16 Lecture 16 Feb 27

Last time

- Confidence intervals for SLR
- Multiple linear regression

Today

- Multiple correlation
- More review
- Dummy variable regression

Multiple correlation, JF 5.2.3

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

$$\begin{aligned} TSS &= \sum (Y_i - \bar{Y})^2 \\ RegSS &= \sum (\hat{Y}_i - \bar{Y})^2 \\ RSS &= \sum (Y_i - \hat{Y}_i)^2 = \sum \hat{\epsilon}_i^2 \end{aligned}$$

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:

$$\begin{aligned} R^2_{adj} &= 1 - \frac{\frac{RSS}{n-p-1}}{\frac{TSS}{n-1}} \\ &= 1 - \left[\frac{(1-R^2)(n-1)}{n-p-1} \right] \end{aligned}$$

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 16.1.

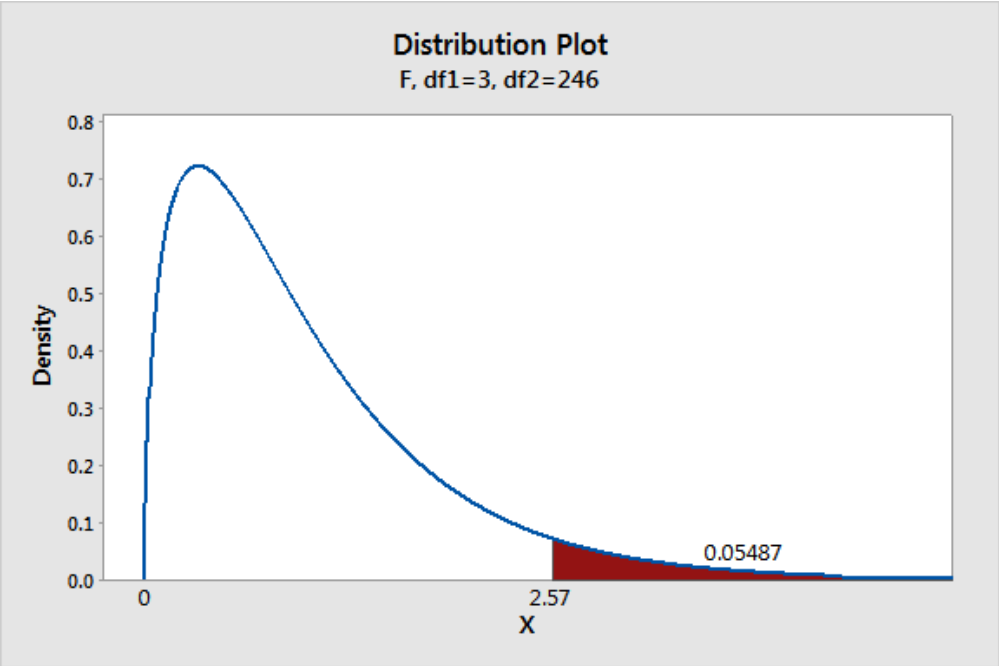

Figure 16.1: An example for p-value for F-statistic value 2.57 with an $F_{3,246}$ distribution

A little more background review

Reference:

- Statistical Inference, 2nd Edition, by George Casella & Roger L. Berger
- Review of Probability Theory by Arian Maleki and Tom Do

Chi-square, t-, and F-Distributions

Let $Z_1, Z_2, \ldots, Z_k \overset{iid}{\sim} N(0,1)$, then $X \equiv Z_1^2 + Z_2^2 + \cdots + Z_k^2 \sim \chi_k^2$ (with k degrees of freedom). If $X \sim \chi_k^2$

$$\mathbf{E}(X) = k$$
$$\mathbf{Var}(X) = 2k.$$

Student's t versus χ^2

If $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, then

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

When σ is unknown,

$$\frac{\bar{X}-\mu}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1}, \quad \text{where } \hat{\sigma} = \sqrt{\frac{\sum(X_i-\bar{X})^2}{n-1}}.$$

Note that

$$\begin{aligned}\frac{\bar{X}-\mu}{\hat{\sigma}/\sqrt{n}} &= \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \cdot \frac{1}{\frac{\hat{\sigma}}{\sigma}} \\ &= Z \cdot \frac{1}{\sqrt{\frac{\sum(X_i-\bar{X})^2}{(n-1)\sigma^2}}} \\ &= \frac{Z}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}}\end{aligned}$$

