Lecture 25

Part 5 Linear Regression

Multiple Linear Regression

R Output

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

Residuals:
Min        1Q   Median       3Q      Max
-1.56859 -0.79732  0.03449  0.47779  1.82480


Coefficients:
Estimate                Std. Error t stat. 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
```

Assessing the Model

- We can use the same approaches used for simple linear regression to assess our multiple linear regression model:

1. Check to see if the model assumptions hold.
2. Test the overall significance of the model.
3. Estimate σ_ϵ^2, the variance of the error variable.
4. Calculate R^2, the proportion of variation in Y explained by the model.

1. Checking the Model Assumptions

- Similar to simple linear regression, we check to see if the residuals e_i satisfy the model assumptions:

  (a). Are they normally distributed?

  - Check histograms (normal shape) and normal probability plots (linear).

  (b). Do they have mean 0 and constant variance?

  - Check scatter plots of residuals against *fitted values* (should be random noise around 0 with no patterns).

  (c). Are they independent?

  - Check plots of residuals against collection order (should be no trends or patterns).

2. Testing Overall Significance of Model

- Recall that for simple linear regression, the following hypotheses were used to test the overall significance of the model:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

- If H_0 is true, then X drops out of the model, but if H_1 is true, then X is linearly related to Y.

2. Testing Overall Significance of Model

- For a multiple linear regression model, the following hypotheses must be used to test the overall significance of the model:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

H_1 : Not all coefficient parameters are equal to 0.

- If H_0 is true, then all independent variables drop out of the model, but if H_1 is true, at least one of them is linearly related to Y.

2. Testing Overall Significance of Model

- If we fit a multiple linear regression model, but H_0 is true, then that model will probably not explain much of the variation in Y.
- If we fit a multiple linear regression model, and H_1 is true, then that model will most likely be able to explain a reasonable amount of variation in Y.
- How do we measure sources and amounts of variation?

Sums of Squares

- Total sum of squares:

$$SS(\text{Total}) = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

- Sum of squares for regression:

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

- Sum of squares for error:

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Sums of Squares

- Formulae are exactly the same as for simple linear regression, with the only difference being in how the fitted values $\hat{Y}_i$ are calculated.
- And the same identity relating the sums of squares still holds:

$$SS(\text{Total}) = SSR + SSE$$

ANOVA Table for Regression

- So to test the overall significance of the multiple linear regression model, we need to construct an ANOVA table for regression:

Source	Sum of squares	Deg. of freedom	Mean squares	F-statistic
Regression	SSR	k	$MSR = \frac{SSR}{k}$	$F = \frac{MSR}{MSE}$
Error (Residual)	SSE	$n-k-1$	$MSE = \frac{SSE}{n-k-1}$	
Total	$SS(\text{Total})$	$n-1$		

Test Statistic

- So we reject H_0 if the model explains a large amount of the variation in Y.
- That is, we reject when the MSR is large, compared to the MSE.
- Just like in ANOVA, the test statistic is the F-statistic:

$$F = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}} = \frac{MSR}{MSE}$$

Decision Rule

- We compare the F-statistic to an F-distribution with k numerator degrees of freedom and $n-k-1$ denominator degrees of freedom.
- At a significance level of α, we reject H_0 if $F > F_{\alpha,k,n-k-1}$.
- Note that if we reject H_0, we are concluding that at least one of the coefficient parameters is not equal to 0.
- But we then need to do some additional tests to determine *which* coefficients are not equal to 0.

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			

- We can test the overall significance of the model from the ANOVA section of the R output.
- Since the p-value of 0.0002649 is very small, we reject H_0 and we conclude that at least one coefficient parameter is not equal to 0.

Testing Individual Coefficient Parameters

- If we reject H_0 and conclude that at least one coefficient parameter is not equal to 0, we next want to test which are not equal to 0.
- For each coefficient parameter, we can test:

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

for $j = 1, \ldots, k$.

Testing Individual Coefficient Parameters

- We use the following test statistic to test these hypotheses:

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

  where $s_{\hat{\beta}_j}$ is the standard error of $\hat{\beta}_j$.
- For our decision rule, we compare the test statistic to a t-distribution with $n-k-1$ degrees of freedom and reject H_0 if $T > t_{\frac{\alpha}{2},n-k-1}$ or $T < -t_{\frac{\alpha}{2},n-k-1}$.

Testing Individual Coefficient Parameters

- Each test of an individual coefficient parameter is conditional on the fact that *all the other independent variables have been included in the model*.
  - If we reject H_0, we would conclude that, *once all the other variables have been considered*, X_j has a significant linear relationship with Y.
  - If we fail to reject H_0, we would conclude that, *once all the other variables have been considered*, X_j does not have a significant linear relationship with Y.

Testing Individual Coefficient Parameters

- The conditional dependence of each test of an individual coefficient parameter is very important to recognize and understand.
- It means that if we were to fit a simple linear regression with only X_j, we might not necessarily make the same conclusion.
- For example, we might conclude that X_j is not linearly related to Y in a multiple linear regression, but based on a simple linear regression with just X_j, we might conclude that X_j is linearly 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

- Based on the p-value of 0.000775, we would reject H_0 and conclude that, once importance of weather is considered, duration of residence still has a significant linear relationship with attitude towards city.

General Test for β_j

- General hypotheses for β_j, $j = 0, \ldots, k$:

$$H_0 : \beta_j = c$$

$$H_1 : \beta_j (\neq, <, >) c$$

- Test statistic:

$$T = \frac{\hat{\beta}_j - c}{s_{\hat{\beta}_j}}$$

- Decision rule:
  - Compare to a t-distribution with $n - k - 1$ degrees of freedom.

3. Estimating σ_ϵ^2

- For multiple linear regression, the standard error of estimate s_ϵ is defined as:

$$s_\epsilon = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-k-1}} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-k-1}}$$

- The appropriate degrees of freedom is now $n-k-1$, where k is the number of independent variables.

3. Estimating σ_ϵ^2

- Note that if $k = 1$ (simple linear regression), we get the usual $n - 2$ in the denominator.
- Since s_ϵ^2 is an estimate of σ_ϵ^2, recall that a small value of s_ϵ indicates a good model.
- But again, if you are only using s_ϵ (or s_ϵ^2) to evaluate models, it's more useful as a comparative tool.

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, $s_\epsilon = 1.243$.

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:

$$s_\epsilon = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{\frac{12.3505}{8}} = 1.243$$