}}).$$

Proposition 18.1 At the n observed design points, the average prediction error variance for independent future responses is

$$\sigma^2 + \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}(\hat{\mu}_i) = \sigma^2 \left[1 + \frac{p+1}{n}\right].$$

Proof. We compute

$$\sigma^2 + \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}(\hat{\mu}_i) = \sigma^2 + \frac{1}{n}\,\mathrm{tr}(\mathrm{Var}(\hat{\boldsymbol{\mu}})) = \sigma^2 + \frac{1}{n}\,\mathrm{tr}\big(\sigma^2 \mathbf{H}\big) = \sigma^2 + \frac{1}{n}\sigma^2(p+1) = \sigma^2\left[1 + \frac{p+1}{n}\right].$$

□

Thus, at the observed design points, each additional predictor increases the average estimation component of prediction variance by σ^2/n; the irreducible σ^2 term is unchanged.

On the other hand, adding predictors can never decrease R^2, so raw R^2 by itself is not a reliable model selection criterion:

Proposition 18.2 If Model 1 contains Model 2 as a special case, then $R_1^2 \geq R_2^2$. Equivalently, among nested models with the same response, R^2 cannot decrease when predictors are added.

Proof. We have $R^2 = 1 - \mathrm{SSE}/\mathrm{SST}$, where

$$\mathrm{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

is fixed across competing models. Thus, comparing R^2 reduces to comparing SSE. Consider Model

1:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i,$$

and Model 2:

$$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i.$$

Then

$$\text{SSE}_1 = \min_{\beta_0,\beta_1,\beta_2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2, \qquad \text{and} \qquad \text{SSE}_2 = \min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1})^2.$$

Model 2 is Model 1 with the constraint $\beta_2 = 0$, so removing that constraint cannot increase the minimum. Therefore $\text{SSE}_1 \leq \text{SSE}_2$, and hence $R_1^2 \geq R_2^2$. □

A standard correction is **adjusted R^2**, defined by

$$R^2_{\text{adj}} = 1 - \frac{\text{SSE}/(n-p-1)}{\text{SST}/(n-1)} = 1 - \frac{n-1}{n-p-1}(1-R^2).$$

Unlike ordinary R^2, adjusted R^2 can decrease when an added predictor does not reduce SSE enough to offset the loss of degrees of freedom. For this reason, among models fit to the same response, adjusted R^2 is often used as a quick goodness-of-fit measure adjusted for complexity, with larger values preferred.

Proposition 18.3 When one predictor is added to a nested model, adjusted R^2 increases if and only if the absolute value of the t-statistic for that added predictor exceeds 1.

Proof. Suppose the reduced model has p predictors and the full model has one additional predictor. Since SST is common to both models, adjusted R^2 increases precisely when

$$\frac{\text{SSE}_{\text{full}}}{n-p-2} < \frac{\text{SSE}_{\text{red}}}{n-p-1}.$$

Rearranging gives

$$\frac{\text{SSE}_{\text{red}} - \text{SSE}_{\text{full}}}{\text{SSE}_{\text{full}}/(n-p-2)} > 1.$$

The left-hand side is the partial F-statistic with one degree of freedom, which equals the square of the t-statistic for the added coefficient. Hence the condition is $t^2 > 1$, or equivalently $|t| > 1$. □

Because R^2 is monotone in model size, we use criteria that explicitly trade off fit and complexity.

One classical option is **Mallows' C_p**:

$$C_p = (n - p - 1)\frac{\hat{\sigma}_p^2}{\hat{\sigma}_{\text{Full}}^2} - (n - 2(p + 1)),$$

where $\hat{\sigma}_p^2 = \text{SSE}_p/(n-p-1)$ is the error variance estimate from a candidate model with p predictors, and $\hat{\sigma}_{\text{Full}}^2$ is the corresponding estimate from a sufficiently rich full model. If the candidate model is approximately unbiased for the true mean structure, then $\hat{\sigma}_p^2 \approx \hat{\sigma}_{\text{Full}}^2$, so C_p is expected to be close to $p + 1$. In practice, we prefer models with relatively small C_p, paying particular attention to those for which C_p is near $p + 1$.

Another widely used criterion is the **Akaike information criterion** (**AIC**). For Gaussian linear models (up to constants common to all candidate models),

$$\text{AIC} = n \log\left(\frac{\text{SSE}}{n}\right) + 2(p + 1).$$

A lower AIC indicates a better tradeoff between goodness of fit and model complexity. This criterion motivates the forward, backward, and stepwise procedures in the next lecture.

