CHAPTER 3

Multiple Regression

3.1 Multiple Linear Regression Model

3.2 Assessing a Multiple Regression Model

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

- In example Porsche price, the price (Y) of a used Porsche may depend on its mileage (X_1), and also may depend on its age (X_2).

Notice that the assumptions are the same for both simple and multiple linear regression.

3.1 Multiple Linear Regression:

We have n observations on k explanatory variables $X_1, X_2, \cdots, X_k$ 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 model is

$$\underbrace{Y}_{\text{Data}} = \underbrace{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k}_{\text{Model}} + \underbrace{\in}_{\text{Error}}$$

Where, $\in \sim N(0, \sigma_{\in})$ and the errors are independent from one another.

The 4 Step Process for Multiple Regression:

Collect data for the response and all predictors.

CHOOSE a form of the model.

Select predictors; possible transform Y.

Choose any function of predictors.

FIT estimate the coefficients $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$.

Estimate the residual standard error $\hat{\sigma}_{\in}$ (RMSE).

Assess the fit.

Test the overall fit: ANOVA, R^2.

Test individual predictors: t-test.

Examine residuals.

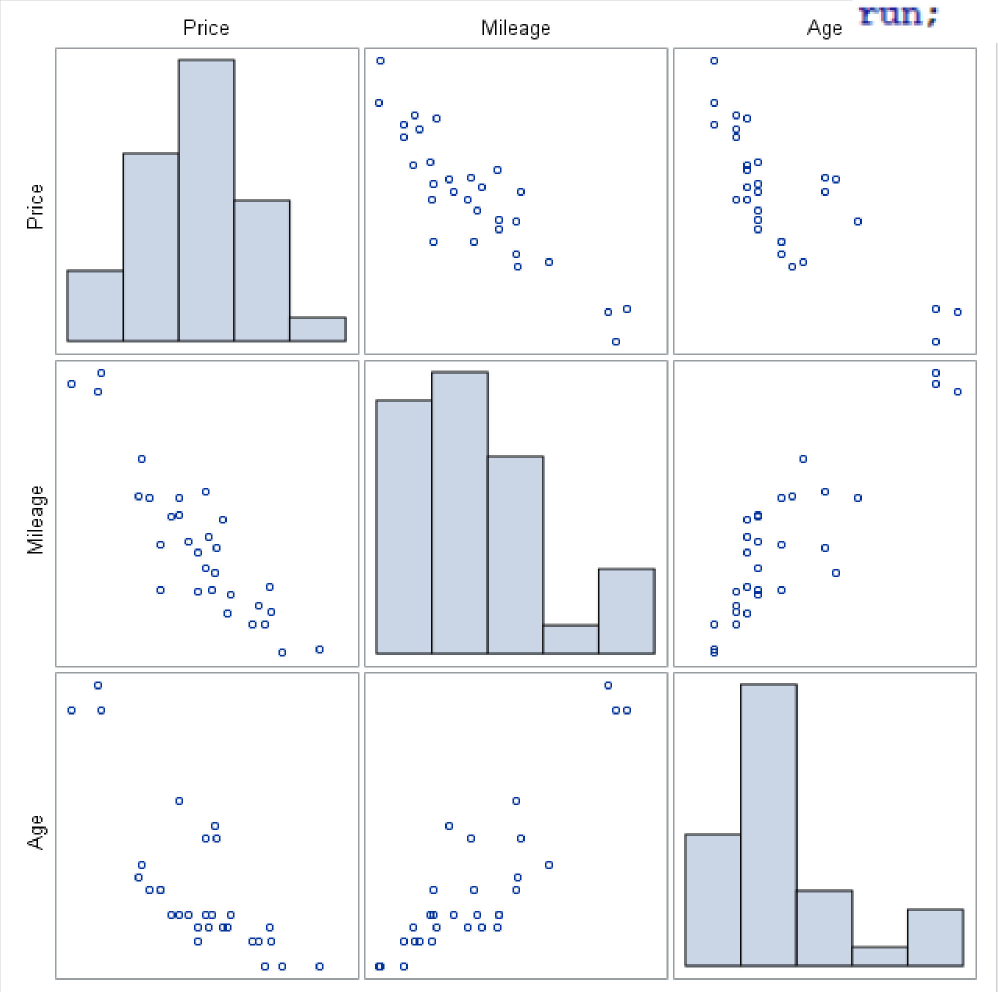
USE Predications, CI's, and PI's.

