

10 Lecture 10: Feb 10

Last time

- SLR questions
- Multiple Linear Regression

Today

- Multiple correlation
- Confidence intervals and hypothesis tests
- R practice with questions

Multiple correlation, JF 5.2.3

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

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

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:

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

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 $0 \leq q_1 < q_2 < \cdots < q_k \leq p$ 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 10.1.

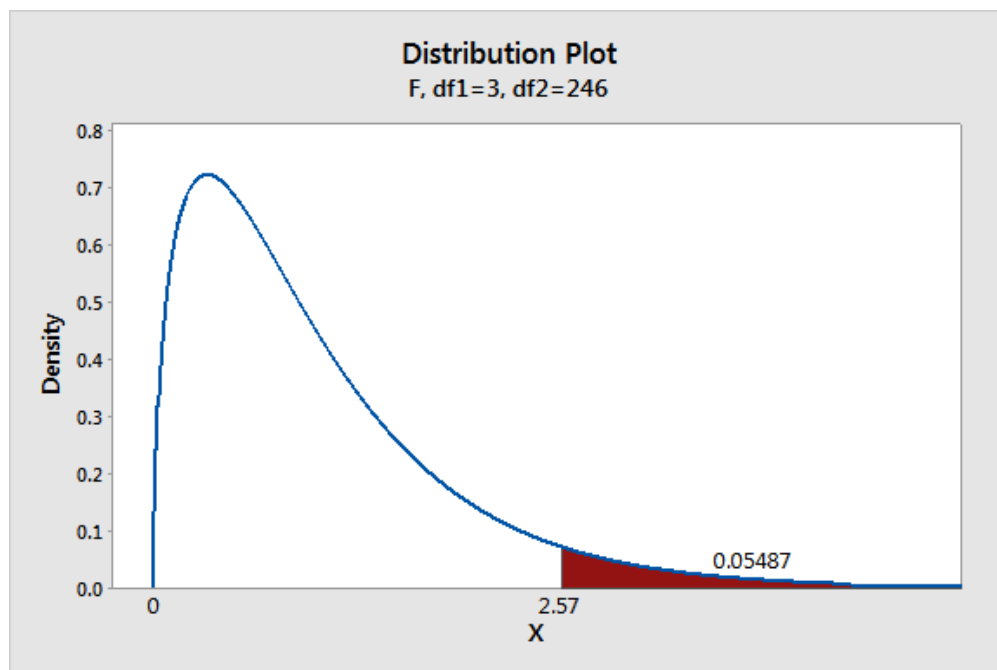


Figure 10.1: An example for p -value for F-statistic value 2.57 with an $F_{3,246}$ distribution

Now, open the `Lecture10_to_fill.Rmd` file and start working on the following 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
7. Test the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$