In **all-subsets regression**, all 2^q models formed from q candidate predictors are fitted and then compared using a criterion such as adjusted R^2, AIC, or Mallows' C_p. This is practical only when q is small enough for exhaustive enumeration.

Lecture 19 — Wednesday, July 16, 2014

Forward, Backward, and Stepwise Selection

In **forward selection**, we start from the model containing only an intercept, $y = \beta_0 + \varepsilon$, and compare it with all models obtained by adding one predictor. If an addition lowers AIC, the candidate with the lowest AIC becomes the new current model. We repeat this comparison one predictor at a time and stop when no available addition reduces AIC.

In **backward selection**, we move in the opposite direction: we start from a full model containing all available predictors, compare models obtained by deleting one term at a time, and remove the term whose deletion gives the best AIC improvement. This process continues until no single deletion lowers AIC.

Stepwise selection combines both directions at each stage. After each move, we check whether adding any excluded predictor or deleting any included predictor can further improve AIC, and then take the best available move. The method stops when neither a one-step addition nor a one-step deletion improves the criterion.

Versions of these algorithms can instead use significance thresholds, usually with a prespecified level between 5% and 15%. Forward selection adds the excluded predictor with the smallest partial test p-value, provided that p-value is at most the entry threshold. Backward elimination removes the included predictor with the largest p-value, provided that p-value exceeds the removal threshold. Stepwise regression alternates these two checks and may use different thresholds for entry and removal.

The procedures that move in only one direction have corresponding limitations. A predictor admitted by forward selection can never be removed, even if it later becomes insignificant, and a predictor discarded by backward elimination cannot return. Forward selection can also miss a group of predictors that is important jointly even though none is sufficiently significant when considered alone. Stepwise regression mitigates the first two restrictions, but it remains a local search rather than an exhaustive comparison.

For out-of-sample assessment, the **PRESS statistic** uses leave-one-out predictions. Let $\hat{y}_{(-i)}$ denote the predicted value for observation i obtained from a model fitted without that observation. Then

$$\text{PRESS} = \sum_{i=1}^{n} \left(y_i - \hat{y}_{(-i)}\right)^2 .$$

Smaller values indicate better leave-one-out predictive performance among models fitted to the same response.

Example 19.1 The saved course project training data contain 2,103 flights and 13 variables. Arrival delay is the response, while the candidate predictors include calendar variables, scheduled and actual timing variables, carrier and airport factors, distance, and taxi times. The following code treats month and day of week as factors, fits a large model, and performs backward AIC selection.

```
flight_data <- read.csv("data/project_train.csv")
flight_data$Month <- factor(flight_data$Month)
flight_data$DayOfWeek <- factor(flight_data$DayOfWeek)
```

```r
full_fit <- lm(
  ArrDelay ~ Month + DayOfWeek + CRSArrTime + UniqueCarrier + DepDelay +
    Origin + Dest + Distance + TaxiIn + TaxiOut,
  data = flight_data
)
selected_fit <- step(full_fit, direction = "backward", trace = 0)

formula(selected_fit)
rbind(
  full = c(AIC = AIC(full_fit), adjusted_R2 = summary(full_fit)$adj.r.squared,
           residual_SE = summary(full_fit)$sigma),
  selected = c(AIC = AIC(selected_fit),
               adjusted_R2 = summary(selected_fit)$adj.r.squared,
               residual_SE = summary(selected_fit)$sigma)
)
selected_fit$anova
```

The full fit has residual standard error 5.687, adjusted $R^2 = 0.9069$, and AIC 13,680.7. Backward selection removes `Dest` as one grouped factor term; the selected model has residual standard error 5.859, adjusted $R^2 = 0.9011$, and AIC 13,638.2. Thus AIC accepts a modest reduction in fit to avoid the 211 destination coefficients. This is a useful reminder that the model preferred by a complexity-penalized criterion need not maximize R^2 or minimize the in-sample residual standard error.

