Lecture 11: Regression: Inference and Hypothesis Testing

Thomas Chadefaux

Quantitative Methods I

1. Hypothesis testing
2. Goodness of fit

Hypothesis testing

Standard error of the coefficients

- We now have the coefficients, but we would like some measure of uncertainty
  - confidence intervals
  - significance levels (p-values)
- Just like we needed the SE for the CI of the mean, we need the standard error of our coefficients for their CI.

Standard error of the coefficients

For the simple linear model:

$$SE_{b_1} = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}},$$

where $s = \sqrt{\frac{\sum e_i^2}{n-2}}$ is the standard deviation about the line.
More generally, in matrix notation:

$$SE_{b_1} = \sigma^2 (X'X)^{-1}$$

Hypothesis testing of the coefficients (slopes)

To test:

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

calculate the t-statistic:

$$t = \frac{b_1}{SE_{b_1}}$$

Confidence interval of the coefficients

$$b_1 \pm t^* SE_{b_1}$$

where t^* is the critical value of the t distribution with $N - 2$ degrees of freedom

An example

Example from Moore et al., p. 584

x	0.543	0.797	1.03
y	-23.4	17.8	67.3

1. Calculate the coefficients

```
x <- c(0.543, 0.797, 1.03)
y <- c(-23.4, 17.8, 67.3)
lm1 <- lm(y ~ x)
coefficients(lm1)
```

```
## (Intercept)           x
##   -126.2802    185.8821
```

An Example

2. Calculate $s = \sqrt{\frac{\sum e_i^2}{n-2}}$

```
sqrt(sum(resid(lm1)^2) / (3-2))
```

```
## [1] 4.983612
```

3. Calculate SE_{b_1}

$$SE_{b_1} = \frac{s}{\sqrt{\sum x_i - \bar{x}}}$$

$$= \frac{4.983612}{\sqrt{(0.543 - 0.79)^2 + (0.797 - 0.79)^2 + (1.03 - 0.79)^2}}$$

$$= \frac{4.983612}{\sqrt{0.118658}}$$

$$= 14.47$$

An Example

4. Calculate t

$$t = \frac{b_1}{SE_{b_1}} = \frac{185.8821}{14.47} = 12.85$$

5. Find the p-value

```
2*(1-pt(12.85, 1))
```

```
## [1] 0.04944275
```

An Example

6. Check with R:

```
summary(lm1)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      1      2      3
##  1.946 -4.068  2.122
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -126.28      11.79  -10.71   0.0592 .
## x             185.88      14.47   12.85   0.0494 *
```

An Example

7. Calculate the confidence interval of b_1

```
# Find the critical $t^*$ value:
tstar <- qt(0.975, df=1)

# Calculate the upper and lower value of the confidence
c(185.88 + tstar * 14.47, 185.88 - tstar * 14.47)
```

```
## [1] 369.738783   2.021217
```

```
# Check with R
confint(lm1)
```

```
##                   2.5 %    97.5 %
## (Intercept) -276.035420  23.47501
## x              2.054169 369.71006
```

Testing multiple coefficients: the F-test

- F-test:
  - H_0: J of the K coefficients are equal to zero ($J < K$).
  - H_a: at least one of these J coefficients is not equal to zero.
- F-test: compare:
  - the sum of squared residuals of the full model
  - the sum of squared residuals of the restricted model
- i.e., we want to compare the residual sum of squares of two models:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \quad (1)$$

$$y = \beta_0 + \beta_1 + \varepsilon \quad (2)$$

Testing multiple coefficients: the F-test

- Let:
  - $RSS_{restricted}$ be the residual sum of squares of (2)
  - $RSS_{unrestricted}$ be the residual sum of squares of (1).
- Then the following statistic follows an F distribution with J and N-K degrees of freedom:

$$\frac{(RSS_{restricted} - RSS_{unrestricted})/J}{RSS_{unrestricted}/(N-K)}$$

F-test: An example

Suppose

$$y = \begin{pmatrix} 5 \\ 7 \\ 12 \\ 2 \end{pmatrix} \text{ and } x_1 = \begin{pmatrix} 3 \\ 2 \\ 5 \\ 3 \end{pmatrix} \text{ and } x_2 = \begin{pmatrix} 5 \\ 9 \\ 1 \\ 2 \end{pmatrix}$$

- The sum of squared residuals for the model $y = \beta_0 + \beta_1 x_1 + + \beta_1 x_2 + \varepsilon$ is 5.36 (in R: sum(resid(lm1)^2))
- The sum of squared residuals for the model $y = \beta_0 + \varepsilon$ is 53. The question then is how bad is 53 compared to 5.36?

