17 Lecture 17: February 26

Last time

- Multiple linear regression

Today

- HW1 review
- Dummy-Variable regression
- Interactions

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 17.1.

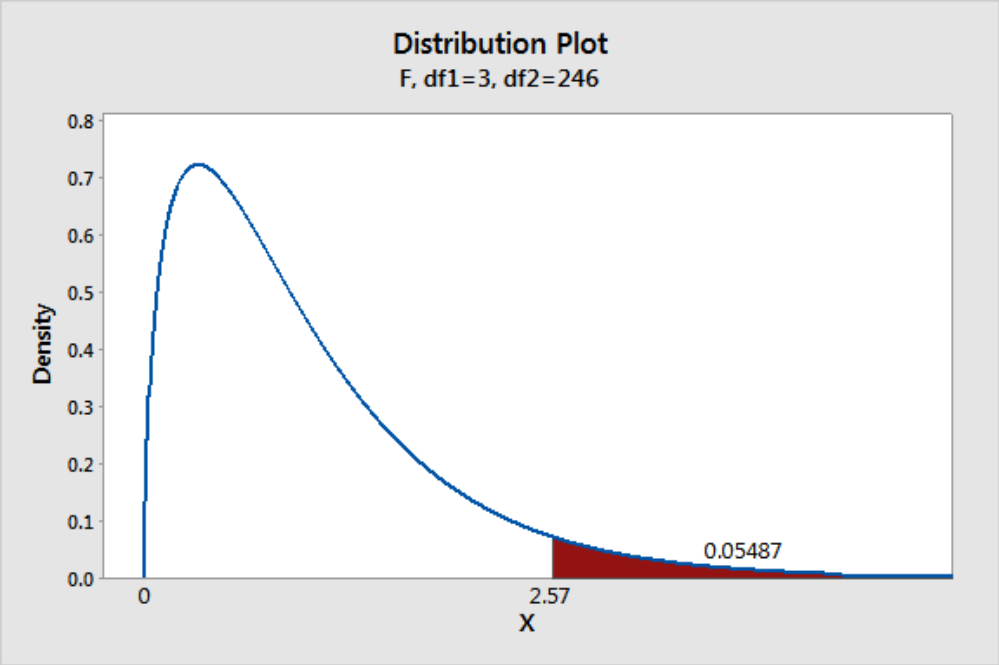

Figure 17.1: An example for p-value for F-statistic value 2.57 with an $F_{3,246}$ distribution

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 17.2 (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.

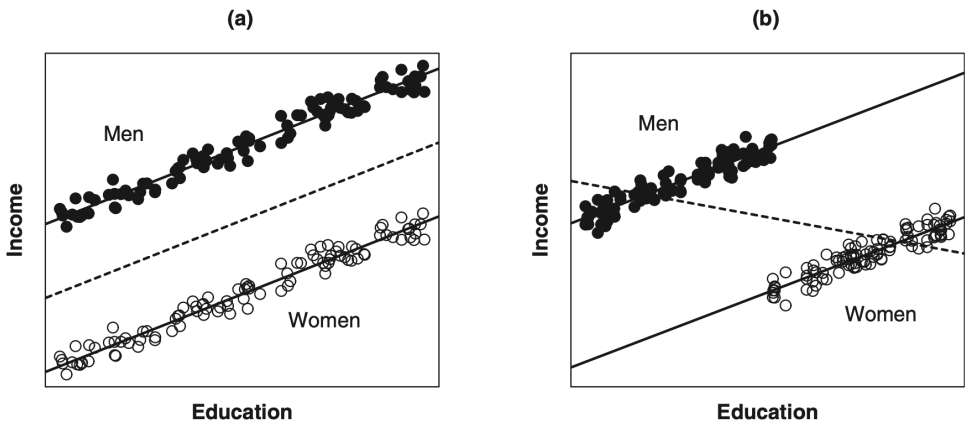

Figure 17.2: Idealized data representing the relationship between income and education for populations of men (filled circles) and women (open circles). In (a), there is no relationship between education and gender; in (b), women have a higher average level of education than men. In both (a) and (b), 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:

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

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, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.

Interactions

Two explanatory variables are said to interact in determining a response variable when the partial effect of one depends on the value of the other. Consider the hypothetical data shown in Figure 17.3.

