STAT 302 - Chapter 3 : Multiple Regression - Part 1

Harsha Perera

Key Topics

- Multiple Linear Regression Model
- Assessing a Multiple Linear Regression Model

Multiple Linear Regression Model

- In Chapters 1 and 2, we studied simple linear regression with a single quantitative predictor. This chapter introduces the more general case of **multiple linear regression**, which allows several explanatory variables to combine in explaining a response variable.

- Now we have k explanatory variables $X_1, X_2, \cdots, X_k$. The model now assumes that the mean response μ_y for a particular set of values of the explanatory variables is a linear combination of those variables:

$$\mu_y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

- As with the simple linear regression case, an actual response value (Y) will deviate around this mean (μ_y) by some random error (ϵ).

- We assume the errors in a multiple linear regression model are **independent** of each other, have a **constant variance**and **normally distributed**. Therefore $\epsilon \sim \text{N}(0, \sigma_\epsilon)$

Multiple Linear Regression Model

THE MULTIPLE LINEAR REGRESSION MODEL

We have n observations on k explanatory variables $X_1, X_2, \ldots, X_k$ and a response variable Y. Our goal is to study or predict the behavior of Y for the given set of the explanatory variables. The multiple linear regression model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

where $\epsilon \sim N(0, \sigma_\epsilon)$ and the errors are independent from one another.

- This model has $k + 2$ unknown parameters that we must estimate from data.
- $k + 1$ coefficients $\beta_0, \beta_1, \beta_2, \cdots, \beta_k$ and the standard deviation of the error σ_ϵ.
- The estimation uses the same procedure of computing the sum of squared residuals, where the residuals are obtained as the differences between the actual Y values and the predicted values $\hat{Y}$.

Example 3.1 and 3.2 : NFL Winning Percentage

- NFL Winning Percentage : Is offense or defense more important in winning football games ? We are interested in using both these scoring variables in a model to predict the winning percentage.

- Below is some computer output for fitting the multiple linear regression model to predict the winning percentages for NFL teams based on both points scored and points allowed.

The regression equation is

WinPct = 0.785 + 0.001699 PointsFor − 0.002482 PointsAgainst

Predictor	Coef	SE Coef	T	P
Constant	0.785	0.154	5.11	0.000
PointsFor	0.001699	0.000263	6.47	0.000
PointsAgainst	-0.002482	0.000320	-7.74	0.000

S = 0.0965319 R-Sq =78.24% R-Sq(adj) = 76.74%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	0.9715	0.485728	52.13	0.000
Error	29	0.2702	0.009318		
Total	31	1.2417			

Example 3.1 and 3.2 : NFL Winning Percentage

- Interpreting the coefficients is a bit trickier in the multiple regression setting because the predictors might be related to each other. We will discuss more details about this in Section 3.5.

- For example, scoring 10 more points increases the predicted winning percentage by 0.001699(10)=0.01699 or about 1.7%. Here we assume that the points allowed did not change.

- The estimate for the standard deviation of the multiple regression model with k predictors is;

$$\hat{\sigma}_\epsilon = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{MSE}$$

- In this example $\hat{\sigma}_\epsilon = \sqrt{\frac{0.2702}{32-2-1}} = 0.0965$

- SSE, Error DF, and MSE are already given in the ANOVA output. In R output, Residual Standard Error ($\hat{\sigma}_\epsilon$) is also given.

Assessing a Multiple Linear Regression Model

- Assessing a multiple linear regression model is similar to what we have already seen for simple linear regression model (sections 1.3, 2.1, and 2.2)
- Are the conditions for the model reasonably met? Is the predictor related to the response variable? Does the model do a good job of explaining variability in the response?
- We use residual plots and normal quantile plots of residuals for checking conditions for the multiple linear regression model.
- The methods for constructing these plots and their interpretation are the same when dealing with multiple predictors.
- Example 3.3 in R ...