F-test: An example

Let us calculate the F statistic:

$$\frac{(53 - 5.36)/2}{5.36/(4-3)} = 4.44$$

Now we look up 4.44 in a table of an F distribution with 2 and 1 degrees of freedom and get a p-value of 0.318. I.e., we fail to reject the null hypothesis that $\beta_1 = \beta_2 = \ldots = 0$. $\}$

Why do we care about F? Here are examples:

- First, suppose we are interested in who voted and who did not. We also have a series of dummy variables that measure whereas an individual identifies herself as Catholic, Protestant, Jewish, Muslim, or affiliated with another religion.
  - We could run a regression with each dummy variable to see the rate at which each group votes. But the coefficients will always be in comparison to the omitted category, which may not be that useful.
  - Most likely, we are interested in whether there is any difference between any of the groups. We can do that by testing the null hypothesis that all of the religion coefficients are equal to 0.

Why do we care about F? Here are examples (cont'd):

- We might also be interested in interaction terms (more on this in the next lecture). We can only rule them out if all of their constituent coefficients are equal to zero.
- Policies might be significant, but only together.
- Your F statistic might be significant, yet none of your coefficient is statistically different from 0. This could be because of multicollinearity (more on this later).

Goodness of fit

R^2

- Now that we have estimated the coefficient parameters, we want to ask how well the estimated regression line fits the observations.
- One way to do this is to calculate the sum of squared residuals. I.e., for each observation, we calculate $\hat{y} - y$, square it, and sum these up.
- But we would like a more meaningful estimate of how good our line is fitting the points. In particular, we want to know what % of the variation in y is explained by the line.

R^2

- First let's think about the total variation in y. That's the sum of squared deviations from the mean, i.e., $\sum_i (y_i - \bar{y})^2$.
- Now, we know that some of the variation is NOT explained by our line. What portion? Our error term. That is, $\sum_i (\hat{y}_i - y)^2 = e_i^2$.
- So the portion of the variation in y that is not explained by our regression line is:

$$\frac{\sum e_i^2}{\sum_i (y_i - \bar{y})^2}$$

- And since we want to know how much IS explained by our regression line, we subtract this from 1. I.e.,

$$R^2 = 1 - \frac{\sum e_i^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{Var[e]}{Var[y]} = \frac{\text{Regression variation}}{\text{Total Variation}}$$

R^2

R^2 should almost never be your guide to select your regression. Why?

- R^2 never decreases as you add variables, so it would be tempting to keep adding them. Remember, however, that this is at the cost of lower t values and hence less significant coefficients
- A 'high' R^2 is dependent on the question you are asking.

R^2 in R

```
x <- rnorm(100)
y <- x + rnorm(100)
lm1 <- lm(y ~ x)

total.var.in.y <- sum((y - mean(y))^2)
variation.not.explained.by.x <- sum(residuals(lm1)^2)
R2 <- 1-( variation.not.explained.by.x / total.var.in.y)
R2
```

```
## [1] 0.5675356
```

```
#let's check:
summary(lm1)$r.squared
```

```
## [1] 0.5675356
```

AIC, BIC

There are other measures of fit that penalises a large number of parameters more heavily. In particular, AIC and BIC are important:

$$AIC = log\frac{1}{N}\sum e_i^2 + \frac{2K}{N}$$
$$BIC = log\frac{1}{N}\sum e_i^2 + \frac{K}{N}log(N)$$

Models with a *lower* AIC or BIC are preferred.