In this unit we were introduced to regression and forecasting. Regression is a statistical procedure that uses data from the past to make a prediction about the future. In this course we will only be concerned with linear regression but be aware there are curvilinear regression equations. We also learned about time-series models of forecasting. The time-series models we studied are moving average, exponential smoothing, trend projections, and decomposition. Regression analysis can be used in trend projections and in one type of decomposition model.

Recall from Unit 1 that we can create a scatterplot for two variables. The X variable is the predictor variable and the Y variable is the predicted variable. Here is the scatter diagram from Unit 1.

Clearly we can see that the data points go from upper left to lower right indicating that there is a negative correlation between study time and errors on the mid-term. We can use that relationship to make predictions using the equation for the regression line, Ŷ = **0 + **1(X) see page 221). For those who are math inclined (not me for sure) you will notice this simply is the equation for a straight line!

We can use Excel to generate not only the scatter diagram but the regression line as well. Here is the same output with the “trendline” added, note the trendline is the regression line. You will see that the line is about as close to all the points as possible, which is what is really does. The distance between the line and any point is prediction error and the sum of the squared errors is as small as mathematically possible. This is known as the “Least Squares Criterion.” You will also notice that the equation and the r2 have been added to the scatter diagram. If you check the Live Binder you will find a video showing you how to do this. The equation is in a slightly different format than the one we use but it is the same formula. Excel uses the equation: Ŷ = **1(X) + **0, so be careful when using it.



To use the equation to make a prediction we simply substitute a value for X (how long did someone study) and then predict how many errors they “MIGHT” make on the midterm examination. For example, let’s say you tell me you studied 9 hours for the midterm and want to know if that is enough to do well on the midterm. Using the equation I substitute 9 for X in our equation:

Ŷ = **0 + **1(X)

Ŷ = ** - **(9)

Ŷ = ** - **

Ŷ = *4.3713 errors*

So I would tell you that I would predict that you might make between 4 and 5 errors on the midterm, based on how students have done in the past who study about 9 hours for the exam. This is a forecast based on the relationship between the two variables.

In real-world applications we rarely have only one predictor variable. The concept of the multiple regression, that is more than one predictor variable, is the same as for simple regression. It is not as easy to graph but the equation for multiple regression is just an expansion of the equation for a single predictor variable.

Ŷ = **0 + **1(X1) + **2(X2) + **3(X3)

To get this equation we cannot use have it added to the chart because, as I said it is not easy to make the chart! What we will use is the Excel Data Analysis procedure “Regression” or the PHStat procedure “Multiple Regression.” In either procedure the predictor variables (X variables) must be in contiguous columns. In the book on page 224 are some sample data from the Jenny Wilson Real Estate Company. Those data are reproduced here:

|  |  |  |
| --- | --- | --- |
| Selling price | Square Footage | Age |
| 95000 | 1926 | 30 |
| 119000 | 2069 | 40 |
| 124800 | 1720 | 30 |
| 135000 | 1396 | 15 |
| 142800 | 1706 | 32 |
| 145000 | 1847 | 38 |
| 159000 | 1950 | 27 |
| 165000 | 2323 | 30 |
| 182000 | 2285 | 26 |
| 183000 | 3752 | 35 |
| 200000 | 2300 | 18 |
| 211000 | 2525 | 17 |
| 215000 | 3800 | 40 |
| 219000 | 1740 | 12 |

Jenny wants to build a model to predict the selling price of the house based on square footage and age. Notice that the two predictor variables, square footage and age are in adjacent columns. Using the Regression procedure from Data Analysis we get the following output.

There is a lot of information in this output that we will NOT use. The things we need to attend to are in red in the output above. The first is the Multiple R. The Multiple R is the same as the correlation r but the capital R indicates that it is between multiple predictors and a single dependent variable. In our case we see the correlation is 0.82 (there is no need to have more than 2 decimal places for the correlation). The R square is the coefficient of determination and is interpreted the same as the r2. In our situation 67% of the variance in selling price is accounted for by the square footage and age.

You should IGNORE the ANOVA output, we just do not have time to cover all of this in just one 10-week term.