Example 1: *(Porsche prices)*

For the same dataset *Porsche prices.csv*.

1. Using SAS, graph the scatterplot of the mileage vs. price and age.

```
proc corr plots = matrix(histogram);
var price mileage age;
run;
```



2. Using SAS, Calculate and interpret the correlation coefficients.

Pearson Correlation Coefficients, N = 30 Prob > |r| under H0: Rho=0			
	Price	Mileage	Age
Price	1.00000	-0.89135 <.0001	-0.78189 <.0001
Mileage	-0.89135 <.0001	1.00000	0.86313 <.0001
Age	-0.78189 <.0001	0.86313 <.0001	1.00000

- The correlation coefficient between the price and mileage is $r = -0.89$, so there is a strong negative relationship between them.
- The correlation coefficient between the price and age is $r = -0.78$, so there is a strong negative relationship between them.

3. State your hypotheses and interpret the p-values of the correlation coefficients.

$$H_0: \rho_{Y,X_1} = 0 \quad vs \quad H_1: \rho_{Y,X_1} \neq 0$$

$$H_0: \rho_{Y,X_2} = 0 \quad vs \quad H_1: \rho_{Y,X_2} \neq 0$$

Decision: Since $p - value < 0.0001 < 0.05$ for the two predictors, so we reject H_0.

Conclusion: The correlation coefficient of the population doesn't equal to 0. Which means that there is a **significant** linear relationship between the mileage (or age) and the price.

4. Fit the regression model.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	70.91916	2.48352	28.56	<.0001
Age	1	-0.13023	0.45684	-0.29	0.7778
Mileage	1	-0.56134	0.11407	-4.92	<.0001

The multiple linear regression model is:

$$\widehat{Price} = 70.92 - 0.13\ Age - 0.56\ Mileage$$

5. Interpret the regression coefficients.

- **Intercept:** The predicted price of a new car (0 year and 0 mile) is $70,919.16.
- **Age coefficient:** For every additional 1 year, when the mileage held constant, the predicted price goes down by $130.
- **Mileage coefficient:** For every additional 1000 miles, when the age held constant, the predicted price goes down by $561.

6. What is the fitted (predicted) value of the price corresponding to 21,500 (21.5) miles and 3 years old.

$$Price = 70.91916 - 0.13023(3) - 0.56134(21.5)$$
$$= \$58.460$$

The predicted value of the price corresponding to 21,500 (21.5) miles and 3 years old is $58,460.

7. What is the residual corresponding 21,500 (21.5) miles and 3 years old.

$$residual = \$69.4 - \$58.460$$

$$= \$10.94$$

8. Using SAS, find the estimate for the standard error of the multiple regression.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	5570.00389	2785.00195	52.39	<.0001
Error	27	1435.24577	53.15725		
Corrected Total	29	7005.24967			

MSE

Root MSE	7.29090	R-Square	0.7951
Dependent Mean	50.53667	Adj R-Sq	0.7799
Coeff Var	14.42695		

$\sqrt{MSE}$

$$\hat{\sigma}_{\epsilon} = \sqrt{MSE} = \sqrt{53.15725} = 7.29090$$

or

$$\hat{\sigma}_{\epsilon} = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{\frac{1435.24577}{30-2-1}} = \sqrt{\frac{1435.24577}{27}} = 7.29090$$

3.2 Assessing a Multiple Linear Regression Model:

ANOVA for a Multiple Regression Model:

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

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

$$vs \qquad H_1: at\ least\ one\ \beta_i \neq 0$$

Source of Variation	Degrees of Freedom	Sums of Squares	Mean Squares	F
Regression	k	SSR	MSR = SSR / k	MSR/ MSE
Residual	n-k-1	SSE	MSE=SSE/(n-k-1)	
Total	n-1	SST		

Number of predictors

Example 2: *(Porsche prices)*

Interpret ANOVA table.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	5570.00389	2785.00195	52.39	<.0001
Error	27	1435.24577	53.15725		
Corrected Total	29	7005.24967			

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

$$vs \qquad H_1: at\ least\ one\ \beta_i \neq 0$$

Decision: since p-value < 0.0001 < 0.05 = α, so we reject H_0

Conclusion: at least one of the predictors, mileage and age, has a significant effect for the explaining variability in price.