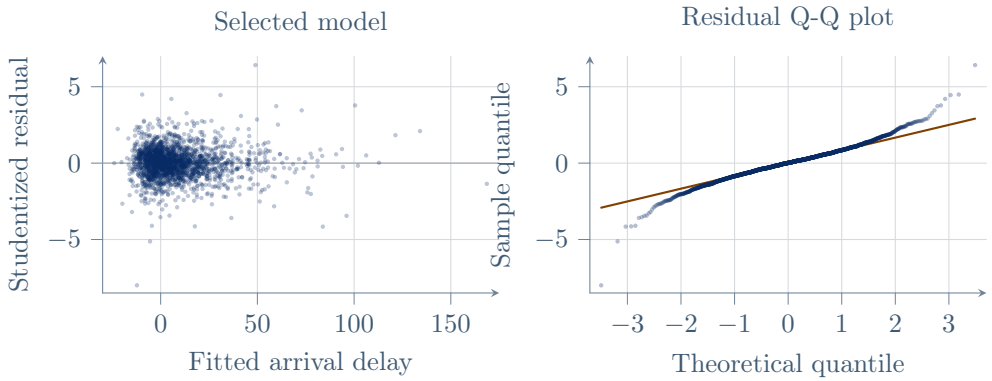


Figure 20

Figure 20 shows residuals against fitted values and normal Q–Q diagnostics for the flight delay model selected by backward AIC, rendered natively with PGFPlots. The central residual distribution is reasonably regular, but the tails contain several influential candidates.

A few applied modeling notes are worth recording. In high-dimensional data, it is helpful to begin with basic screening and data quality checks, since some predictors (for example, scheduling code variables) may contribute little information or have problematic sparsity.

For exploratory diagnostics, we can inspect variable cardinality with `apply(train, 2, function(x) length(unique(x)))` and examine a few pairwise relationships with `pairs(train[c("ArrDelay", "DepDelay", "Distance")])`.

When a level-specific effect appears (for example, `DestEWR`), it is often useful to create a targeted indicator such as `train$OnlyDest <- train$Dest == "EWR"` and then compare its interpretation and stability against related variables, such as origin indicators. These simple checks and recodings often make automated selection more stable and the resulting model easier to interpret.

Lecture 20 — Friday, July 18, 2014

The General F-Test

Example 20.1 As a motivating example, suppose we compare treatment and control via

$$y = \beta_1 \cdot (\text{treatment}) + \beta_2 \cdot (\text{control}) + \varepsilon.$$

If

$$H_0 : \beta_1 = \beta_2 = \beta^*,$$

then

$$y = \beta_1^* \cdot (\text{treatment} + \text{control}) + \varepsilon.$$

Treating trt and ctrl as indicators,

$$\text{trt} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \qquad \text{and} \qquad \text{ctrl} = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

and $\text{trt} + \text{ctrl} = \mathbf{1}$, so under H_0 this reduces to $y = \beta^* + \varepsilon$.

The **additional sum of squares principle** compares nested models, where the reduced model is obtained by imposing linear restrictions on the full model.

For example, consider Model 1: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ (full model) and Model 2: $y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \varepsilon$ (reduced model). The reduced model is obtained by setting $\beta_2 = 0$ in the full model. The question is whether the reduced model is adequate, and we can get a data-driven answer using a hypothesis test called an ***F*-test.**

As a non-nested example, let Model 1 be $y = \beta_0 + \beta_1 x_1 + \beta_2 x_3 + \varepsilon$ and let Model 2 be $y = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$. Neither can be obtained by constraining the other, so they are non-nested.

Theorem 20.2 Under the normal multiple regression model, let $\boldsymbol{\beta}$ be a $(p+1) \times 1$ vector of regression coefficients, let $\mathbf{A}$ be a full-row-rank $r \times (p+1)$ matrix of constants, and let $\boldsymbol{c}$ be a $r \times 1$ vector of constants. To test $H_0 : \mathbf{A}\boldsymbol{\beta} = \boldsymbol{c}$, define $\boldsymbol{\theta} = \mathbf{A}\boldsymbol{\beta}$ and $\hat{\boldsymbol{\theta}} = \mathbf{A}\hat{\boldsymbol{\beta}}$. Then

$$\hat{\boldsymbol{\theta}} \sim \mathcal{N}(\boldsymbol{\theta}, \mathbf{A}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{A}^\top \sigma^2).$$

Under H_0,

$$\frac{(\hat{\boldsymbol{\theta}} - \boldsymbol{c})^\top (\mathbf{A}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{A}^\top)^{-1}(\hat{\boldsymbol{\theta}} - \boldsymbol{c})}{r\hat{\sigma}^2} \sim F(r, n-p-1).$$

Equivalently,

$$F^* = \frac{(\text{SSE}_{\text{red}} - \text{SSE}_{\text{full}})/r}{\text{SSE}_{\text{full}}/(n-p-1)} \geq 0.$$

