

## **CHAPTER 11 – MULTIPLE LINEAR REGRESSION**

In chapter 10, we discussed simple linear regression, a method which uses a single explanatory variable to explain or predict a single response variable.

In chapter 11, we will be learning about **multiple linear regression**, which uses **several explanatory** variables to explain or predict a single response variable.

In the multiple linear regression setting, the response variable  $y$  depends on  $p$  explanatory variables, rather than a single one.

These explanatory variables will be denoted by  $x_1, x_2, \dots, x_p$ .

The equation for our MLR model is:  $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon_i$ , where the deviations ( $\epsilon_i$ ) are normally distributed with mean 0 and standard deviation  $\sigma$  and  $i=1, \dots, n$ .

The mean response is expressed as  $\mu_y = E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$

So our MLR model consists of our mean response expression and a term for random error.

As with simple linear regression, we can describe our model as:

DATA = FIT + RESIDUAL, where FIT is our mean response and RESIDUAL represents the variation of observations about the means.

### Example #1

So for our football example, we have  $p = 4$  explanatory variables or predictors:  $x_1 = \text{temp}$ ,  $x_2 = \text{humid}$ ,  $x_3 = \text{home}$ ,  $x_4 = \text{injury}$ . Our MLR model would therefore be:

**Solution:** Our MLR model is  $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i$ , where

$y_i$  denotes the number of games won by the  $i$ th team,

$x_{i1}$  denotes the average daily temperature for the  $i$ th team,

$x_{i2}$  denotes the average humidity level for the  $i$ th team,

$x_{i3}$  denotes the number of important home games for the  $i$ th team,

$x_{i4}$  denotes the number of team members with preexisting injuries for the  $i$ th team, and

$\varepsilon_i$  is the error term for the  $i$ th team.

### Data Layout

In simple linear regression we had two variables and treated our data as pairs of  $x$ 's and  $y$ 's,  $(x_i, y_i)$ . So our  $n$  observations would be:

$$\begin{aligned} & (x_1, y_1) \\ & (x_2, y_2) \\ & \cdot \\ & \cdot \\ & \cdot \\ & (x_n, y_n) \end{aligned}$$

In multiple linear regression, because we have multiple explanatory variables, we treat each observation as a vector of my different  $x$ 's and  $y$  values. So for MLR our  $n$  observations would be:

$$\begin{aligned} & (x_{11}, x_{12}, x_{13}, \dots, x_{1p}, y_1) \\ & (x_{21}, x_{22}, x_{23}, \dots, x_{2p}, y_2) \\ & (x_{31}, x_{32}, x_{33}, \dots, x_{3p}, y_3) \\ & \cdot \\ & \cdot \\ & \cdot \\ & (x_{n1}, x_{n2}, x_{n3}, \dots, x_{np}, y_n) \end{aligned}$$

## PARAMETER ESTIMATES

As with SLR, this model represents the true and accurate linear relationship between a response variable and several explanatory variables from a population. This relationship is unknown, however, and we must therefore pick an SRS, find the linear relationship between the response and explanatory variables in our sample and use it to estimate the relationship for our population.

Therefore,  $\sigma$ ,  $\beta_0$ ,  $\beta_1$ , ...,  $\beta_p$  all represent our **population parameters**, whereas  $s$ ,  $b_0$ ,  $b_1$ , ...,  $b_p$  represent our **sample statistics**.

The calculations of  $b_0$ ,  $b_1$ , ...,  $b_p$  are more complicated in MLR and require matrix algebra, therefore we will use statistical software to provide them for us.

### Example # 2

A brand manager for a new pizza flavor collected data on  $y$  = brand recognition (percent of potential customers who can describe what the product is),  $x_1$  = length in seconds of an introductory TV commercial and  $x_2$  = number of repetitions of the commercial over a 2 week period. What does the brand manager **assume** if a MLR model  $\hat{y} = 0.31 + 0.042x_1 + 1.41x_2 + \varepsilon$  is used to predict  $y$ ?

Assumptions:

The  $\varepsilon$ 's ( i.e. the deviations/errors/residuals) are independent and normally distributed with mean 0 and standard deviation  $\sigma$ .

For the  $i^{\text{th}}$  observation, the **predicted response** is  $\hat{y}_i = b_0 + b_1x_{i1} + b_2x_{i2} + \dots + b_px_{ip}$ .