ANOVA for Multiple Regression

- For simple linear regression the t-test for slope and ANOVA F-test for regression are equivalent.
- In multiple linear regression model, t-tests will assess the importance of each predictor individually in the model, while the ANOVA F-test will check how the predictors do as a group. We often look at the ANOVA first ("Is the model effective ?") before checking t-tests ("Which predictors are helpful?")
- To test the effectiveness of the model as a whole we return to the idea of partitioning the variability in the data into : **SSTotal = SSModel + SSError**

ANOVA for a Multiple Regression Model

ANOVA FOR A MULTIPLE REGRESSION MODEL

To test the effectiveness of the multiple linear regression model, the hypotheses are

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$
H_a : at least one $\beta_i \neq 0$

and the ANOVA table is

Source	Degrees of Freedom	Sum of Squares	Mean Square	F-statistic
Model	k	$SSModel$	$MSModel = \frac{SSModel}{k}$	$F = \frac{MSModel}{MSE}$
Error	$n-k-1$	SSE	$MSE = \frac{SSE}{n-k-1}$	
Total	$n-1$	$SSTotal$		

If the conditions for the multiple linear regression model, including normality, hold, we compute the P-value using the upper tail of an F-distribution with k and $n-k-1$ degrees of freedom.

If the p-value is small, we conclude that one or more of the predictors is effective in the model. But the ANOVA analysis doesn't identify which predictors are significant. That is the role for the individual t-tests. Now read Example 3.6 and the R output ...

Coefficient of Multiple Determination - R^2

- Coefficient of Multiple Determination is a measure of the percentage of total variability in the response that is explained by the regression model. This concept applies equally well in the setting of multiple regression so that :

  R^2 = variability explained by the model/ Total variability in y

  R^2 = SSModel / SSTotal = 1- (SSE/SSTotal)

- Using the ANOVA table information we can calculate $R^2 = 0.9715/1.2417 = 0.782$

- That means we conclude, 78% of the variability in winning percentage of NFL teams for the 2016 regular season can be explained by the regression model based on the points scored and points allowed.

Adjusted Coefficient of Determination

- Adding a new predictor to a multiple linear regression model always **increase** the percentage of variability explained by the model regardless it is effective or not.
- But does that increase reflect important new information provided by the new predictor or just extra variability explained due to random chance ?
- To answer this we can use an **Adjusted Coefficient of Determination** which reflects the number of predictors in the model as well as the amount of variability explained.

Adjusted Coefficient of Determination

ADJUSTED COEFFICIENT OF DETERMINATION

The **adjusted** R^2, which helps account for the number of predictors in the model, is computed as

$$R^2_{adj} = 1 - \frac{SSE/(n-k-1)}{SSTotal/(n-1)} = 1 - \frac{\hat{\sigma}^2_\epsilon}{S^2_Y}$$

- Adjusted R^2 value might go down when a weak predictor is added to a model.
- Adjusted R^2 is specially useful, as we will compare models with different numbers of predictors.
- Example 3.7 ...

t-tests for Coefficients

- After doing the ANOVA test and if it is significant we need to ask the question of whether or not an individual predictor is helpful to include in the model.

- We can test this by seeing if the coefficient for the predictor is significantly different from zero.

INDIVIDUAL *t*-TESTS FOR COEFFICIENTS IN MULTIPLE REGRESSION

To test the coefficient for one of the predictors, X_i, in a multiple regression model, the hypotheses are

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

and the test statistic is

$$t = \frac{\text{parameter estimate}}{\text{standard error of estimate}} = \frac{\hat{\beta}_i}{SE_{\hat{\beta}_i}}$$

If the conditions for the multiple linear model hold (including normality of errors), we compute the *P*-value for the test statistic using a *t*-distribution with $n - k - 1$ degrees of freedom.

Example 3.4

- The **parameter estimates**, **standard errors of the estimates**, **test statistics values**, and **p-values** for t-tests for individual predictors appear in standard regression output.