Proof. Under the normal linear model,

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}\right).$$

Define $\boldsymbol{\theta} = \mathbf{A}\boldsymbol{\beta}$ and $\hat{\boldsymbol{\theta}} = \mathbf{A}\hat{\boldsymbol{\beta}}$. Since $\hat{\boldsymbol{\theta}}$ is a linear transformation of $\hat{\boldsymbol{\beta}}$,

$$\hat{\boldsymbol{\theta}} \sim \mathcal{N}\left(\boldsymbol{\theta}, \sigma^2\mathbf{A}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{A}^\top\right).$$

Let

$$\mathbf{P} = \left[\mathbf{A}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{A}^\top\right]^{-1/2}.$$

Then

$$\frac{1}{\sigma}\mathbf{P}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_r),$$

so

$$\frac{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top \left[\mathbf{A}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{A}^\top\right]^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})}{\sigma^2} \sim \chi^2_r.$$

Also,

$$\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} = \frac{\text{SSE}_{\text{full}}}{\sigma^2} \sim \chi^2_{n-p-1}.$$

Under normality, $\hat{\boldsymbol{\beta}}$ (hence $\hat{\boldsymbol{\theta}}$) is independent of $\hat{\sigma}^2$, so the two chi-square variables above are independent. Therefore,

$$\frac{\dfrac{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top \left[\mathbf{A}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{A}^\top\right]^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})}{r\sigma^2}}{\dfrac{(n-p-1)\hat{\sigma}^2}{(n-p-1)\sigma^2}} \sim F(r, n-p-1),$$

Under H_0, we have $\boldsymbol{\theta} = \boldsymbol{c}$, so this gives

$$\frac{(\hat{\boldsymbol{\theta}} - \boldsymbol{c})^\top (\mathbf{A}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{A}^\top)^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{c})}{r\hat{\sigma}^2} \sim F(r, n-p-1).$$

For completeness, the sum-of-squares identity follows directly from constrained least squares. The normal equations give

$$S(\boldsymbol{\beta}) = \text{SSE}_{\text{full}} + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top \mathbf{X}^\top\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}).$$

Minimizing the quadratic term subject to $\mathbf{A}\boldsymbol{\beta} = \boldsymbol{c}$ by Lagrange multipliers gives

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{A}^\top \left[\mathbf{A}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{A}^\top\right]^{-1} (\mathbf{A}\hat{\boldsymbol{\beta}} - \boldsymbol{c}).$$

Substitution yields

$$\text{SSE}_{\text{red}} - \text{SSE}_{\text{full}} = (\hat{\boldsymbol{\theta}} - \boldsymbol{c})^\top \left[\mathbf{A}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{A}^\top\right]^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{c}).$$

Hence the equivalent statistic is

$$F^* = \frac{(\text{SSE}_{\text{red}} - \text{SSE}_{\text{full}})/r}{\text{SSE}_{\text{full}}/(n-p-1)} \geq 0.$$

□

Thus we reject H_0 if $F^* > F_\alpha(r, n-p-1)$. For example, suppose $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$.

1) Suppose we wish to test

$$H_0 : \beta_2 = \beta_3 = 0.$$

Then $H_0 : \mathbf{A}\boldsymbol{\beta} = \boldsymbol{c}$ where $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$, $\mathbf{A} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ and $\boldsymbol{c} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. Alternatively, we may take $\mathbf{A} = \begin{bmatrix} 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$, $\boldsymbol{c} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

2) Now suppose we wish to test

$$H_0 : \beta_2 = 0 \qquad \text{and} \qquad \beta_1 = \beta_3.$$

Then $\mathbf{A} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}$ and $\boldsymbol{c} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

Lecture 21 — Wednesday, July 23, 2014

For a general F-test of linear hypotheses, we compare a full and a reduced model using

$$F^* = \frac{(\text{SSE}_{\text{red}} - \text{SSE}_{\text{f}})/r}{\text{SSE}_{\text{f}}/\text{df}_{\text{f}}} = \frac{(\text{SSE}_{\text{red}} - \text{SSE}_{\text{f}})/r}{\hat{\sigma}^2_{\text{f}}},$$

and reject H_0 if $F^* > F_\alpha$.

For example, consider the grouped data setup from Example 12.2. The response is continuous weight gain, $n = 10$, and diet is categorical, with observations y_1, y_2, y_3 on Diet #1, observations y_4, y_5, y_6 on Diet #2, and observations y_7, y_8, y_9, y_{10} on Diet #3. The model is

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon,$$

with $x_1 = \mathbb{1}_{\{i \in \{1,2,3\}\}}$, $x_2 = \mathbb{1}_{\{i \in \{4,5,6\}\}}$, and $x_3 = \mathbb{1}_{\{i \in \{7,\ldots,10\}\}}$.

