# Chapter 4

# Multiple Regression Analysis - Inference

Jim Bang

## Occupation and Wages

1. Re-estimate `wage.lm7`, which from the previous tutorial regressed wage on education, experience, experience squared, job tenure, and occupation type.

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 tenure +$$
$$\beta_5 profocc + \beta_6 clerocc + \beta_7 servocc$$

2. Summarize the result.

```
wage.lm7 <- lm(wage ~ educ + exper + I(exper^2) + tenure + profocc + clerocc + servocc, data = wage1)
summary(wage.lm7)
```

```
##
## Call:
## lm(formula = wage ~ educ + exper + I(exper^2) + tenure + profocc +
##     clerocc + servocc, data = wage1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.0847 -1.7196 -0.3174  1.0726 13.4667
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.378355   0.757306  -1.820   0.0693 .
## educ         0.393551   0.057420   6.854 2.05e-11 ***
## exper        0.181219   0.035033   5.173 3.30e-07 ***
## I(exper^2)  -0.003865   0.000759  -5.092 4.97e-07 ***
## tenure       0.150322   0.020402   7.368 6.88e-13 ***
```

```
## profocc      1.512422   0.354797   4.263 2.40e-05 ***
## clerocc     -0.674670   0.387684  -1.740   0.0824 .
## servocc     -0.946641   0.408266  -2.319   0.0208 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.887 on 518 degrees of freedom
## Multiple R-squared:  0.3969, Adjusted R-squared:  0.3888
## F-statistic: 48.71 on 7 and 518 DF,  p-value: < 2.2e-16
```

**Question**

- The p-Value to test whether the wages of clerical occupations differ from the baseline group (manufacturing occupations) is:
  a. 0.0208
  b. 0.0824
  c. 0.9176
  d. 0.3876

## Testing Linear Combinations of Multiple Coefficients

Suppose we want to test whether clerical and service occupations differ *from each other*. One way to write this is:

$$H_0 : \beta_6 = \beta_7$$
$$H_1 : \beta_6 \neq \beta_7$$

Another way to write this is:

$$H_0 : \beta_6 - \beta_7 = 0$$
$$H_1 : \beta_6 - \beta_7 \neq 0$$

Rearranging them this way has a purpose: it places all unknown parameters on the left and numerical constants on the right.

### Sampling distribution of $\beta_j - \beta_l$

Under $H_0$, $\beta_6 - \beta_7 = 0$

If the OLS assumptions hold, then under $H_0$,

1. $E(\hat{\beta}_6 - \hat{\beta}_7) = 0$
2. $Var(\hat{\beta}_6 - \hat{\beta}_7) = Var(\hat{\beta}_6) + Var(\hat{\beta}_7) - 2Cov(\hat{\beta}_6, \hat{\beta}_7)$ $se_{\hat{\beta}_6 - \hat{\beta}_7} = \sqrt{Var(\hat{\beta}_6) + Var(\hat{\beta}_7) - 2Cov(\hat{\beta}_6, \hat{\beta}_7)}$

Since the parameter estimates are normally distributed, and the variances are $\chi^2$, the test statistic,

$$t_{\hat{\beta}_6 - \hat{\beta}_7} = \frac{(\hat{\beta}_6 - \hat{\beta}_7) - 0}{s_{\hat{\beta}_6 - \hat{\beta}_7}} \sim t(n - k - 1)$$

### Comparing Two Occupational Groups' Wages

Use the `glht` function from the `multcomp` package (preloaded with this tutorial) to test the hypothesis that clerical occupations have the same wages as service occupations against the alternative that they differ. Pipe the test to a `summary` to give the full table of test statistics.