F versus χ^2

$$F_{ndf,ddf} \equiv \frac{\chi^2_{ndf}/ndf}{\chi^2_{ddf}/ddf}$$

t versus F

$$\begin{aligned}t_k &= \frac{Z}{\sqrt{\chi^2_k/k}} \\ &= \frac{\sqrt{\chi^2_1/1}}{\sqrt{\chi^2_k/k}} \\ &= \sqrt{F_{1,k}}\end{aligned}$$

or, in other words, $t^2_k = F_{1,k}$

Random vectors and matrices

The cdf for random vector

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \text{ is } F_{\mathbf{Y}}(\mathbf{y}) = \Pr(Y_1 \leqslant y_1, Y_2 \leqslant y_2, \ldots, Y_n \leqslant y_n)$$

If a joint pdf exists, then $f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{Y}}(y_1, \ldots, y_n)$ and

$$F_{\mathbf{Y}}(\mathbf{y}) = \int_{-\infty}^{y_1}\int_{-\infty}^{y_2}\cdots\int_{-\infty}^{y_n} f_{\mathbf{Y}}(\mathbf{t})d\mathbf{t}$$

Moments

$$\begin{aligned}
\mathbf{E}(\mathbf{Y}) = \boldsymbol{\mu}_{\mathbf{Y}} &= \begin{bmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_n) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} \\
\mathbf{Var}(\mathbf{Y}) &= \mathbf{E}\left((\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}})(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}})^T\right) \\
&= \mathbf{E}\left(\begin{bmatrix} (Y_1 - \mu_1)^2 & (Y_1 - \mu_1)(Y_2 - \mu_2) & \dots \\ (Y_2 - \mu_2)(Y_1 - \mu_1) & (Y_2 - \mu_2)^2 & \dots \\ \dots & & \end{bmatrix}\right) \\
&= \mathbf{E}\left([(Y_i - \mu_i)(Y_j - \mu_j), i = 1, 2, \dots, n, j = 1, 2, \dots, n]\right) \\
&= (\sigma_{ij})_{i=1,2,\dots,n; j=1,2,\dots,n}
\end{aligned}$$

where $\sigma_{ij} = Cov(Y_i, Y_j)$

Linear functions

Let $\mathbf{X} \in \mathbb{R}^{k \times 1}$, $\mathbf{Y} \in \mathbb{R}^{n \times 1}$ and $\mathbf{A} \in \mathbb{R}^{k \times 1}$, $\mathbf{B} \in \mathbb{R}^{k \times n}$ be non-random, then

$$\begin{aligned}
\underset{k \times 1}{\mathbf{X}} &= \underset{k \times 1}{\mathbf{A}} + \underset{k \times n}{\mathbf{B}} \underset{n \times 1}{\mathbf{Y}} \\
\mathbf{E}(\mathbf{X}) &= \mathbf{A} + \mathbf{B}\mathbf{E}(\mathbf{Y}) \\
\mathbf{Var}(\mathbf{X}) &= \mathbf{B}\mathbf{Var}(\mathbf{Y})\mathbf{B}^T
\end{aligned}$$

Sums of random vectors

$$\begin{aligned}
\underset{n \times 1}{\mathbf{X}} &= \underset{n \times 1}{\mathbf{Y}} + \underset{n \times 1}{\mathbf{Z}} \\
\mathbf{E}(\mathbf{X}) &= \mathbf{E}(\mathbf{Y}) + \mathbf{E}(\mathbf{Z}) = \mathbf{E}(\mathbf{Y} + \mathbf{Z})
\end{aligned}$$

Note that there is no independence assumed above.

$$\mathbf{Var}(\mathbf{X}) = \mathbf{Var}(\mathbf{Y} + \mathbf{Z}) = \mathbf{Var}(\mathbf{Y}) + \mathbf{Var}(\mathbf{Z}) + Cov(\mathbf{Y}, \mathbf{Z}) + Cov(\mathbf{Z}, \mathbf{Y})$$

If $\mathbf{Y}, \mathbf{Z}$ are uncorrelated, then $\mathbf{Var}(\mathbf{X}) = \mathbf{Var}(\mathbf{Y}) + \mathbf{Var}(\mathbf{Z})$