The question here is whether weight gained depends on diet. Thus, we want to test

$$H_0 : \beta_1 = \beta_2 = \beta_3.$$

Note that the standard ANOVA F-test in this coding tests whether all coefficients are zero, which is

different. To answer the question at hand, write $H_0 : \mathbf{A}\boldsymbol{\beta} = \boldsymbol{c}$, with $\boldsymbol{\beta} = \begin{bmatrix}\beta_1\\ \beta_2\\ \beta_3\end{bmatrix}$, $\mathbf{A} = \begin{bmatrix}1 & -1 & 0\\ 0 & 1 & -1\end{bmatrix}$, $\boldsymbol{c} = \begin{bmatrix}0\\ 0\end{bmatrix}$. Here $r = 2$, so

$$F^* = \frac{(\text{SSE}_{\text{red}} - \text{SSE}_{\text{f}})/2}{\hat{\sigma}^2_{\text{f}}},$$

compared against $F_\alpha(2, 10 - 3 = 7)$.

For the setup of Example 20.1, the model is

$$\text{weight} = \beta_1 \cdot \text{treatment} + \beta_2 \cdot \text{control} + \varepsilon$$

and we want to test $H_0 : \beta_1 = \beta_2$. Equivalently, $H_0 : \mathbf{A}\boldsymbol{\beta} = \boldsymbol{c}$ with $\mathbf{A} = \begin{bmatrix}1 & -1\end{bmatrix}$, $\boldsymbol{c} = \begin{bmatrix}0\end{bmatrix}$, $r = 1$. Then

$$F^* = \frac{\text{SSE}_{\text{red}} - \text{SSE}_{\text{f}}}{\hat{\sigma}^2_{\text{f}}}.$$

The exact data from the plant weight example and the nested model comparison are reproduced by the following code.

```
control <- c(4.17, 5.58, 5.18, 6.11, 4.50,
             4.61, 5.17, 4.53, 5.33, 5.14)
treatment <- c(4.81, 4.17, 4.41, 3.59, 5.87,
               3.83, 6.03, 4.89, 4.32, 4.69)
plant_data <- data.frame(
  weight = c(control, treatment),
  group = factor(rep(c("Control", "Treatment"), each = 10))
)

full_fit <- lm(weight ~ group - 1, data = plant_data)
reduced_fit <- lm(weight ~ 1, data = plant_data)

coef(full_fit)
anova(reduced_fit, full_fit)
c(SSE_reduced = deviance(reduced_fit), SSE_full = deviance(full_fit))
```

From the output, we see that

$$\hat{\sigma}_{\mathrm{f}} = 0.6964 \quad \text{and} \quad \mathrm{SSE}_{\mathrm{f}} = (0.6964)^2 \cdot 18,$$
$$\hat{\sigma}_{\mathrm{r}} = 0.704, \quad \text{and} \quad \mathrm{SSE}_{\mathrm{red}} = (0.704)^2 \cdot 19.$$

Hence

$$F^* = \frac{\mathrm{SSE}_{\mathrm{red}} - \mathrm{SSE}_{\mathrm{f}}}{\hat{\sigma}_{\mathrm{f}}^2} = 1.4191.$$

Since `qf(0.95, 1, 18)` returns $F_{0.05}(1, 18) = 4.413873$, we have $F^* < F_{0.05}$, so we do not reject H_0. The command `anova(reduced_fit, full_fit)` gives $F^* = 1.4191$ and $p = 0.2490$, agreeing with the direct calculation.

Note that R's multiple R-squared is $R^2 = 1 - \mathrm{SSE}/\mathrm{SST}$, where $\mathrm{SSE} = \hat{\sigma}^2 \cdot df$ and SST is obtained from the base model ($y \sim 0$); that is,

$$\sum_{i=1}^{n} (y_i - 0)^2$$

for $y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$. For $y = \beta_0 + \beta_1 x_1 + \varepsilon$, the base model is $y \sim 1$, so

$$\mathrm{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

Theorem 10.3 is a special case of the general F-test with

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0.$$

Taking $\mathbf{A} = [\mathbf{0}\ \mathbf{I}_p]$, $\boldsymbol{c} = \mathbf{0}$, and $r = p$ imposes these p restrictions while leaving the intercept unrestricted. The general F-statistic is

$$F^* = \frac{(\mathrm{SSE}_{\mathrm{red}} - \mathrm{SSE}_{\mathrm{f}})/p}{\mathrm{SSE}_{\mathrm{f}}/(n - p - 1)}.$$

