12 Lecture 12: Feb 21

Last time

- Inference of SLR model

Today

- remember to send me the link to HW1 tag
- Lab 2 review
- 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

<div align="center">Multiple linear regression - an example</div>

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.
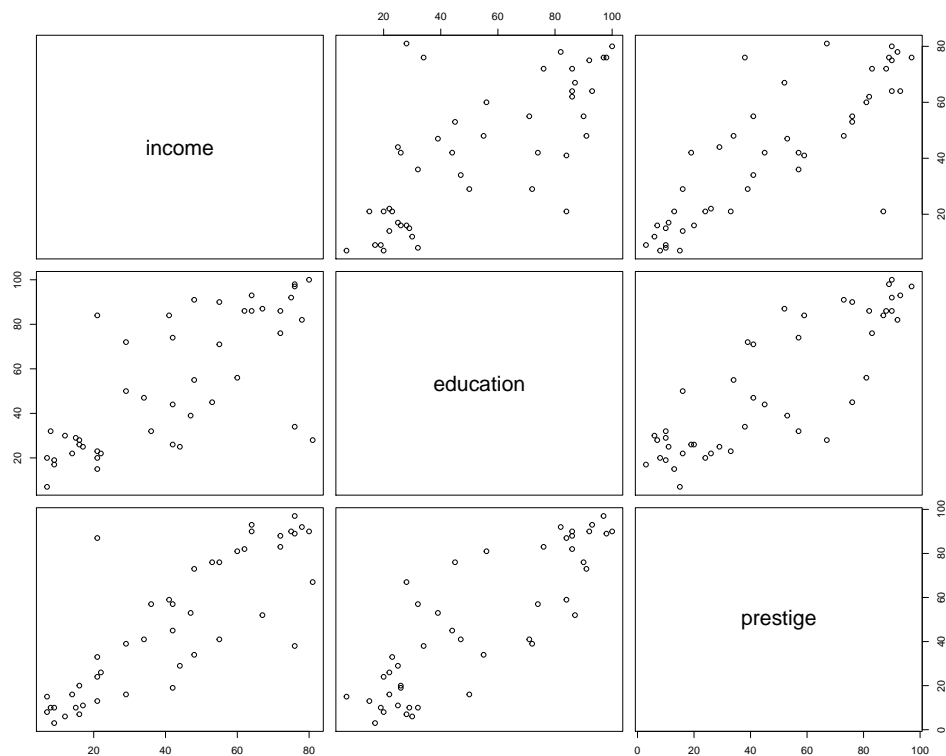

Figure 12.1: Scatterplot matrix for occupational prestige, level of education, and level of income of 45 U.S. occupations in 1950.

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

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

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

- 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 <u>with all other independent variables held fixed.</u>

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

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

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 <u>error vector</u> 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,

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

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

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

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 $\hat{\beta}$

- **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 <u>fitted</u> or <u>predicted values</u>
- $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is called the <u>hat matrix</u>
- $\hat{\boldsymbol{\epsilon}}$ is the vector of <u>residuals</u>

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

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

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

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

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

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

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

Multiple correlation, JF 5.2.3

The sums of squares in multiple regression are defined in the same manner as in SLR:

$$TSS = \sum (Y_i - \bar{Y})^2$$
$$RegSS = \sum (\hat{Y}_i - \bar{Y})^2$$
$$RSS = \sum (Y_i - \hat{Y}_i)^2 = \sum \epsilon_i^2$$

Not surprisingly, we have a similar analysis of variance for the regression:

$$TSS = RegSS + RSS$$

The squared multiple correlation R^2, representing the proportion of variation in the response variable captured by the regression, is defined in terms of the sums of squares:

$$R^2 = \frac{RegSS}{TSS} = 1 - \frac{RSS}{TSS}.$$

Because there are several slope coefficients, potentially with different signs, the *multiple correlation coefficient* is, by convention, the positive square root of R^2. The multiple correlation is also interpretable as the simple correlation between the fitted and observed Y values, i.e. $r_{\hat{Y}Y}$.

