12.3: Multiple Linear Regression

A multiple linear regression line describes how two or more predictor variables affect the response variable y. An equation of a line relating p independent variables to y is of the form for the population as: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$, where $\beta_1, \beta_2, \ldots, \beta_p$ are the slopes, β_0 is the y-intercept and ε is called the error term.

We use sample data to estimate this equation using the predicted value of y as $\hat{y}$ with the regression equation (also called the line of best fit or least squares regression line) as:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

where $b_1, b_2, \ldots, b_p$ are the slopes, and b_0 is the y-intercept

For example, if we had two independent variables, we would have a 3-dimensional space as in Figure 12-25 where the red dots represent the sample data points and the equation would be a plane in the space represented by $y = b_0 + b_1 x_1 + b_2 x_2$.

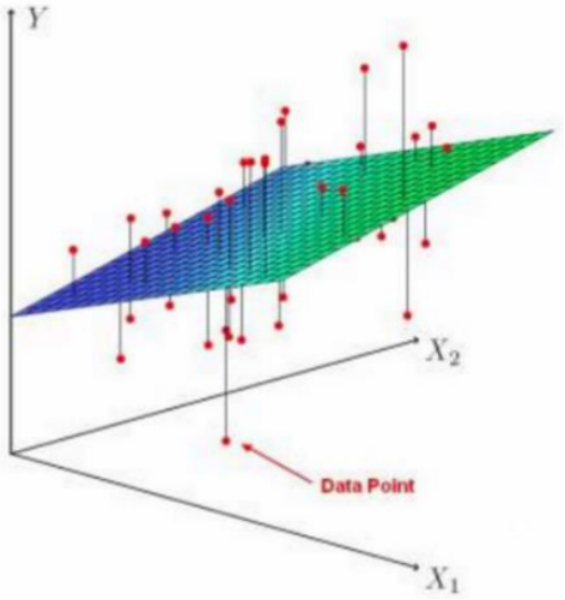

Figure 12-25: Multiple linear regression with 2 independent variables. This photo by unknown author is licensed under CC BY-SA-NC.

The calculations use matrix algebra, which is not a prerequisite for this course. We will instead rely on a computer to calculate the multiple regression model.

If all the population slopes were equal to zero, the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$ would not be significant and should not be used for prediction. If one or more of the population slopes are not equal to zero then the model will be significant, meaning there is a significant relationship between the independent variables and the dependent variable and we may want to use this model for prediction. There are other statistics to look at to decide if this would be the best model to use. Those methods are discussed in more advanced courses.

The hypotheses will always have an equal sign in the null hypotheses.

The hypotheses are:

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$
$H_1 :$ At least one slope is not zero.

Note that the alternative hypothesis is not written as $H_1 : \beta_1 \neq \beta_2 \neq \cdots \neq \beta_p \neq 0$. This is because we just want one or more of the independent variables to be significantly different from zero, not necessarily all the slopes unequal to zero.

Use the F-distribution with degrees of freedom for regression = $df_R = p$, where p = the number of independent variables (predictors), and degrees of freedom for error = $df_E = n - p - 1$, where n is the number of pairs. This is always a right-tailed

ANOVA test, since we are testing if the variation in the regression model is larger than the variation in the error.

The test statistic and p-value are the last two values on the right in the ANOVA table. The p-value rule is easiest to use since the p-value is part of the outcome, but a critical value can be found using the invF program on your calculator or in Excel using =F.INV.RT(α, df_R, df_E) We can also single out one independent variable at a time and use a t-test to see if the variable is significant by itself in predicting y.

This would have hypotheses:

$H_0 : \beta_i = 0$
$H_1 : \beta_i \neq 0$
where i is a placeholder for whichever independent variable is being tested.

This t-test is found in the same row as the coefficient that you are testing.