Also,

$$\mathrm{SSE}_{\mathrm{red}} = \mathrm{SSE}(y = \beta_0 + \varepsilon) = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \mathrm{SST}.$$

Therefore,

$$F^* = \frac{(\mathrm{SST} - \mathrm{SSE}_{\mathrm{f}})/p}{\mathrm{SSE}_{\mathrm{f}}/(n - p - 1)} = \frac{\mathrm{SSR}_{\mathrm{f}}/p}{\mathrm{SSE}_{\mathrm{f}}/(n - p - 1)},$$

which is the ANOVA statistic.

Lecture 22 — Friday, July 25, 2014

6. Logistic Regression

Binary Response Modelling

We now turn to logistic regression for binary responses. In earlier lectures, the response y_i was continuous. For a binary response model, let

$$y_i = \begin{cases} 1, & \text{if the event of interest occurs for observation } i, \\ 0, & \text{otherwise,} \end{cases}$$

where the augmented predictor vector for observation i is $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})^\top$.

If we naively apply linear regression (for example $y \sim x_1$), fitted values can fall outside $[0, 1]$, as shown in Figure 21:

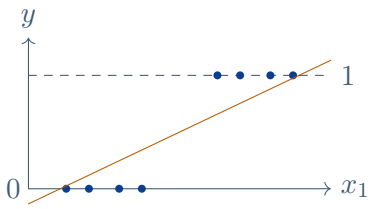


Figure 21

For conditionally independent $y_i \sim \text{Bernoulli}(\pi_i)$, we have $\mathbb{E}[y_i \mid \boldsymbol{x}_i] = \pi_i$, so we model π_i through

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \qquad \text{and} \qquad g(\pi_i) = \eta_i,$$

where the **link function** g maps $(0, 1)$ to $\mathbb{R}$. This defines a Bernoulli generalized linear model; the term **logistic regression** refers specifically to the logit link. Common links are:

1. The **logit link**: $g(\pi_i) = \log(\pi_i/(1 - \pi_i))$, the **log-odds**.
2. The **probit link**: $g(\pi_i) = \Phi^{-1}(\pi_i)$, where $\Phi(z) = \mathbb{P}(Z \leq z)$ for $Z \sim \mathcal{N}(0, 1)$.
3. The **complementary log-log link** $g(\pi_i) = \log(-\log(1 - \pi_i))$.

Under the logit link,

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \boldsymbol{x}_i^\top \boldsymbol{\beta},$$

so

$$\pi_i = \frac{e^{\eta_i}}{1+e^{\eta_i}} = \frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}}.$$

Holding the other predictors fixed, β_j is the change in the log-odds for a one-unit increase in x_j. If π_i^+ denotes the probability after that increase, then

$$\beta_j = \log\left(\frac{\pi_i^+}{1-\pi_i^+} \Big/ \frac{\pi_i}{1-\pi_i}\right),$$

so e^{β_j} is the corresponding **odds ratio**.

To estimate $\boldsymbol{\beta}$, we can use maximum likelihood. Since $y_i \sim \text{Bernoulli}(\pi_i)$,

$$\mathbb{P}(y_i = j) = \pi_i^j (1-\pi_i)^{1-j}, \quad j = 0, 1,$$

and

$$L(\boldsymbol{\beta}) = \prod_{i=1}^n \pi_i^{y_i}(1-\pi_i)^{1-y_i} \qquad \text{and} \qquad \ell(\boldsymbol{\beta}) = \sum_{i=1}^n \left(y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i)\right).$$

Lecture 23 — Wednesday, July 30, 2014

Recall the logistic regression setup:

$$\pi_i = \mathbb{E}\left[y_i \mid \boldsymbol{x}_i\right], \qquad g(\pi_i) = \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}, \qquad \text{and} \qquad g : (0,1) \to \mathbb{R}.$$

For the logit link,

$$g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \log(\text{odds of success}), \qquad \text{and} \qquad \pi_i = \frac{e^{\eta_i}}{1+e^{\eta_i}},$$

and the log-likelihood is

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i)\}.$$

Proposition 23.1 For each $j = 0, \ldots, p$, with $x_{i0} = 1$,

$$\frac{\partial \ell}{\partial \beta_j} = \boldsymbol{x}_j^\top (\boldsymbol{y} - \boldsymbol{\pi}).$$

In matrix form,

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \mathbf{X}^\top (\boldsymbol{y} - \boldsymbol{\pi}) = 0,$$

which defines the score equations for a finite interior maximum likelihood estimate $\hat{\boldsymbol{\beta}}$, when one exists.