Adjusted-R^2

Because the multiple correlation can only rise, never decline, when explanatory variables are added to the regression equation (HW1), investigators sometimes penalize the value of R^2 by a "correction" for degrees of freedom. The corrected (or "adjusted") R^2 is defined as:

$$R_{adj}^2 = 1 - \frac{\frac{RSS}{n-p-1}}{\frac{TSS}{n-1}}$$
$$= 1 - \left[\frac{(1 - R^2)(n - 1)}{n - p - 1} \right]$$

Confidence intervals

Confidence intervals and hypothesis tests for individual coefficients closely follow the pattern of simple-regression analysis:

1. substitute an estimate of the error variance (MSE) for the unknown σ^2 into the variance term of $\hat{\beta}_i$

2. find the estimated standard error of a slope coefficient $\widehat{SE}(\hat{\beta}_i)$

3. $t = \frac{\hat{\beta}_i - \beta_i}{\widehat{SE}(\hat{\beta}_i)}$ follows a t-distribution with degrees of freedom as associated with SSE.

Therefore, we can construct the $100(1 - \alpha)\%$ confidence interval for a single slope parameter by (why?):

$$\hat{\beta}_i \pm t(n - p - 1, \alpha/2)\widehat{SE}(\hat{\beta}_i)$$

Hand-waving proof:

Hypothesis tests

We first test the null hypothesis that all population regression slopes are 0:

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

The test statistics,

$$F = \frac{RegSS/p}{RSS/(n - p - 1)}$$

follows an F-distribution with p and $n - p - 1$ degrees of freedom.

We can also test a null hypothesis about a *subset* of the regression slopes, e.g.,

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

Or more generally, test the null hypothesis

$$H_0 : \beta_{q_1} = \beta_{q_2} = \cdots = \beta_{q_k} = 0$$

where $0 \leqslant q_1 < q_2 < \cdots < q_k \leqslant p$ is a subset of k indices. To get the F-statistic for this case, we generally perform the following steps:

1. Fit the *full* ("unconstrained") model, in other words, model that provides context for H_0. Record SSR_{full} and the associated df_{full}

2. Fit the *reduced* ("constrained") model, in other words, full model constrained by H_0. Record SSR_{red} and the associated df_{red}

3. Calculate the F-statistic by

$$F = \frac{[SSR_{red} - SSR_{full}]/(df_{red} - df_{full})}{SSR_{full}/df_{full}}$$

4. Find p-value (the probability of observing an F-statistic that is at least as high as the value that we obtained) by consulting an F-distribution with numerator df $(ndf) = df_{red} - df_{full}$ and denominator df $(ddf) = df_{full}$. Notation: $F_{ndf,ddf}$, see Figure 12.2.

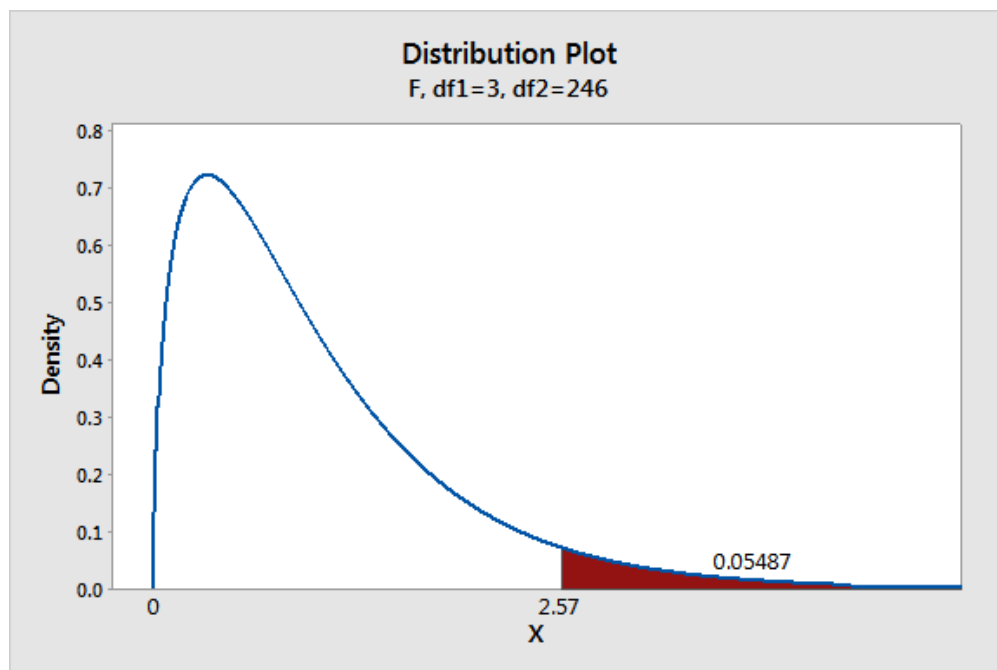Figure 12.2: An example for p-value for F-statistic value 2.57 with an $F_{3,246}$ distribution