12 Lecture 12 Feb 15

Last time

- Properties of the Least-Squares estimator

Today

- HW1 due Feb 17
- Inference of SLR model
- Multiple linear regression

Properties of the Least-Squares estimator

Under the strong assumptions of the simple regression model, the sample least squares coefficients $\hat{\beta}_{ls}$ have several desirable properties as estimators of the population regression coefficients β_0 and β_1:

- The least-squares intercept and slope are *linear estimators*, in the sense that they are linear functions of the observations y_i.
  Proof:

- The sample least-squares coefficients are *unbiased estimators* of the population regression coefficients:
$$\mathbf{E}\left(\hat{\beta}_0\right) = \beta_0$$
$$\mathbf{E}\left(\hat{\beta}_1\right) = \beta_1$$
  Proof:

- Both $\hat{\beta}_0$ and $\hat{\beta}_1$ have simple sampling variances:
$$\text{Var}(\hat{\beta}_0) = \frac{\sigma_\epsilon^2 \sum x_i^2}{n\sum(x_i - \bar{x})^2}$$
$$\text{Var}(\hat{\beta}_1) = \frac{\sigma_\epsilon^2}{\sum(x_i - \bar{x})^2}$$
  Proof:

- Rewrite the formula for $\text{Var}(\hat{\beta}_1) = \frac{\sigma_\epsilon^2}{(\text{n}-1)\text{S}_\text{X}^2}$, we see that the sampling variance of the slope estimate will be small when
  - The error variance σ_ϵ^2 is small

  - The sample size n is large
  - The explanatory-variable values are spread out (i.e. have a large variance, S_X^2)

- (Gauss-Markov theorem) Under the assumptions of linearity, constant variance, and independence, the least-squares estimators are BLUE (Best Linear Unbiased Estimator), that is they have the smallest sampling variance and are unbiased. (show this)
  Proof:

- Under the full suite of assumptions, the least-squares coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ are the maximum-likelihood estimators of β_0 and β_1. (show this)
  Proof:

- Under the assumption of normality, the least-squares coefficients are themselves normally distributed. Summing up,

$$\hat{\beta}_0 \sim N(\beta_0, \frac{\sigma_\epsilon^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2})$$
$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma_\epsilon^2}{\sum (x_i - \bar{x})^2})$$

Statistical inference of the SLR model

Now we have the distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\hat{\beta}_0 \sim N(\beta_0, \frac{\sigma_\epsilon^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2})$$
$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma_\epsilon^2}{\sum (x_i - \bar{x})^2}).$$

However, σ_ϵ is never known in practice. Instead, an *unbiased* estimator of σ_ϵ^2 is given by

$$\hat{\sigma_\epsilon}^2 = MS[E] = \frac{SS[E]}{n-2}.$$

show that $\mathbf{E}\left(\sum (y_i - \hat{y}_i)^2\right) = \sigma_\epsilon^2 (n-2)$.
Proof:

Confidence intervals

Now we substitute $\hat{\sigma}_\epsilon^2$ into the distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma_\epsilon^2}{\sum (x_i - \bar{x})^2})$$
$$\hat{\beta}_0 \sim N(\beta_0, \frac{\sigma_\epsilon^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2})$$

to get the estimated standard errors:

$$\widehat{SE}(\hat{\beta}_1) = \sqrt{\frac{MS[E]}{\sum(x_i - \bar{x})^2}}$$

$$\widehat{SE}(\hat{\beta}_0) = \sqrt{MS[E]\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right)}$$

And the $100(1-\alpha)\%$ confidence intervals for β_1 and β_0 are given by

$$\hat{\beta}_1 \pm t(n-2, \alpha/2)\sqrt{\frac{MS[E]}{S_{xx}}}$$

$$\hat{\beta}_0 \pm t(n-2, \alpha/2)\sqrt{MS[E]\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}$$

where $S_{xx} = \sum(x_i - \bar{x})^2$

Confidence interval for $\mathbf{E}(Y|X = x_0)$

The conditional mean $\mathbf{E}(Y|X = x_0)$ can be estimated by evaluating the regression function $\mu(x_0)$ at the estimates $\hat{\beta}_0$, $\hat{\beta}_1$. The conditional variance of the expression isn't too difficult (already shown):

$$\mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 \mathrm{x}_0 | \mathrm{X} = \mathrm{x}_0) = \sigma^2(\frac{1}{\mathrm{n}} + \frac{(\mathrm{x}_0 - \bar{\mathrm{x}})^2}{\mathrm{S}_{\mathrm{xx}}})$$

This leads to a confidence interval of the form

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t(n-2, \alpha/2)\sqrt{MS[E]\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

Prediction interval

Often, prediction of the response variable Y for a given value, say x_0, of the independent variable of interest. In order to make statements about future values of Y, we need to take into account

- the sampling distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$
- the randomness of a future value Y.