Proof. For each observation,

$$\ell_i(\boldsymbol{\beta}) = y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i),$$

so

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}).$$

Fix $j \in \{0, \ldots, p\}$. By the chain rule,

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \pi_i} \frac{\partial \pi_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}.$$

Now

$$\frac{\partial \ell_i}{\partial \pi_i} = \frac{y_i}{\pi_i} - \frac{1 - y_i}{1 - \pi_i}, \qquad \text{and} \qquad \frac{\partial \eta_i}{\partial \beta_j} = x_{ij},$$

and for the logistic mean function $\pi_i = e^{\eta_i}/(1 + e^{\eta_i})$,

$$\frac{\partial \pi_i}{\partial \eta_i} = \frac{e^{\eta_i}}{(1 + e^{\eta_i})^2} = \pi_i(1 - \pi_i).$$

Therefore,

$$\begin{aligned}\frac{\partial \ell}{\partial \beta_j} &= \sum_{i=1}^{n} \left(\frac{y_i}{\pi_i} - \frac{1 - y_i}{1 - \pi_i} \right) \pi_i (1 - \pi_i) x_{ij} \\ &= \sum_{i=1}^{n} \big[y_i (1 - \pi_i) - (1 - y_i) \pi_i \big] x_{ij} \\ &= \sum_{i=1}^{n} (y_i - \pi_i) x_{ij} = \boldsymbol{x}_j^\top (\boldsymbol{y} - \boldsymbol{\pi}).\end{aligned}$$

Stacking these equations for all j yields

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \mathbf{X}^\top (\boldsymbol{y} - \boldsymbol{\pi}).$$

At a finite interior maximum likelihood estimate $\hat{\boldsymbol{\beta}}$, the score is zero, so

$$\mathbf{X}^\top (\boldsymbol{y} - \boldsymbol{\pi}) = 0.$$

□

When a finite maximum likelihood estimator exists, $\hat{\boldsymbol{\beta}}$ solves

$$\mathbf{X}^\top (\boldsymbol{y} - \boldsymbol{\pi}) = 0,$$

which generally has no closed-form solution. We therefore use numerical methods, most commonly iterative reweighted least squares, a version of Newton's method. This idea is developed in STAT 431.

Example 23.2 Two small R examples make these comparisons concrete. The first uses the two binary predictors `type` and `group`; the second uses a simulated continuous predictor to compare the logit, probit, and complementary log-log links.

```
# Two categorical predictors from the lecture handout.
y <- c(1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1)
type <- factor(c(1, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0))
group <- factor(c(rep(1, 6), rep(0, 6)))
categorical_fit <- glm(y ~ type + group, family = binomial(link = "logit"))
new_groups <- expand.grid(type = levels(type), group = levels(group))
new_groups$probability <- predict(categorical_fit, new_groups, type = "response")
coef(summary(categorical_fit))
```

```
new_groups

# One continuous predictor, comparing the three links used in lecture.
set.seed(1000)
y_continuous <- y
x <- rnorm(length(y_continuous), 0, 5)
x[y_continuous == 1] <- x[y_continuous == 1] + 4
link_names <- c("logit", "probit", "cloglog")
link_fits <- lapply(
  link_names,
  function(link) glm(y_continuous ~ x, family = binomial(link = link))
)
sapply(link_fits, coef)

# Move one failure left to make the groups more nearly separated.
x_near_separation <- x
x_near_separation[4] <- x_near_separation[4] - 7
near_separation_fits <- lapply(
  link_names,
  function(link) {
    glm(y_continuous ~ x_near_separation,
        family = binomial(link = link))
  }
)
sapply(near_separation_fits, coef)

# Moving a second failure left produces complete separation.
x_complete_separation <- x_near_separation
x_complete_separation[8] <- x_complete_separation[8] - 7
complete_separation_fits <- lapply(
  link_names,
  function(link) {
    suppressWarnings(
      glm(y_continuous ~ x_complete_separation,
          family = binomial(link = link))
    )
  }
)
sapply(complete_separation_fits, coef)
sapply(complete_separation_fits, function(fit) fit$converged)
```

For the categorical model, the fitted linear predictor is

$$\hat{\eta} = 0.3573 - 0.7146\,\mathbb{1}_{\{\texttt{type}=1\}} + 0.7146\,\mathbb{1}_{\{\texttt{group}=1\}}.$$

The four fitted probabilities for $(\texttt{type}, \texttt{group}) = (0,0), (1,0), (0,1), (1,1)$ are 0.5884, 0.4116, 0.7450, and 0.5884, respectively. In the continuous example, all three links map the real-valued linear predictor into $(0,1)$, but their shapes and tail behaviour differ slightly. With only 12 observations, the three fitted curves are quite similar over the well-populated portion of the predictor range.