Dummy-variable regression

For categorical data (factor), we use dummy variable regression:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \epsilon_i$$

where D, called a dummy variable regressor or an indicator variable, is coded 1 for one level and 0 for all others,

$$D_i = \begin{cases} 1 & \text{for men} \\ 0 & \text{for women} \end{cases}.$$

Therefore, for women, the model becomes

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

and for men

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 + \epsilon_i = (\beta_0 + \beta_2) + \beta_1 X_i + \epsilon_i$$

For example, Figure 16.2 (a) and (b) represents two small (idealized) populations. In both cases, the within-gender regressions of income on education are parallel. Parallel regressions imply additive effects of education and gender on income: Holding education constant, the "effect" of gender is the vertical distance between the two regression lines, which, for parallel lines, is everywhere the same.

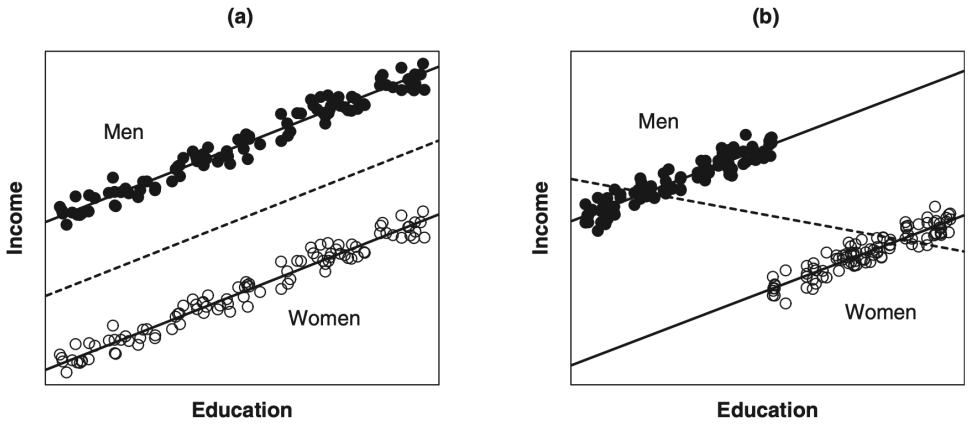

Figure 16.2: Idealized data representing the relationship between income and education for populations of men (filled circles) and women (open circles). In (a), there is no relationship between education and gender; in (b), women have a higher average level of education than men. In both (a) and (b), the within-gender (i.e., partial) regressions (solid lines) are parallel. In each graph, the overall (i.e. marginal) regression of income on education (ignoring gender) is given by the broken line. JF Figure 7.1.

Multi-level factor

We can model the effects of classification factors with m categories (levels) by using $m - 1$ indicator variables.

For example, the three-category occupational-type factor can be represented in the regression equation by introducing two dummy regressors:

Category	D_1	D_2
Professional and managerial	1	0
White collar	0	1
Blue collar	0	0

A model for the regression of prestige on income, education, and type of occupation is then

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \epsilon_i$$

where X_1 is income and X_2 is education. This model describes three parallel regression planes, which can differ in their intercepts:

$$\begin{array}{ll} \text{Professional:} & Y_i = (\beta_0 + \gamma_1) + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \\ \text{White collar:} & Y_i = (\beta_0 + \gamma_2) + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \\ \text{Blue collar:} & Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \end{array}$$

Therefore, the coeficient β_0 gives the intercept for blue-collar occupations; γ_1 represents the constant vertical difference between the parallel regression planes for professional and blue-collar occupations (fixing the values of education and income); and γ_2 represents the constant vertical distance between the regression planes for white-collar and blue-collar occupations (again, fixing education and income).

In the above prestige example, we chose "blue collar" as the baseline category. Sometimes, it is natural to pick a particular category as the baseline category, for example, the "control group" in an experiment. However, in most applications, the choice of a baseline category is entirely arbitrary.

Matrix representation

For the above prestige model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \epsilon_i$$

we have the design matrix $\mathbf{X}$ as

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & D_{11} & D_{12} \\ 1 & X_{21} & X_{22} & D_{21} & D_{22} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & D_{n1} & D_{n2} \end{bmatrix}$$

and the vector of coefficients β is

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \gamma_1 \\ \gamma_2 \end{bmatrix}$$

such that we have (again) the linear model in matrix form:

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

where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, in other words, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.