The  $i^{\text{th}}$  **residual**, the difference between the observed and predicted response is

$$\begin{aligned} e_i &= \text{observed response} - \text{predicted response} \\ &= y_i - \hat{y}_i \\ &= y_i - b_0 - b_1x_{i1} - b_2x_{i2} - \dots - b_px_{ip} \end{aligned}$$

The **method of least squares** chooses the values of the  $b$ 's that make the sum of the squares of the residuals as small as possible. That is, we want to choose values for  $b_0, b_1, \dots, b_p$  that make  $\sum (y_i - \hat{y}_i)^2$ , which is equal to  $\sum (y_i - b_0 - b_1x_{i1} - b_2x_{i2} - \dots - b_px_{ip})^2$ , as **small as possible**.

The parameter  $\sigma^2$  measures the **variability of the responses about the population regression equation**. As in the case of SLR, we estimate  $\sigma^2$  by  $s^2$  ( $\sigma$  by  $s$ ), an average of the squared residuals.

$$s^2 = \frac{\sum e_i^2}{n-p-1} = \frac{\sum (y_i - \hat{y}_i)^2}{n-p-1} \quad \text{where } n-p-1 \text{ is the degrees of freedom associated with } s^2.$$

Example #3

Detailed interviews were conducted with over 1000 street vendors in Puebla, Mexico in order to study the factors influencing vendor's incomes. Vendors were defined as individuals working in the street, and included vendors with carts and stands on wheels and excluded beggars, drug dealers, and prostitutes. The researchers collected data on gender, age, hours worked per day, annual earnings, and education level. A sample of these data appears in the following table:

| Vendor Number | Annual Earnings<br>$y$ | Age $x_1$ | Hours Worked per Day $x_2$ |
|---------------|------------------------|-----------|----------------------------|
| 1             | \$2841                 | 29        | 12                         |
| 2             | 1876                   | 21        | 8                          |
| 3             | 2934                   | 62        | 10                         |
| 4             | 1552                   | 18        | 10                         |
| 5             | 3065                   | 40        | 11                         |
| 6             | 3670                   | 50        | 11                         |
| 7             | 2005                   | 65        | 5                          |
| 8             | 3215                   | 44        | 8                          |
| 9             | 1930                   | 17        | 8                          |
| 10            | 2010                   | 70        | 6                          |
| 11            | 3111                   | 20        | 9                          |
| 12            | 2882                   | 29        | 9                          |
| 13            | 1683                   | 15        | 5                          |
| 14            | 1817                   | 14        | 7                          |
| 15            | 4066                   | 33        | 12                         |

- a) Write a MLR model for annual earnings ( $y$ ) as a function of age ( $x_1$ ) and hours worked ( $x_2$ ).

We fit our model using **R**, a statistical software.

```
Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept) -20.352   652.745  -0.031  0.97564
age          13.350    7.672   1.740  0.10738
hours       243.714   63.512   3.837  0.00236 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Null deviance: 8618428 on 14 degrees of freedom
Residual deviance: 3600196 on 12 degrees of freedom
```

- b) Find the least squares regression equation (or prediction equation) from the data above.
- c) Interpret the estimated  $\beta$  coefficients in your model.
- d) Is age ( $x_1$ ) a statistically useful predictor of annual earnings? Test using  $\alpha = 0.01$ .
- e) Find a 95% confidence interval for  $\beta_2$ . Interpret the interval in the words of the problem.

**Solution:**

- a) The MLR model is  $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$ , where  $y_i$  denotes the annual income of the  $i$ th vendor,  $x_{i1}$  denotes the age of the  $i$ th vendor,  $x_{i2}$  denotes the hours worked per day of the  $i$ th vendor, and  $\varepsilon_i$  is the error term for the  $i$ th vendor.
- b) The prediction equation from the data above is  $y_i(\text{hat}) = -20.352 + 13.35x_{i1} + 243.714x_{i2}$ .

- c) The intercept parameter -20.352 is the predicted value of yearly income if  $x_1 = x_2 = 0$ . The age coefficient had a value of 13.35, so this means that for each increase in age by 1 year, a vendor's yearly income is predicted to increase by 13.35. The hours coefficient is 243.714, which means that for each increase in hours worked per day by 1 hour, a vendor's income is predicted to increase by 242.714.
- d) This is a test of  $H_0: \beta_1 = 0$  versus  $H_a: \beta_1 \neq 0$ . The test statistic for this test is  $t_o = b_1/SE(b_1) = 13.35/7.672 = 1.74$  with  $df = 12$ . The test statistic 1.74 with  $df = 12$  yields a p-value of 0.107. Hence we fail to reject the null hypothesis.
- e) A 95% CI for  $\beta_2$  is (105.32, 382.11). We are 95% confident that for each increase in hours worked per day by one hour, a vendor's annual income will increase by some value in the interval (105.32, 382.11).