The question now is:

Do both predictor variables provide significant information about the price?

If not.

Which predictor variable is providing significant information about the price?

We can answer this question by using the individual t-test.

Individual t-Test for Coefficients in Multiple Regression :

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

$$H_0: \beta_i = 0 \quad vs \quad H_1: \beta_i \neq 0, \qquad i = 1,2,\cdots,k$$

and the test statistic is

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

Example 3: *(Porsche prices)*

Test the hypotheses $(\beta_1 = \beta_{age})$

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

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	70.91916	2.48352	28.56	<.0001
Age	1	-0.13023	0.45684	-0.29	0.7778
Mileage	1	-0.56134	0.11407	-4.92	<.0001

Decision: since p-value = 0.7778 > 0.05 = α, so we fail to reject H_0.

Conclusion: we do not have an **evidence** to say that the car age has a significant effect for the explaining variability in price.

Note: we should to drop the age from the model.

Example 3: *(Porsche prices)*

The **full** model is

$$\widehat{Price} = 70.92 - 0.13\ Age - 0.56\ Mileage$$

The **reduced** model is

$$\widehat{Price} = 71.09 - 0.59\ Mileage$$

Note: if we fit a simple linear regression model between the price and the age, the relationship will be **significant**, but since the mileage by itself can fit the data well.

Confidence Interval for a Multiple Regression Coefficients:

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 t-table with degrees of freedom $= \mathrm{n} - \mathrm{k} - 1$.
The value of the standard error of the coefficient, $SE_{\hat{\beta}_i}$, is obtained from computer output.

Example 4: *(Porsche prices)*

Find the 95% confidence interval for the population age and mileage coefficients.

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	70.91916	2.48352	28.56	<.0001	65.82341	76.01491
Age	1	-0.13023	0.45684	-0.29	0.7778	-1.06759	0.80714
Mileage	1	-0.56134	0.11407	-4.92	<.0001	-0.79538	-0.32729

The 95% confidence interval for the population age is

$$(-1.06759, 0.80714)$$

The 95% confidence interval for the population mileage is

$$(-0.79539, -0.32729)$$

Coefficient of Multiple Determination:

The coefficient of determination, R^2, uses as a measure of the percentage of total variability in the response that is explained by the regression model.

$$R^2 = \frac{Variability\ explained\ by\ the\ model}{Total\ variability\ in\ Y}$$

$$= \frac{SSModel}{SSTotal} = 1 - \frac{SSE}{SSTotal}$$

- In general, adding a new predictor will increase R^2.

Adjusted Coefficient Determination :

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

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

- The R^2_{adj} value might go down when a weak predictor is added to a model.
- In general $R^2_{adj} \leq R^2$.

Example 4: *(Porsche prices)*

Find R^2 and R^2_{adj} of the full model.

Root MSE	7.29090	R-Square	0.7951
Dependent Mean	50.53667	Adj R-Sq	0.7799
Coeff Var	14.42695		

$$R^2 = 0.7951 \approx 79.51\%$$

and

$$R^2_{adj} = 0.7799 \approx 77.99\%$$

Can you explain whey $R^2_{adj} < R^2$?

Because we add a weak variable to our model.

2. Determine a 95% confidence interval for the average price of Porsche with 50,000 mileage and 6 years old.

($37,469, $46,673)

3. Determine a 95% prediction interval for the price of Porsche with 50,000 mileage and 6 years old.

($26,420, $57,723)

Obs	Price	Age	Mileage	predicted	Lower_Mean	Upper_Mean	Lower_Prediction	Upper_Prediction
31	.	6	50	42.0710	37.4692	46.6728	26.4195	57.7225

Full model vs. reduced model:

- The **full model** contains all predictors (explanatory variables) listed in the dataset.
- The **reduced model** contains only the significant predictors.

Reading Assignment

Read section 3.1 and 3.2