We have seen the predicted value of Y based on the linear regression is given by $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

The 95% prediction interval has the form

$$\hat{Y}_0 \pm t(n-2, \alpha/2)\sqrt{MS[E]\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}.$$

Hypothesis test

To test the hypothesis $\boxed{H_0 : \beta_1 = \beta_{slope}}$ that the population slope is equal to a specific value β_{slope} (most commonly, the null hypothesis has $\beta_{slope} = 0$), we calculate the test statistic (T-statistics) with $df = n - 2$

$$t_0 = \frac{\hat{\beta}_1 - \beta_{slope}}{\widehat{SE}(\hat{\beta}_1)} \sim t_{n-2}$$

Some questions to answer using regression analysis:

1. What is the meaning, in words, of β_1?

2. True/False: (a) β_1 is a statistic (b) β_1 is a parameter (c) β_1 is unknown.

3. True/False: (a) $\hat{\beta}_1$ is a statistic (b) $\hat{\beta}_1$ is a parameter (c)$\hat{\beta}_1$ is unknown

4. Is $\hat{\beta}_1 = \beta_1$?

Multiple linear regression

JF 5.2+6.2

Multiple linear regression - an example

An example on the prestige, education, and income levels of 45 U.S. occupations (Duncan's data):

	income	education	prestige
accountant	62	86	82
pilot	72	76	83
architect	75	92	90
author	55	90	76
chemist	64	86	90
minister	21	84	87
professor	64	93	93
dentist	80	100	90
reporter	67	87	52
engineer	72	86	88
lawyer	76	98	89
teacher	48	91	73

"prestige" represents the percentage of respondents in a survey who rated an occupation as "good" or "excellent" in prestige, "education" represents the percentage of incumbents in the

occupation in the 1950 U.S. Census who were high school graduates, and "income" represents the percentage of occupational incumbents who earned incomes in excess of $3,500.

Using the `pairs` command in R, we can look at the pairwise scatter plot between the three variables as in Figure 12.1.

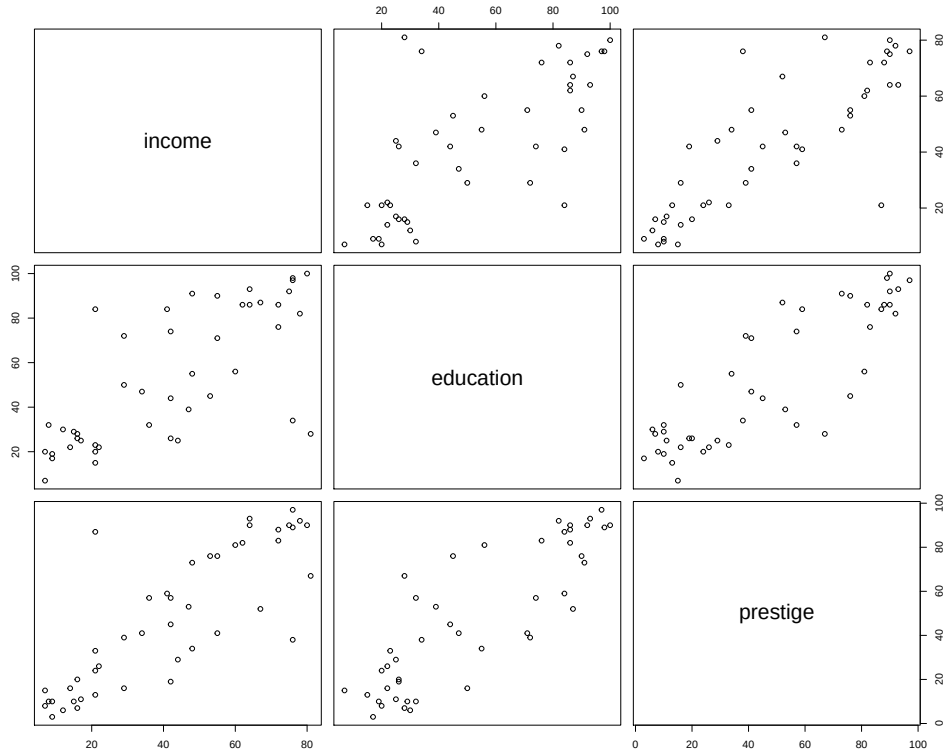

Figure 12.1: Scatterplot matrix for occupational prestige, level of education, and level of income of 45 U.S. occupations in 1950.

Consider a regression model for the "prestige" of occupation i, Y_i, in which the mean of Y_i is a linear function of two predictor variables $X_{i1} = income, X_{i2} = education$ for occupations $i = 1, 2, \ldots, 45$:

$$Y = \beta_0 + \beta_1 income + \beta_2 education + error$$

or

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

or

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \epsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \epsilon_2 \\ \vdots &= \vdots \\ Y_{45} &= \beta_0 + \beta_1 X_{45,1} + \beta_2 X_{45,2} + \epsilon_{45} \end{aligned}$$

A multiple linear regression (MLR) model w/ p independent variables

Let p independent variables be denoted by $x_1, \ldots, x_p$.

- Observed values of p independent variables for i^{th} subject from sample denoted by $x_{i1}, \ldots, x_{ip}$
- response variable for i^{th} subject denoted by Y_i
- For $i = 1, \ldots, n$, MLR model for Y_i:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

- As in SLR, $\epsilon_1, \ldots, \epsilon_n \overset{iid}{\sim} N(0, \sigma^2)$

Least squares estimates of regression parameters minimize $SS[E]$:

$$SS[E] = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$$

$$\boxed{\hat{\sigma}^2 = \frac{SS[E]}{n-p-1}}$$

Interpretations of regression parameters:

- σ^2 is unknown error variance parameter
- $\beta_0, \beta_1, \ldots, \beta_p$ are $p + 1$ unknown regression parameters:
  - β_0: average response when $x_1 = x_2 = \cdots = x_p = 0$
  - β_i is called a partial slope for x_i. Represents mean change in y per unit increase in x_i *with all other independent variables held fixed.*