Lesson 14. Inference for Multiple Linear Regression – Part 1

Note. In Part 2 of this lesson, you can run the R code that generates the outputs in here Part 1.

1 Overview

- Recall the multiple linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \cdots \beta_k X_k + \varepsilon \quad \text{where} \quad \varepsilon \sim N(0, \sigma_\varepsilon^2)$$

- This is a population-level model
- We want to **infer** something about the population based on our sample
- Many of these upcoming inference topics will be familiar
    - We have seen them before in the context of simple linear regression

2 t-tests for coefficients

- Question: **Is an individual explanatory variable X_i helpful to include in the model, if the other explanatory variables are still there?**
- In other words: after we account for the effects of all the other predictors, does the predictor of interest X_i have a significant association with Y?
- Formal steps:

1. State the hypotheses:

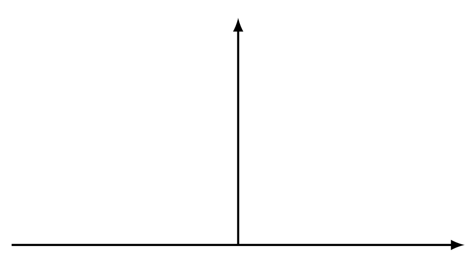
$$H_0 : \beta_i = 0$$
$$H_A : \beta_i \neq 0$$

2. Calculate the test statistic:

$$t = \frac{\hat{\beta}_i}{SE_{\hat{\beta}_i}}$$

3. Calculate the p-value:
    - If the conditions for multiple linear regression hold, then the sampling distribution of the test statistic under the null hypothesis is the t-distribution with

      degrees of freedom



4. State your conclusion, based on the given significance level α:

   If we reject H_0 (p-value $\leq \alpha$):

   > We see evidence that, after accounting for the other explanatory variables, X_i is significantly associated with Y.

   If we fail to reject H_0 (p-value $> \alpha$):

   > We do not see evidence that X_i is significantly associated with Y after accounting for the other explanatory variables.

   The highlighted parts above should be rephrased to correspond to the context of the problem

Example 1. After accounting for the size of a house, is its price related to its proximity to bike trails?

Use the `RailsTrails` data in the `Stat2Data` package to fit a multiple linear regression model predicting *Price2014* (price in thousands of dollars) from *SquareFeet* (size of house, in thousands of ft^2) and *Distance* (miles to nearest bike trail). Assume that the regression conditions are met.

We run the following R code:

```
fit <- lm(Price2014 ~ SquareFeet + Distance, data = RailsTrails)
summary(fit)
```

We obtain the following output:

```
Call:
lm(formula = Price2014 ~ SquareFeet + Distance, data = RailsTrails)

Residuals:
    Min      1Q  Median      3Q     Max
-152.15  -30.27   -4.14   25.75  337.93

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   78.985     25.607   3.085  0.00263 **
SquareFeet   147.920     12.765  11.588  < 2e-16 ***
Distance     -15.788      7.586  -2.081  0.03994 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 65.55 on 101 degrees of freedom
Multiple R-squared:  0.6574,	Adjusted R-squared:  0.6506
F-statistic: 96.89 on 2 and 101 DF,  p-value: < 2.2e-16
```

a. State the population-level model.

b. State the fitted model.

c. What do we learn from the estimated coefficient of *Distance*?

d. Is the association between *Distance* and *Price2014* statistically significant, after accounting for house size? Use a significance level of 0.05 to test whether the coefficient of *Distance* is 0. (Report the relevant values from the summary output.)

3 Confidence intervals for coefficients

- Goal: **We want to provide a range of plausible values for β_i, instead of just a point estimate**
- Formula:
  - If the conditions for multiple linear regression are met, then we can form a CI for β_i with the following formula

$$\hat{\beta}_i \pm t_{\alpha/2, n-(k+1)} SE_{\hat{\beta}_i}$$

- Interpretation:

  We are 95% confident that the true coefficient of X_i is between lower endpoint of CI and upper endpoint of CI.

- Taking the interpretation even further:

  We are 95% confident that, holding the other explanatory variables constant, a one unit increase in X_i is associated with an average decrease/increase of between smaller magnitude of CI and larger magnitude of CI units in the response variable.

- The highlighted parts above should be rephrased to correspond to the context of the problem

Example 2. Continuing with Example 1...

a. Based on the reported degrees of freedom for the residual standard error, what must n (the number of observations) be?

b. Use the R output to form a 95% confidence interval for the coefficient of *Distance*.
Note that $t_{0.05/2,101}$ = `qt(1 - 0.05/2, df = 101)` = 1.984.

c. Interpret your CI in the context of this problem.

4 ANOVA for multiple linear regression

- In addition to testing the individual explanatory variables one-by-one, we could also ask...
- Question: **Is the model as a whole effective?**
- In other words: is the model with all the explanatory variables better than a model with none of the explanatory variables?
- To answer this question, we return to the idea of partitioning variability:

$$SSTotal \quad = \quad SSModel \quad + \quad SSE$$

where

$$SSTotal = \sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad SSModel = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \qquad SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

4.1 ANOVA table for multiple regression

Source	DF	Sum of Squares	Mean Square	F-Statistic
Model		$SSModel$	$MSModel =$	$F = \frac{MSModel}{MSE}$
Error		SSE	$MSE =$	
Total		$SSTotal$		

- Unfortunately, the R function `anova()` does not create the above ANOVA table
- Instead, we can use the following R code to fill in the blanks of the above ANOVA table:

```
y <- RailsTrails$Price2014
n <- 104
k <- 2

SSModel <- sum( (predict(fit) - mean(y))^2 )
SSE <- sum( (y - predict(fit))^2 )
SSTotal <- SSModel + SSE

MSModel <- SSModel / k
MSE <- SSE / (n - (k + 1))

F <- MSModel / MSE
```

4.2 ANOVA test steps

1. State the hypotheses:

   Note that the alternative is not that every predictor has a non-zero coefficient

2. Calculate the test statistic:

$$F = \frac{MSModel}{MSE}$$

3. Calculate the p-value:

   - If the conditions for multiple linear regression hold, then the sampling distribution of the test statistic under the null hypothesis is the F-distribution with

     degrees of freedom

4. State your conclusion, based on the given significance level α:

   If we reject H_0 **(**p**-value** $\leq \alpha$**):**

   We see significant evidence that the model as a whole is effective.

   If we fail to reject H_0 **(**p**-value** $> \alpha$**):**

   We do not see sufficient evidence to conclude that the model is effective.

   The underlined parts above should be rephrased to correspond to the context of the problem

Example 3. Continuing Examples 1 and 2...

Use the output in Example 1 to perform an ANOVA test that determines whether the multiple linear regression model that uses *SquareFeet* and *Distance* to predict *Price2014* is effective as a whole.