14   Lecture 14 Feb 17

Last time

- Inference of SLR model

Today

- Confidence intervals for SLR
- Review Lab 1 and 2
- Multiple linear regression

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 $\mathbf{E}\left(Y|X = x_0\right)$

The conditional mean $\mathbf{E}\left(Y|X = x_0\right)$ 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):

$$\text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0 | X = x_0) = \sigma^2(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}})$$

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 14.1.
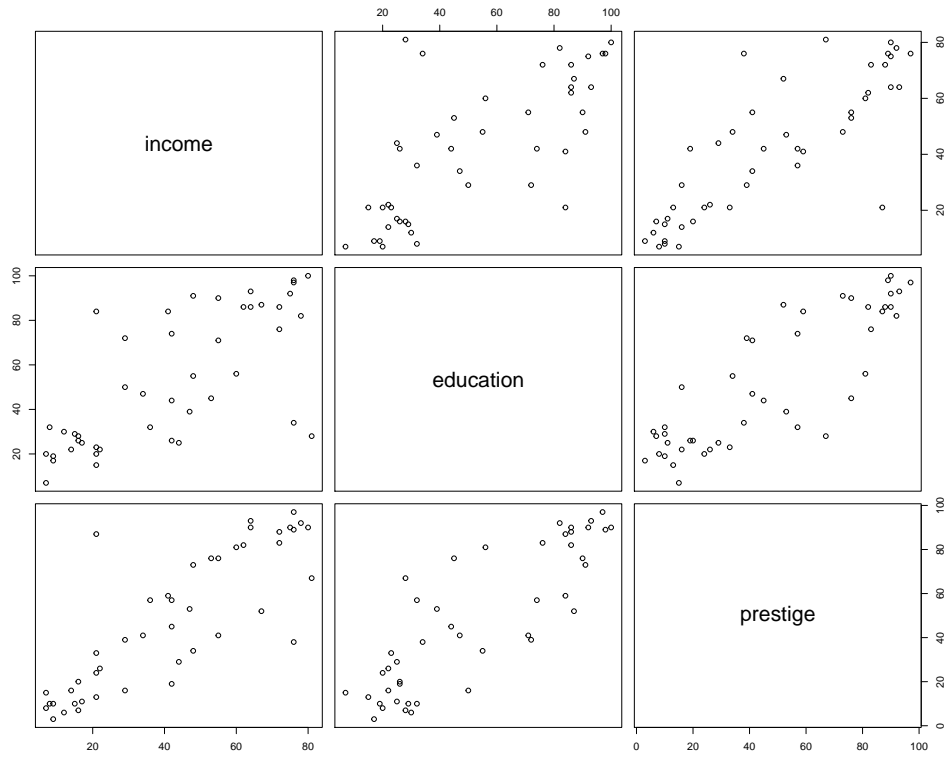
3

Figure 14.1: 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

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

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 $x_{i1}, \ldots, x_{ip}$

4

- 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 <u>error variance</u> 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 <u>partial slope</u> for x_i. Represents mean change in y per unit increase in x_i *with all other independent variables held fixed.*

Matrix formulation of MLR

Let a vector for p observed independent variables for individual i be defined by

$$x_{i\cdot} = (1, x_{i1}, x_{i2}, \ldots, x_{ip}).$$

The MLR model for $Y_1, \ldots, Y_n$ is given by

$$Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \cdots + \beta_p X_{1p} + \epsilon_1$$
$$Y_2 = \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \cdots + \beta_p X_{2p} + \epsilon_2$$
$$\vdots = \vdots$$
$$Y_n = \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \cdots + \beta_p X_{np} + \epsilon_n$$

This system of n equations can be expressed using matrices:

$$\boxed{\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}}$$

where

- $\mathbf{Y}$ denotes a <u>response vector</u> of size $n \times 1$

- $\mathbf{X}$ denotes a <u>design matrix</u> of size $n \times (p+1)$

- $\boldsymbol{\beta}$ denotes a vector of <u>regression parameters</u> of size $(p+1) \times 1$

- $\boldsymbol{\epsilon}$ denotes an error vector of size $n \times 1$

Here, the error vector ϵ is assumed to follow a multivariate normal distribution with variance-covariance matrix $\sigma^2 \mathbf{I}_n$. For individual i,

$$y_i = \mathbf{x}_{i.}\boldsymbol{\beta} + \epsilon_i.$$

Some simplified expressions: ($\mathbf{a}$ is a known $p \times 1$ vector)

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\
\mathbf{Var}\left(\hat{\boldsymbol{\beta}}\right) &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \\
&= \boldsymbol{\Sigma} \\
\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) &= MS[E](\mathbf{X}^T\mathbf{X})^{-1} \\
&= \widehat{\boldsymbol{\Sigma}} \\
\widehat{\mathrm{Var}}(\mathbf{a}^T\hat{\boldsymbol{\beta}}) &= \mathbf{a}^T\widehat{\boldsymbol{\Sigma}}\mathbf{a}
\end{aligned}$$

Question: what are the dimensions of each of these quantities?

- $(\mathbf{X}^T\mathbf{X})^{-1}$ may be verbalized as " x transposed x inverse"

- $\widehat{\boldsymbol{\Sigma}}$ is the estimated variance-covariance matrix for the estimate of the regression parameter vector $\hat{\beta}$

- $\mathbf{X}$ is assumed to be of full *rank*.

Some more simplified expressions:

$$\begin{aligned}
\hat{\mathbf{Y}} &= \mathbf{X}\hat{\boldsymbol{\beta}} \\
&= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\
&= \mathbf{H}\mathbf{Y} \\
\hat{\boldsymbol{\epsilon}} &= \mathbf{Y} - \hat{\mathbf{Y}} \\
&= \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} \\
&= (\mathbf{I} - \mathbf{H})\mathbf{Y}
\end{aligned}$$

- $\hat{\mathbf{Y}}$ is called the vector of fitted or predicted values

- $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is called the hat matrix

- $\hat{\boldsymbol{\epsilon}}$ is the vector of residuals

For the Duncan's data example on income, education and prestige, with $p = 2$ independent variables and $n = 45$ observations,

$$\mathbf{X} = \begin{bmatrix} 1 & 62 & 86 \\ 1 & 72 & 76 \\ \vdots & \vdots & \vdots \\ 1 & 8 & 32 \end{bmatrix}$$

and

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 45 & 1884 & 2365 \\ 1884 & 105148 & 122197 \\ 2365 & 122197 & 163265 \end{bmatrix}$$

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 0.10211 & -0.00085 & -0.00084 \\ -0.00085 & 0.00008 & -0.00005 \\ -0.00084 & -0.00005 & 0.00005 \end{bmatrix}$$

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \begin{bmatrix} -6.0646629 \\ 0.5987328 \\ 0.5458339 \end{bmatrix} =?$$

$$SS[E] = \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = (\mathbf{Y} - \hat{\mathbf{Y}})^T(\mathbf{Y} - \hat{\mathbf{Y}}) = 7506.7$$

$$MS[E] = \frac{SS[E]}{df} = \frac{7506.7}{45 - 2 - 1} = 178.73$$

$$\widehat{\boldsymbol{\Sigma}} = MS[E](\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 18.249481 & -0.151845008 & -0.150706025 \\ -0.151845 & 0.014320275 & -0.008518551 \\ -0.150706 & -0.008518551 & 0.009653582 \end{bmatrix}$$