Lecture 26

Part 5 Linear Regression

Multiple Linear Regression

4. Calculating R^2

- Since we now have more than one independent variable, we cannot define the coefficient of determination, R^2, to be the square of the correlation coefficient.
- Instead, we use the other definition of R^2 given in the previous topic:

$$R^2 = \frac{SSR}{SS(Total)}$$

- It still represents the proportion of total variation in Y that is explained by the model.

4. Calculating R^2

- We can also express R^2 as:

$$R^2 = \frac{SSR}{SS(Total)}$$
$$= \frac{SS(Total) - SSE}{SS(Total)}$$
$$= 1 - \frac{SSE}{SS(Total)}$$

- Problem with R^2 in multiple linear regression: It will always increase as we add more independent variables to the model, even if they are not related to Y!

4. Calculating R^2

- To deal with this, we often use the **adjusted** R^2, defined as:

$$\text{adjusted } R^2 = 1 - \frac{\frac{SSE}{n-k-1}}{\frac{SS(Total)}{n-1}}$$

- If we use adjusted R^2, it tends to be smaller than R^2 when we have added independent variables which are not related to Y.

Example

```
Call:
lm(formula = attitude ~ duration + weather, data = city.dat)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.45755    0.94094   0.486  0.639817
duration    0.46751    0.08907   5.249  0.000775
weather     0.26344    0.11784   2.236  0.055810

Residual standard error: 1.243 on 8 degrees of freedom
Multiple R-squared:  0.8724,    Adjusted R-squared: 0.8405
F-statistic: 27.35 on 2 and 8 DF,  p-value: 0.0002649
```

- From the regression output, $R^2 = 0.8724$ and adjusted $R^2 = 0.8405$.

Example

```
Analysis of Variance Table
Response: attitude
            Df  Sum Sq  Mean Sq  F value    Pr(>F)
Regression   2 84.4586 42.2293  27.3538  0.0002649
Residuals    8 12.3505  1.5438
Total       10 96.8091
```

- From the ANOVA table:

$$\text{adjusted } R^2 = 1 - \frac{\frac{SSE}{n-k-1}}{\frac{SS(Total)}{n-1}} = 1 - \frac{\frac{12.3505}{8}}{\frac{96.8091}{10}} = 0.8405$$

Using the Model

- Using the estimated model, we can obtain a point estimate to predict the value of Y for a new observation from our population in the natural way:

$$\hat{y}_g = \hat{\beta}_0 + \hat{\beta}_1 x_{1g} + \cdots + \hat{\beta}_k x_{kg}$$

- We will rely on software to calculate confidence intervals for a particular value of Y and for the expected value of Y.

Multicollinearity

- Because it causes the parameter estimates of the correlated independent variables to become unstable and have large standard errors (i.e., large $s_{\hat{\beta}_j}$).
- To illustrate, let's consider a model with two independent variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

- Now let's also suppose that X_1 and X_2 are perfectly correlated with each other.

Multicollinearity

- In fact, let's assume that $X_1 = X_2$ (a very extreme situation).
- Consider the following two estimated models:

$$\text{Model 1: } \hat{Y} = 2 + 100X_1 + 2X_2$$
$$\text{Model 2: } \hat{Y} = 2 + 2X_1 + 100X_2$$

What do you notice about the two models?

Multicollinearity

- They are actually the same and are both equal to:

$$\hat{Y} = 2 + 102X_1$$

- So two very different pairs of parameter estimates can result in the exact same model.
- This leads to huge variability in both $\hat{\beta}_1$ and $\hat{\beta}_2$, meaning $s_{\hat{\beta}_1}$ and $s_{\hat{\beta}_2}$ will both become very large.

Multicollinearity

- This will affect the T-statistics for the tests of these individual coefficients:

$$T_j = \frac{\hat{\beta}_j}{s_{\hat{\beta}_j}}$$

- We can end up making the wrong conclusion.
- That is, based on tests using the T-statistics, we might decide that there is no linear relationship between Y and a particular X_j, when in fact there is.

Multicollinearity

- Fortunately, multicollinearity does not affect the F-statistic for testing the overall significance of the model.
- Multicollinearity is difficult to deal with - one way is to try to only use independent variables that are uncorrelated with each other.

Multiple Linear Regression with Categorical Independent Variables

- Categorical independent variables can be incorporated into a multiple linear regression model by coding them as *indicator* or *dummy* variables.
- Recall that an indicator variable only takes two values (usually 0 and 1), where a 1 indicates the existence of a condition and 0 indicates the absence of the condition.

Example

- We have collected house prices for a sample of 50 houses and want to examine whether a linear relationship exists between price (Y) and two independent variables, house size in square metres (X), and whether or not the house has a pool.
- Both price and size are clearly continuous variables.

Example

- To include the categorical variable reflecting whether or not the house has a pool, we define the indicator variable W as follows:

$$W = \begin{cases} 1 & \text{if the house has a pool} \\ 0 & \text{if the house does not have a pool} \end{cases}$$

- Now, let's say that you had initially decided to fit the following model:

$$Y = \beta_0 + \beta_1 X + \beta_2 W + \epsilon$$

Example

- For houses with a pool, $W = 1$ and the model takes the form:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + \beta_2 \times 1 + \epsilon \\ &= (\beta_0 + \beta_2) + \beta_1 X + \epsilon \end{aligned}$$

- For houses without a pool, $W = 0$ and the model takes the form:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + \beta_2 \times 0 + \epsilon \\ &= \beta_0 + \beta_1 X + \epsilon \end{aligned}$$

Example

- That is, the model we initially specified:

$$Y = \beta_0 + \beta_1 X + \beta_2 W + \epsilon$$

  allows the intercept to vary depending on whether or not the house has a pool.
- But the slope, which reflects the relationship between price and house size, remains the same.
- But what if the relationship between price (Y) and size (X) depends on whether the house has a pool?

Example

- To test whether the presence of a pool requires a different intercept and/or a different slope for size, we need to fit the following model:

$$Y = \beta_0 + \beta_1 X + \beta_2 W + \beta_3 (X \times W) + \epsilon$$

- The above model explicitly includes the *interaction* between size and whether or not the house has a pool.

Example

- For houses with a pool, $W = 1$ and the model takes the form:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + \beta_2 \times 1 + \beta_3 (X \times 1) + \epsilon \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X + \epsilon \end{aligned}$$

- For houses without a pool, $W = 0$ and the model takes the form:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + \beta_2 \times 0 + \beta_3 (X \times 0) + \epsilon \\ &= \beta_0 + \beta_1 X + \epsilon \end{aligned}$$

Example

- This model allows both the intercept and the slope of size to change depending on whether the house has a pool.
- Testing whether $\beta_2 = 0$ tests whether different intercepts are required, and testing whether $\beta_3 = 0$ tests whether different slopes are required.

More than Two Categories

- A categorical variable with k categories can be coded with $k-1$ indicator variables (one for each of the first $k-1$ categories).
- Suppose $k=3$ (e.g., no garage, single garage, double garage).

	Garage		
	None	Single	Double
W_1	1	0	0
W_2	0	1	0
W_3	0	0	1

More than Two Categories

- But we don't need all three indicator variables.

	Garage		
	None	Single	Double
W_1	1	0	0
W_2	0	1	0
$\cancel{W_3}$	~~0~~	~~0~~	~~1~~

- W_1 and W_2 uniquely identify all three categories.
- In fact, $W_3 = 1 - (W_1 + W_2)$.