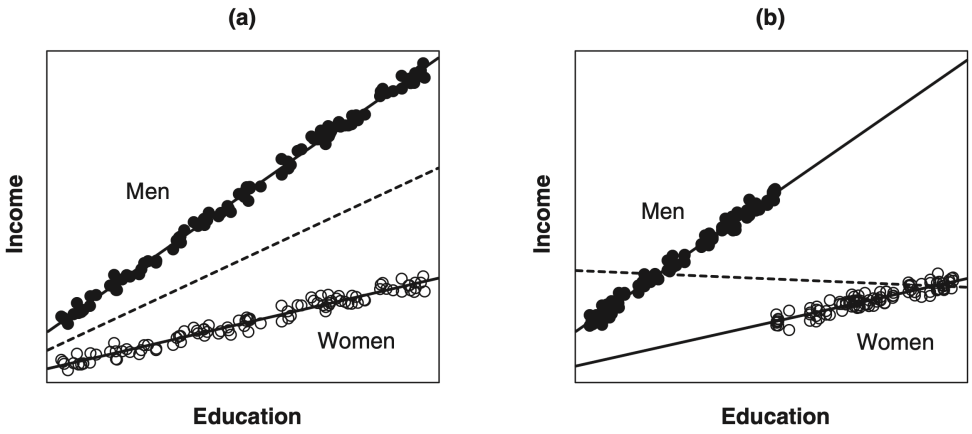

Figure 17.3: Idealized data representing the relationship between income and education for populations of men (filled circles) and women (open circles). In (a), there is no relationship between education and gender; in (b), women have a higher average level of education than men. In both (a) and (b), the within-gender (i.e., partial) regressions (solid lines) are not parallel. The slope for men is greater than the slope for women, and consequently education and gender interact in affecting income. In each graph, the overall regression of income on education (ignoring gender) is given by the broken line. JF Figure 7.7.

It is apparent in both Figure 17.3 (a) and (b) the within-gender regressions of income on education are not parallel: In both cases, the slope for men is larger than the slope for women.

Modeling interactions

We accommodate the interaction of education and gender by:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i D_i) + \epsilon_i$$

where we introduce the interaction regressor XD into the regression equation. For women, the model becomes

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \beta_2 \cdot 0 + \beta_3 (X_i \cdot 0) + \epsilon_i \\ &= \beta_0 + \beta_1 X_i + \epsilon_i \end{aligned}$$

and for men

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \beta_2 \cdot 1 + \beta_3 (X_i \cdot 1) + \epsilon_i \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_i + \epsilon_i \end{aligned}$$

The parameters β_0 and β_1 are, respectively, the intercept and slope for the regression of income on education among women (the baseline category for gender); β_2 gives the difference in intercepts between the male and female groups; and β_3 gives the difference in slopes between the two groups.

Usual guidance: Models that include an interaction between two predictors should also include the individual predictors by themselves regardless of the statistical significance of the associated β's.

Test for the interaction

We can simply test the hypothesis $H_0 : \beta_3 = 0$ and construct the test statistic $t = \frac{\hat{\beta}_i - 0}{\widehat{SE}(\hat{\beta}_i)} \sim t_{n-4}$ $(p = 3)$.

Interactions with multi-level factor

We can easily extend the method for modeling interactions by forming product regressors to multi-level factors, to several factors, and to several quantitative explanatory variables. Using the occupational prestige example, the occupational type could possibly interact both with income (X_1) and with education (X_2):

$$\begin{aligned} Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 D_{i1} + \gamma_2 D_{i2} \\ + \delta_{11} X_{i1} D_{i1} + \delta_{12} X_{i1} D_{i2} + \delta_{21} X_{i2} D_{i1} + \delta_{22} X_{i2} D_{i2} + \epsilon_i \end{aligned}$$

The model therefore permits different intercepts and slopes for the three types of occupations:

$$\begin{array}{lrrrrr} \text{Professional:} & Y_i = & (\beta_0 + \gamma_1) + & (\beta_1 + \delta_{11}) X_{i1} + & (\beta_2 + \delta_{21}) X_{i2} + & \epsilon_i \\ \text{White collar:} & Y_i = & (\beta_0 + \gamma_2) + & (\beta_1 + \delta_{12}) X_{i1} + & (\beta_2 + \delta_{22}) X_{i2} + & \epsilon_i \\ \text{Blue collar:} & Y_i = & \beta_0 + & \beta_1 X_{i1} + & \beta_2 X_{i2} + & \epsilon_i \end{array}$$