### **A SIDE NOTE ABOUT DEGREES OF FREEDOM**

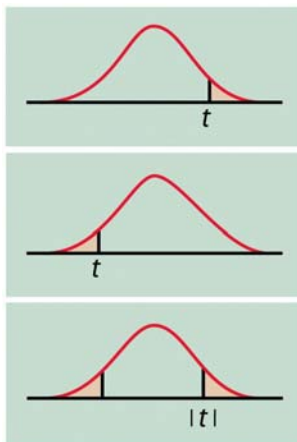
Suppose we know that for a sample size of  $n=3$  with  $\bar{x} = 6$ ,  $x_1 = 5$  and  $x_2 = 8$ . Then we also know that  $x_3$  must be equal to 5 in order to get a mean of 6. Therefore,  $x_3$  was not free to be any number, it was only able to take on the value of 5. Therefore, when we use  $\bar{x}$  in our calculation of  $s$ ,  $s$  will lose a degree of freedom and will then have  $n-1$  degrees of freedom.

**So in our calculations for  $s^2$ , since our sample size is  $n$  and there are  $p+1$  parameters ( $\beta$ 's) that must be estimated, we will be left with  $n - (p + 1) = n-p-1$  degrees of freedom.**

### **CONFIDENCE INTERVALS AND SIGNIFICANCE TESTS FOR $\beta_j$**

A level  $C$  confidence interval for  $\beta_j$  is  $\beta_j \pm t^* SE_{\beta_j}$ , where  $SE_{\beta_j}$  is the standard error of  $\beta_j$  and  $t^*$  is the value for the  $t$ -distribution with  $n-p-1$  degrees of freedom that has an area of  $C$  between  $-t^*$  and  $t^*$ .

To test the hypothesis  $H_0: \beta_j = 0$ , compute the  $t$  statistic  $t = \frac{\beta_j}{SE_{\beta_j}}$ .



For a test of  $H_0: \beta_j = 0$  against:

- i)  $H_a: \beta_j > 0$  our p-value is  $P(T \geq t)$
- ii)  $H_a: \beta_j < 0$  our p-value is  $P(T \leq t)$
- iii)  $H_a: \beta_j \neq 0$  our p-value is  $2P(T \geq |t|)$

Where  $T$  is a random variable from the  $t$  distribution with  $n-p-1$  degrees of freedom.

***ANOVA FOR SIMPLE LINEAR REGRESSION (see Ch 10 notes)***

| Source | Degrees Of Freedom | Sums of Squares                        | Mean Square | F       |
|--------|--------------------|--|-------------|---------|
| Model  | 1                  | $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ | SSM/DFM     | MSM/MSE |
| Error  | n-2                | $\sum_{i=1}^n (y_i - \hat{y}_i)^2$     | SSE/DFE     |         |
| Total  | n-1                | $\sum_{i=1}^n (y_i - \bar{y})^2$       | SST/DFT     |         |

Example #4

Fill in the following blanks:

| Source | Degrees Of Freedom | Sums of Squares | Mean Square | F |
|--------|--------------------|-----------------|-------------|---|
| Model  | 1                  |                 |             |   |
| Error  |                    | 56              |             |   |
| Total  | 19                 | 200             |             |   |

**ANOVA FOR MULTIPLE LINEAR REGRESSION**

In Simple linear regression, the ANOVA F test is equivalent to a two-sided t test of the hypothesis that the slope of the regression line is 0 ( $H_0: \beta_1 = 0$  vs.  $H_a: \beta_1 \neq 0$ ).

For multiple linear regression, the ANOVA F test is a bit different. It tests whether all of the regression coefficients (with the exception of the intercept  $\beta_0$ ) are equal to 0.

**$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$  vs.  $H_a$ : at least one of the  $\beta_j$ 's is not equal to 0.**

The MLR ANOVA table is similar to that of SLR, the only difference being the degrees of freedom. Our **model** now has p degrees of freedom since we now have p explanatory variables instead of just one.

ANOVA table

| Source | Degrees Of Freedom | Sums of Squares                        | Mean Square | F       |
|--------|--------------------|--|-------------|---------|
| Model  | p                  | $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ | SSM/DFM     | MSM/MSE |
| Error  | n-p-1              | $\sum_{i=1}^n (y_i - \hat{y}_i)^2$     | SSE/DFE     |         |
| Total  | n-1                | $\sum_{i=1}^n (y_i - \bar{y})^2$       | SST/DFT     |         |

**R<sup>2</sup>**

As with SLR, we can compute R<sup>2</sup> from our sums of squares.  $R^2 = \frac{SSM}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$

**The difference with the R<sup>2</sup> for multiple linear regression is that it measures the fraction of the variation in the response variable y that is explained by the multiple explanatory variables x<sub>1</sub>, x<sub>2</sub>, ..., x<sub>p</sub>.**