Use the `lht` function from the `car` package (also preloaded). You do not need to `summary()` the results to get the output you want to see.

```
glht(wage.lm7, linfct = "clerocc - servocc = 0") |>
  summary()
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = wage ~ educ + exper + I(exper^2) + tenure + profocc +
##     clerocc + servocc, data = wage1)
##
## Linear Hypotheses:
##                        Estimate Std. Error t value Pr(>|t|)
## clerocc - servocc == 0   0.2720     0.4633   0.587    0.557
## (Adjusted p values reported -- single-step method)
```

```
lht(wage.lm7, "clerocc - servocc",  0)
```

```
## Linear hypothesis test
##
## Hypothesis:
## clerocc - servocc = 0
##
## Model 1: restricted model
## Model 2: wage ~ educ + exper + I(exper^2) + tenure + profocc + clerocc +
##     servocc
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    519 4321.0
## 2    518 4318.1  1    2.8727 0.3446 0.5574
```

Notice that the test statistic for `lht` is exactly the value of that for `glht` *squared*. That's because `lht` uses an F-Test for everything, and the F-distribution (ratio of two Chi-Squareds) is exactly the same as the distribution of the square of a t-distributed random variable (ratio of a normal and the square root of a Chi-Squared). `glht` is cleaner for single restrictions; `lht` is more flexible.

Warning: "The Difference Between 'Significant' and 'Insignificant' is not Itself Statistically Significant" - Andrew Gelman

## Testing Joint Significance of Multiple Coefficients

We want to test the hypothesis that all of the occupational groups have the same wage against the alternative that *at least* one group has different wages.

This involves (a vector of) multiple restrictions and we cannot test this hypothesis using a simple t-Test. We must use an F-test (like you might have if you have studied ANoVA).

$$H_0 : \begin{pmatrix} \beta_{profocc} \\ \beta_{servocc} \\ \beta_{clerocc} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

$$H_1 : \begin{pmatrix} \beta_{profocc} \\ \beta_{servocc} \\ \beta_{clerocc} \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

### Calculating the F-Statitic

There are a couple of different (but equivalent) ways to calculate the F-Statistic for the joint significance of a group of variables. They involve the same basic steps.

1. Estimate the unrestricted model (the full model) and store the $R^2$ *or* the residual sum of squares (SSR).
2. Estimate the restricted model (which *excludes* the variables you wish to test) and store the $R^2$ or SSR.
3. Calculate the F statistic:

a. Formula using SSR:

$$F = \frac{(SSR_r - SSR_u)/q}{SSR_u/(n-k-1)}$$

b. Formula using $R^2$:

$$F = \frac{(R_u^2 - R_r^2)/q}{(1 - R_u^2/(n-k-1)}$$

c. These formulas are equivalent since

$$R^2 = 1 - \frac{SSR}{SST}$$

### Differences in Wages among All Occupational Groups

Test the joint significance of occupational choice on wages "by hand" using the $R^2$ formula:

1. Estimate the unrestricted model, `wage.lm7` and the restricted model (without occupations), `wage.lm8`.
2. Extract the R-squareds as `r2.u` and `r2.r`.
3. Compute the F statistic using the equation from 3(b) above and name it `wage.F.occ`.
4. Calculate and print the p-value.

```
wage.lm7 <- lm(wage ~ educ + exper + I(exper^2) + tenure + profocc + clerocc + servocc, data = wage1)
wage.lm8 <- lm(wage ~ educ + exper + I(exper^2) + tenure, data = wage1)
r2.u <- summary(wage.lm7)$r.squared
r2.r <- summary(wage.lm8)$r.squared
wage.F.occ <- ((r2.u-r2.r)/(wage.lm8$df.residual - wage.lm7$df.residual)) / ((1-r2.u)/wage.lm7$df.residual)
1 - pf(wage.F.occ, wage.lm8$df.residual -wage.lm7$df.residual, wage.lm7$df.residual)
```

```
## [1] 1.152133e-09
```

Note that the numerator degrees of freedom equal the residual degrees of freedom for the restricted model minus the residual degrees of freedom for the unrestricted model; the denominator degrees of freedom are the residual degrees of freedom for the unrestricted model.

### Warning about Independent t-Tests and a Note on the *Regression F-Statistic*

- Suppose we test these three parameters independently and reject just one of them at
  $alpha$
  $le$0.05 (we do not know the actual p-value or exact t-statistic, only the result of the test for some reason). What is the probability of a type I error if we use this method to test joint significance?
  a. 0.000125
  b. 0.05
  c. 0.10
  d. 0.142625

The *regression F-Statistic* (the F-Statistic reported when you `summary()` a model) is simply a joint significance test where the number of restrictions is all of the independent variables. The restricted model is the unconditional mean (regression on a constant). Its formula is:

$$F = \frac{SSR/k}{SST/(n-k-1)}$$