Assumptions for Multiple Linear Regression

When doing multiple regression, the following assumptions need to be met:

1. The residuals of the model are approximately normally distributed.
2. The residuals of the model are independent (not autocorrelated) and have a constant variance (homoscedasticity).
3. There is a liner relationship between the dependent variable and each independent variable.
4. Independent variables are uncorrelated with each other (no multicollinearity).

The following is a schematic for the regression output for Microsoft Excel. Other software usually has a similar output but may have numbers in slightly different places. The blue spaces have the descriptions of the corresponding numbers.

Regression Statistics	
Multiple R	Multiple Correlation Coefficient = $\|r\|$
R Square	Coefficient of Determination = R^2
Adjusted R Square	R^2 Adjusted for degrees of freedom for Multiple Regression
Standard Error	Standard Error of Estimate (Residual Standard Deviation) = s
Observations	Number of data pairs = n

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	p=number of independent variables	SSR=Regression Sum of Squares	MSR=SSR/k	Test Statistic F=MSR/MSE	p-value for whole model
Residual	n – p – 1	SSE=Error Sum of Squares	MSE= SSE/(n–p–1)		
Total	n – 1	SST=Total Sum of Squares			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Y-Intercept	b_0	Standard Deviation of b_0	Test Stat for b_0	p-value for b_0
1st X Variable	b_1	Standard Deviation of b_1	Test Stat for b_1	p-value for b_1
2nd X Variable	b_2	Standard Deviation of b_2	Test Stat for b_2	p-value for b_2
3rd X Variable	b_3	Standard Deviation of b_3	Test Stat for b_3	p-value for b_3

Figure 12-26: Excel output for multiple linear regression.

The coefficients column gives the numeric values to find the regression equation $y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$. The p-values for b_i should be investigated to see if the variable is statistically significant. One should also be careful that the independent variables are not significantly correlated amongst themselves. Correlated independent variables may give unexpected outcomes in the overall regression model and actually flip the sign on a coefficient.

A sample of 30 homes that were recently on the market were selected. The listing price in $1,000's of the home, the livable square feet of the home, the lot size in 1,000's of square feet and the number of bathrooms in the home were recorded. A multiple linear regression was done in Excel with the following output. Test to see if there is a significant relationship between the listing price of a home with the livable square feet, lot size, and number of bathrooms. If there is a relationship, then use the regression model to predict the listing price for a home that has 2,350 square feet, 3 bathrooms and has a 5,000 square foot lot. Use α = 0.05.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.977709
R Square	0.955915
Adjusted R Square	0.950828
Standard Error	20.3056
Observations	30

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	232450.1	77483.36	187.9217	9.74E-18
Residual	26	10720.25	412.3173		
Total	29	243170.3			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-28.8477	29.71148	-0.97093	0.34053	-89.9206	32.22507
Square Feet	0.170908	0.015448	11.06366	2.48E-11	0.139155	0.202661
Lot Size	6.777705	1.421295	4.768683	6.19E-05	3.856191	9.699218
Bathrooms	15.5347	9.20827	1.687038	0.10356	-3.39317	34.46257

Solution

First, we need to test to see if the overall model is significant.

The hypotheses are:

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$
H_1 : At least one slope is not zero.

The test statistic is $F = 187.9217$ and the p-value = $9.74E-18 \sim 0$

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	232450.1	77483.36	187.9217	9.74E-18
Residual	26	10720.25	412.3173		
Total	29	243170.3			

We reject H_0, since the p-value is less than α = 0.05. There is enough evidence to support the claim that there is a significant relationship between the number of bathrooms and lot size of a home with its listing price. Since we reject H_0, we can use the regression model for prediction.

The question asked to predict the listing price for a home that has 2,350 square feet, 3 bathrooms and has a 5,000 square foot lot. This gives us $x_1 = 2350$, $x_2 = 5$ (5,000 square feet), and $x_3 = 3$.

