9 Lecture 9: Feb 8

Last time

- Inference of SLR model
- Lab 1

Today

- SLR questions
- Multiple Linear Regression

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

<u>Multiple linear regression - an example</u>

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 9.1.

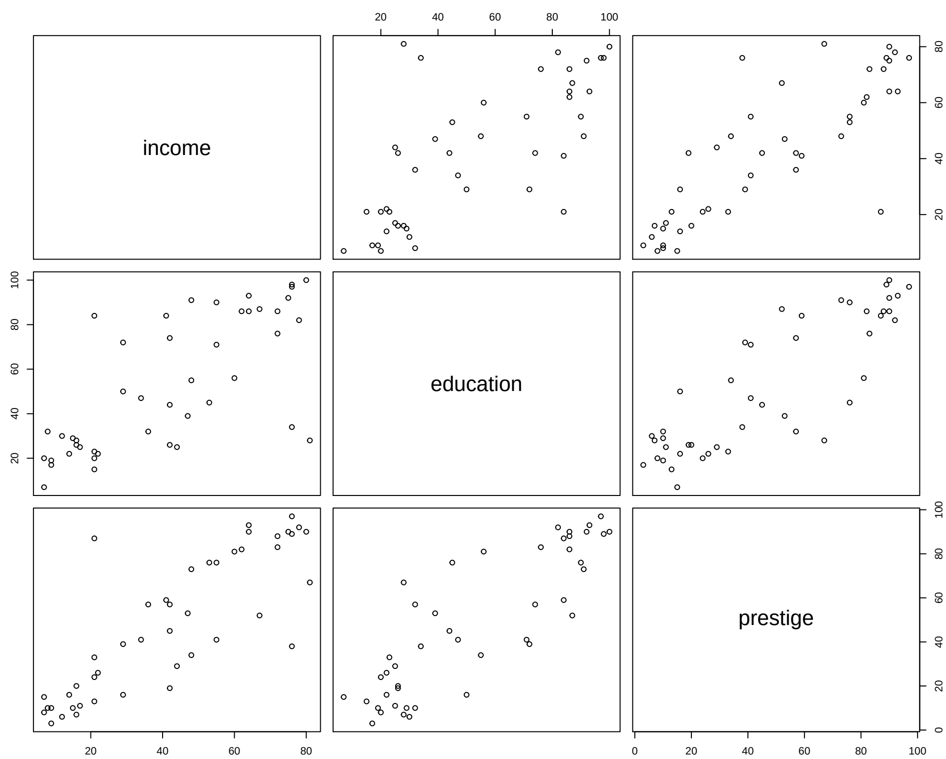

Figure 9.1: Scatterplot matrix for occupational prestige, level of education, and level of income of 45 U.S. occupations in 1950.

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 $X_{i1} = \textit{income}, X_{i2} = \textit{education}$ for occupations $i = 1, 2, \ldots, 45$:

$$Y = \beta_0 + \beta_1 \textit{income} + \beta_2 \textit{education} + \textit{error}$$

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.*

Matrix formulation of MLR

Let a $(1 \times (p+1))$ 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

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

This system of n equations can be expressed using matrices:

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

where

- $\mathbf{Y}$ denotes a response vector of size $n \times 1$
- $\mathbf{X}$ denotes a design matrix of size $n \times (p+1)$
- β denotes a vector of regression parameters of size $(p+1) \times 1$
- ϵ 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,

$$Y_i = x_{i\cdot}\beta + \epsilon_i.$$

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

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

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

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

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

- $\hat{\mathbf{Y}}$ 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
- ϵ 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

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

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

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

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

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

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

Some questions:

1. What is the estimate of β_1? Interpretation?

2. What is the standard error of $\hat{\beta}_1$?

3. Is $\beta_1 = 0$ plausible, while controlling for possible linear associations between Prestige and Education? $(t(0.025, 42) = 2.02)$

4. Estimate the mean prestige among the population of ALL occupations with $income = 42$ and $education = 84$.

5. Report a standard error

6. Report a 95% confidence interval