The Coefficients are listed in the last table of the output. The three elements of our equation are listed in the first column. The coefficient for the Intercept is the 0 in our equation. The coefficient for square footage is the 1 in our equation and the coefficient for age is the 2 in our equation. So, the equation we would use is:

Ŷ = **0 + **1(X1) + **2(X2)

Substituting our data from the output we have:

Ŷ = *146,630 .89* + *43.82* (X1) – *2,898.69* (X2), where X1 is the square footage, X2 is the age and Ŷ is the predicted selling price. So if we have a house that has 2500 square feet and is 10 years old we would have:

Ŷ = *146,630 .89* + *43.82* (2500) – *2,898.69* (10)

Ŷ = *146,630 .89* + *109,550* – *28,986.9*

Ŷ = *146630 .89* + *80,536.10*

Ŷ = *227,193.99*

So for our house with 2500 square feet that is 10 years old we predict the selling price will be around $227,194. This is really all there is to making predictions with regression. If you think about the equation we have just generated and used, it makes perfect sense. We start with a base selling price and then as the square footage goes up the selling price also goes up (1 is positive) but as the age increases the selling price goes down (2 is negative).

**Measures of Forecast Accuracy**

The next concept we discussed is measures of forecast accuracy. To assess the accuracy of a forecast we will compare the forecasted values with the actual or observed values. The forecast error is defined as:

**Forecast error = Actual value – Forecast value.**

There are three measures of forecast error that we will discuss. The first is the Mean Absolute Deviation (MAD), which is the just what it says, the mean absolute deviation.

**MAD** =

The second is the Mean Squared Error (MSE), which is:

**MSE** =

The third measure of forecast error is the Mean Absolute Percent Error (MAPE), which is:

**MAPE** = x 100%

The choice of which to use is up to the decision maker and the situation. In most cases I will use the MAD, but when doing so must keep in mind the units of the forecast.

**Time-Series Forecasts**

Time-series models attempt to predict the future by using historical data. They are based on the old adage, the best predictor of the future is the past! The three different models we covered are (1) Moving Averages, (2) Exponential Smoothing, and (3) Trend projections. We will not spend a lot of time on the decomposition model. As you suspect, we will use technology to do all the math for us; we will simply enter the beginning data and click Solve!

**Moving Averages** is a process of creating a moving average over a specified time period. They are useful in a situation where we can assume that market demands will stay fairly steady over time. An n-period moving average forecast is:

Moving Average Forecast =

I will QM for Windows to show you how to get the moving average forecast. We will start with some historical data. These are the data from page 247 in the text.

|  |  |
| --- | --- |
| Month | Actual Shed Sales |
| January | 10 |
| February | 12 |
| March | 13 |
| April | 16 |
| May  | 19 |
| June | 23 |
| July | 26 |
| August | 30 |
| September | 28 |
| October | 18 |
| November | 16 |
| December | 14 |
| January  | This is what we want to forecast! |

Here is the output from Excel QM for a three-month moving average. You will need to use Excel QM and not QM for Windows for forecasting!

Notice that all three measures of forecast error are shown and the predicted sales or demand for the next month (January) is 16. Compare this to Table 6.7 in your text. Pay particular attention to the Num pds above the data. The default is 1 and we wanted a three period moving average so I changed that to 3.

The next model is the **weighted moving average**. The moving average gives the same weight to each of the past observations, whereas the weighted average allows different weights to be assigned to the previous observations. This may be desirable if we believe that the most recent sales should have more weight than the earlier months. Here is the output for the weighted moving average.

Compare this output to Table 6.8 in the text. Also note that you enter the weights for the most recent periods not beside them but just in order at the top. This will put more weight on Months 11 and 12. Again the three different measures of forecast error are provided and the predicted sales for January are now 15.33, compared with a prediction of 16 using the simple three-month moving average.

The last time-series forecast we will study is **exponential smoothing**. Exponential smoothing uses a smoothing constant () that will have a value between 0 and 1, inclusive. Here is the output from the Port of Baltimore example in the text using  = .10. Following this is the same analysis using  = .50. We will then compare the two forecasts using our measures of forecast error to see which would be the better level for .

To compare these two different levels of , compare the MAD for each one.

Using  = 0.10 the MAD is 9.51 and using  = 0.5 the MAD is 12.05. Clearly using  = 0.10 yields a smaller MAD and would be the preferred smoothing constant.

The concept of trend projections is the multiple regression presented earlier in my summary. If you have any questions please let me know.

Dr. Bob