Example #5

Nutrition facts labels provide consumers with information about the nutritional value of food products that they buy. A study of these labels collected data from 152 consumers who were sent information about a frozen chicken dinner. Each subject was asked to give an overall product nutrition score, and also evaluated each of 10 nutrients on a 9-point scale with higher values indicating that the product has a healthy value for the given nutrient. Composite scores for favorable nutrients (such as protein and fiber) and unfavorable nutrients (such as fat and sodium) were used in a multiple regression to predict the overall product nutrition score. The following was reported in a table:

| Explanatory Variables | B    | SE   | T      | Model F | R <sup>2</sup> |
|-----------------------|------|------|--------|---------|----------------|
| Unfavorable Nutrients | 0.82 | 0.12 | 6.8**  | 33.7**  | 0.31           |
| Favorable Nutrients   | 0.57 | 0.10 | 5.5**  |         |                |
| Constant              | 3.33 | 0.13 | 26.1** |         |                |
| ** p<0.01             |      |      |        |         |                |

- a) What is the equation of the least-squares line?
- b) Give the null and alternative hypothesis associated with the entry labeled “Model F” and interpret this result.

c) The column labeled “t” contains three entries. Explain what each of these means.

d) What are the degrees of freedom associated with the t statistics that you explained in part (c)?

**Solution:**

a) The equation of the least-squares line is  $\hat{y}_i = 3.33 + 0.82x_{i1} + 0.57x_{i2}$ , where  $y_i$  denotes the overall nutrition rating from the  $i$ th consumer,  $x_{i1}$  denotes the unfavorable nutrients rating by the  $i$ th consumer, and  $x_{i2}$  denotes the favorable nutrients rating by the  $i$ th consumer.

b)  $H_0: \beta_1 = \beta_2 = 0$  versus  $H_a$ : not all  $\beta_j$ 's equal 0. This F-value comes from MSM/MSE and has an F distribution with 2 numerator degrees of freedom and 149 denominator degrees of freedom. The high value of this F statistic yields a small p-value. We have evidence to reject the null hypothesis.

c) These numbers are the t-statistics that you obtain in order to test whether or not the factor has an effect.

d) The degrees of freedom are 149 for all three of them.

### Example #6

Consider the regression problem of predicting GPA from the three high school grade variables: average high school grades in mathematics (HSM), science (HSS) and English (HSE). The computer output appears in the following figure.

Output

```
Dependent Variable: GPA

                    Analysis of Variance

Source              DF          Sum of          Mean
                    Squares          Square          F Value          Prob>F

Model                3          27.71233          9.23744          18.861          0.0001
Error              220          107.75046          0.48977
C Total            223          135.46279

Root MSE          0.69984          R-Square          0.2046
Dep Mean          2.63522          Adj R-sq          0.1937
C.V.              26.55711

                    Parameter Estimates

Variable           DF          Parameter          Standard          T for H0:
                    Estimate          Error          Parameter=0          Prob > |T|

INTERCEP           1          0.589877          0.29424324          2.005          0.0462
HSM                 1          0.168567          0.03549214          4.749          0.0001
HSS                 1          0.034316          0.03755888          0.914          0.3619
HSE                 1          0.045102          0.03869585          1.166          0.2451
```

- Write the estimated regression equation.
- What is the value of  $s$ , the estimate of  $\sigma$ ?
- State  $H_0$  and  $H_a$  tested by the ANOVA F statistic for this problem. After stating the hypotheses in symbols, explain them in words.
- What is the distribution of the F statistic under  $H_0$ ? What conclusion do you draw from the F test?
- What percent of the variation in GPA is explained by these three high school grade variables?

**Solution:**

- a) The estimated regression equation is  $\hat{y}_i = 0.590 + 0.169x_{i1} + 0.034x_{i2} + 0.045x_{i3}$ , where  $x_{i1}$ ,  $x_{i2}$ , and  $x_{i3}$  denote the student's average grades in mathematics, science, and English, respectively.
- b) The value of  $s$  is 0.69984
- c)  $H_0: \beta_1 = \beta_2 = \beta_3 = 0$  versus  $H_a$ : not all  $\beta_j$ 's are 0. The null hypothesis says that a student's average grades in mathematics, science, and English have no effect on a student's GPA. The alternative hypothesis says that at least one of these three factors does have an effect on a student's GPA.
- d) The distribution of the F statistic under  $H_0$  is the F distribution with 3 numerator degrees of freedom and 220 denominator degrees of freedom. We get an F statistic value of 18.861 with 3 numerator df and 220 denominator df. This yields a p-value of less than 0.0001, which gives us evidence to reject the null hypothesis. We have reason to believe that at least one of math, science, and English grades have an effect on a student's GPA.
- e) We obtain an  $R^2$  value of 0.2046 in this example.