Moving one observed failure to the left reduces the overlap between the two response groups. The fitted logit slope increases from 0.3135 to 0.5232, and all three curves become steeper. Moving a second failure to the left produces **complete separation**: every failure lies below every success. In this final configuration, R reports nonconvergence and fitted probabilities numerically equal to 0 or 1; the logit coefficients grow to approximately 50.15 and 30.49. This illustrates why ordinary finite maximum likelihood estimates do not exist under complete separation.

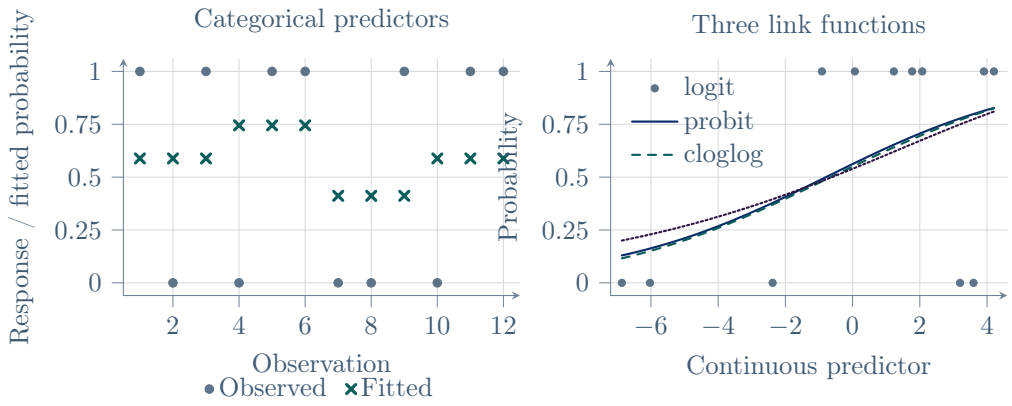


FIGURE 22

Figure 22 shows the observed and fitted values for the categorical example, followed by the fitted logit, probit, and complementary log-log curves for the continuous example.

Figure 23 shows two perturbations of the continuous predictor example. As overlap decreases, the fitted links steepen; under complete separation, numerical fitting drives the curves toward a step function. The fitted coordinates are generated in R and rendered natively with PGFPlots.

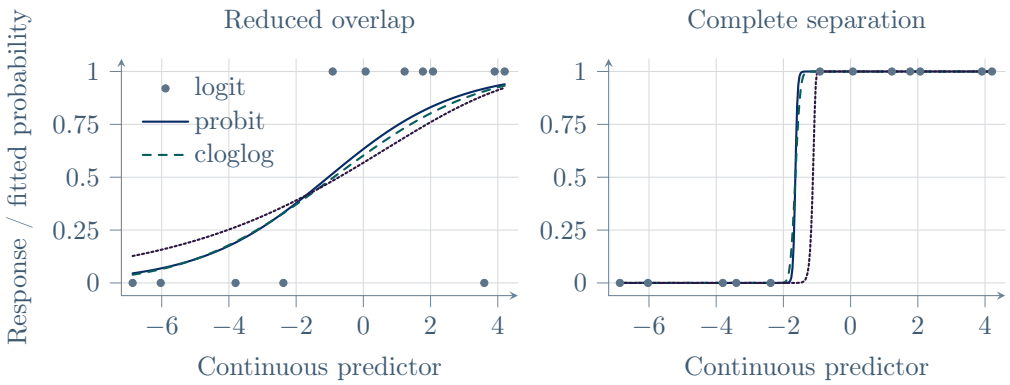

Reduced overlap
Complete separation
Response / fitted probability
logit
probit
cloglog
1
0.75
0.5
0.25
0
−6
−4
−2
0
2
4
Continuous predictor
Continuous predictor

Figure 23

Appendix: Lecture and Tutorial Index