The regression equation is

WinPct = 0.785 + 0.001699 PointsFor − 0.002482 PointsAgainst

Predictor	Coef	SE Coef	T	P
Constant	0.785	0.154	5.11	0.000
PointsFor	0.001699	0.000263	6.47	0.000
PointsAgainst	-0.002482	0.000320	-7.74	0.000

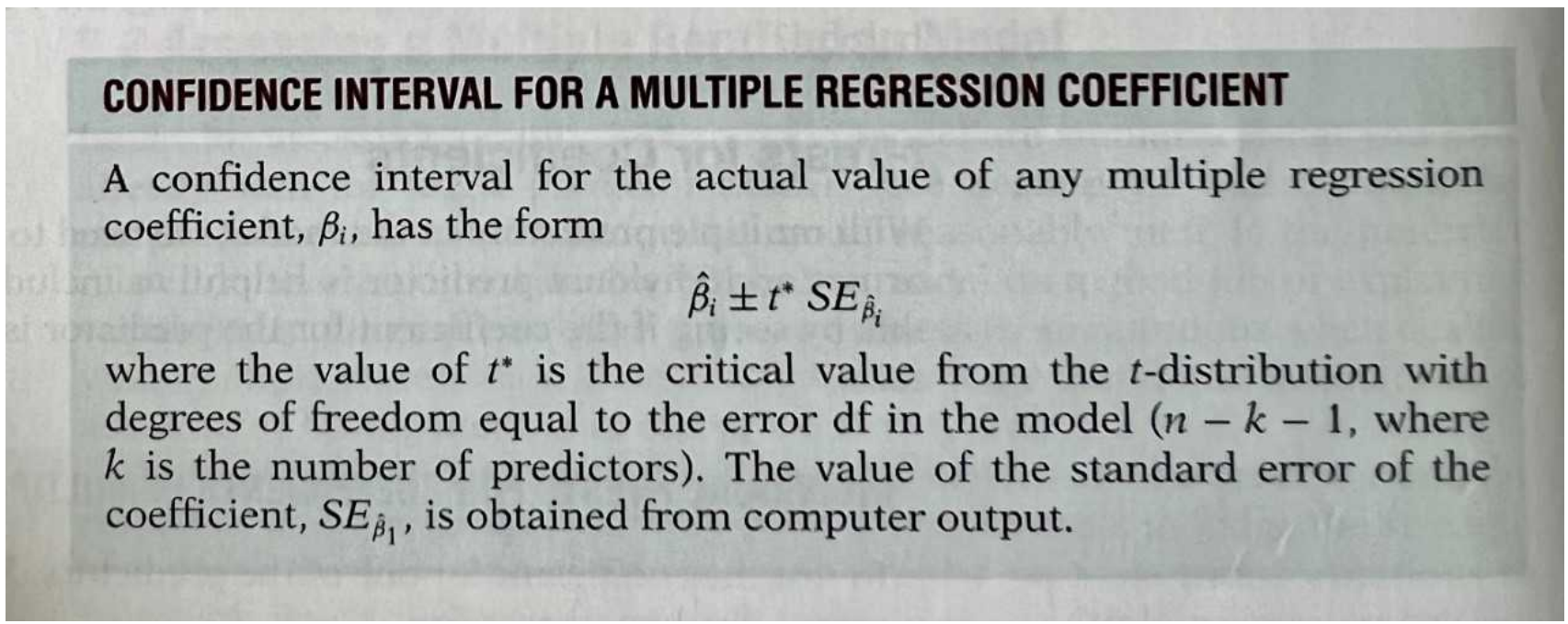
S = 0.0965319 R-Sq =78.24% R-Sq(adj) = 76.74%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	0.9715	0.485728	52.13	0.000
Error	29	0.2702	0.009318		
Total	31	1.2417			

Confidence Intervals for Coefficients

- In addition to the estimate and the hypothesis test for regression coefficients, we may be interested in producing a confidence interval for one or more of the coefficients.



CONFIDENCE INTERVAL FOR A MULTIPLE REGRESSION COEFFICIENT

A confidence interval for the actual value of any multiple regression coefficient, β_i, has the form

$$\hat{\beta}_i \pm t^* \, SE_{\hat{\beta}_i}$$

where the value of t^* is the critical value from the t-distribution with degrees of freedom equal to the error df in the model ($n - k - 1$, where k is the number of predictors). The value of the standard error of the coefficient, $SE_{\hat{\beta}_1}$, is obtained from computer output.

- Example 3.5 R output ...

Confidence and Prediction Intervals

- Just as in section 2.4, we can obtain interval estimates for the mean response or a future individual case given any combination of predictor values.
- The text book doesn't provide the details of these formulas since computing the standard errors is more complicated with multiple predictors. Therefore we rely on R to mange the calculations.
- Example 3.8 R output ...