The coefficients column has the values that correspond to the y-intercept and slopes gives the regression equation: $\hat{y} = -28.8477 + 0.170908 \cdot x_1 + 6.7705 \cdot x_2 + 15.5347 \cdot x_3$.

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-28.8477	29.71148	-0.97093	0.34053	-89.9206	32.22507
Square Feet	0.170908	0.015448	11.06366	2.48E-11	0.139155	0.202661
Lot Size	6.777705	1.421295	4.768683	6.19E-05	3.856191	9.699218
Bathrooms	15.5347	9.20827	1.687038	0.10356	-3.39317	34.46257

Substitute the three given x values into the equation in the correct order and you get $\hat{y} = -28.8477 + 0.170908 \cdot 2350 + 6.7705 \cdot 5 + 15.5347 \cdot 3 = 453.2787$

This then gives a predicted listing price of $453,278.

Note that our sample size is very small and we really need to check assumptions in order to use this predicted value with any reliability.

Is this the best model to use? Note that not all the p-values for each of the individual slope coefficients are significant. The number of bathrooms has a t-test statistic = 1.687038 and p-value = 0.10356, which is not statistically significant at the 5% level of significance.

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-28.8477	29.71148	-0.97093	0.34053	-89.9206	32.22507
Square Feet	0.170908	0.015448	11.06366	2.48E-11	0.139155	0.202661
Lot Size	6.777705	1.421295	4.768683	6.19E-05	3.856191	9.699218
Bathrooms	15.5347	9.20827	1.687038	0.10356	-3.39317	34.46257

We may want to rerun the regression model without the number of bathroom variables and see if we get a higher R^2 and a lower standard error of estimate. Ideally, we would try all the different combinations of independent variables and see which combination gives the best model. This is a lot of work to do if you have many independent variables. Most statistical software packages have built in functions that find the best fit.

Adjusted Coefficient of Determination

When we add more predictor variables into the model, this inflates the coefficient of variation, R^2. In multiple regression, we adjust for this inflation using the following formula for adjusted coefficient of variation.

Adjusted Coefficient of Determination

$$R^2_{adj} = 1 - \left(\frac{(1 - R^2)(n-1)}{(n-p-1)} \right)$$

Use the previous example to verify the value of the adjusted coefficient of determination starting with the regular coefficient of determination $R^2 = 0.955915$.

Solution

First identify in the Excel output $R^2 = 0.955915$, $n-1 = df_T = 29$, and $n-p-1 = df_E = 26$.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.977709
R Square	0.955915
Adjusted R Square	0.950828
Standard Error	20.3056
Observations	30

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	232450.1	77483.36	187.9217	9.74E-18
Residual	26	10720.25	412.3173		
Total	29	243170.3			

Substitute these values in and we get $R^2_{adj} = 1 - \left(\frac{(1-0.955915)(29)}{(26)}\right) = 0.950828$. This is the same value as the adjusted R^2 reported in the Excel output.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.977709
R Square	0.955915
Adjusted R Square	0.950828

The Excel output has both the adjusted coefficient of determination and the regular coefficient of determination. However, you may need the equation for the adjusted coefficient of determination depending on what information is given in a problem.

There are more types of regression models and more that should be done for a complete regression analysis. Ideally, you would find several models and pick the one with no outliers, the smallest standard error of estimate, a good residual plot, and the highest adjusted R^2 and check the assumptions behind each model before using for prediction. More advanced techniques are discussed in a regression course.

> *"Well, I was in fact, I was moving backwards in time. Hmmm. Well, I think we've sorted all that out now. If you'd like to know, I can tell you that in your universe you move freely in three dimensions that you call space. You move in a straight line in a fourth, which you call time, and stay rooted to one place in a fifth, which is the first fundamental of probability. After that it gets a bit complicated, and there's all sorts of stuff going on in dimensions 13 to 22 that you really wouldn't want to know about. All you really need to know for the moment is that the universe is a lot more complicated than you might think…"*
>
> *(Adams, 2002)*