John Wolters

Franco Porto

Linear modeling

Linear modeling is a powerful tool in data analysis designed to use regression to demonstrate a relationship between a dependent variable and one or multiple independent variables. Simple regression involves an x-y relationship however multiple regressions allows for any number of x's and one y. Multiple regression is incredibly useful for analyzing very large data sets with a multitude of variables. At times a data set has more variables than are necessarily predictive of the dependent variable. There are methods that can be used to reduce a linear model to a smaller more optimal variation.

- To begin let's look at the following data set:
  Data=read. table ("ZarEX7.4.txt",header=T)

Faculty Salary Simulated Data

Faculty	Salary	Gender	Rank	Dept	Years	Merit
1	38	0	3	1	0	1.47
2	58	1	2	2	8	4.38
3	80	1	3	1	9	3.65
4	30	1	1	1	0	1.64
5	50	1	1	3	0	2.54
6	49	1	1	3	1	2.06
7	45	0	3	1	4	4.76
8	42	1	1	2	0	3.05
9	59	0	3	3	3	2.73
10	47	1	2	1	0	3.14
11	34	0	1	1	3	4.42
12	53	0	2	3	0	2.36
13	35	1	1	1	1	4.29
14	42	0	1	2	2	3.81
15	42	0	1	2	2	3.84
16	51	0	3	2	7	3.15
17	51	1	2	1	8	5.07
18	40	0	1	2	3	2.73
19	48	1	2	1	1	3.56
20	34	1	1	1	7	3.54
21	46	1	2	1	2	2.71
22	45	0	1	2	6	5.18
23	50	1	1	3	2	2.66
24	61	0	3	3	3	3.7

25	62	1	3	1	2	3.75
26	51	0	1	3	8	3.96
27	59	0	3	3	0	2.88
28	65	1	2	3	5	3.37
29	49	0	1	3	0	2.84
30	37	1	1	1	9	5.12

Here is a description on what the numbers mean

- **Salary**
- **Gender** (0=Male, 1=Female)
- **Rank** (1=Assistant, 2=Associate, 3=Full)
- **Dept** Department (1=Family Studies, 2=Biology, 3=Business)
- **Years** since making Rank
- Average **Merit** Ranking

*Also note there are many different variables, and some of these variables are factors; treatment classes with multiple levels. The key to what these different factors mean is presented to the right.

For this example Salary will be used as the dependent variable while the other variables will be used as independent variables to predict the values of Salary.

Simple Regression:
To begin let us look at how a simple regression using this data set might appear.
Salary will be the dependent variable and Gender will be used as the independent variable.
For this question we might ask if the variable Gender shows a significant correlation to Salary. Such a correlation would show a disparity in pay between different genders, a question that some might ask as a means to demonstrate inequality in the workplace. Worded as a null hypothesis:

Ho: B = 0, Gender has no correlation to Salary
H1: B ≠ 0, there is a correlation between Gender and Salary.

The first step is to actually generate the linear model. This is done very simply in R via the linear model function

```
>attach(Data)
>X=factor(Gender)
>Y=Salary
>LM=lm(Y~X)   ←           #The variable LM is now assigned to our linear model
```

```
>LM
Call:lm(formula = Y ~ X)

Coefficients:
(Intercept)        X1
47.786       1.214
```

To analyze the linear model we can do any of three things, and it is generally useful to do all three.

*The first would be a graphical representation, but as Gender is a factor with only two levels this is not a useful analysis here. It is also less useful in that it is difficult to apply to multiple regressions.

* The second is the summary function.

```
>summary(LM)
Residuals:
   Min     1Q  Median     3Q    Max
-19.000  -6.696  -0.500   4.714  31.000
Coefficients:
         Estimate Std. Error t value Pr(>|t|)
(Intercept)   47.786     2.930  16.307 7.94e-16 ***
X1           1.214     4.013   0.303    0.764
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 10.96 on 28 degrees of freedom
Multiple R-squared: 0.00326,    Adjusted R-squared: -0.03234
F-statistic: 0.09157 on 1 and 28 DF,  p-value: 0.7644.
```

Interpretation:

For the null hypothesis asked above we are concerned with the t-test for β = 0. The values for this test are found in the row labeled X1. The t-value was .303 and the p-value was .764. The p-value is greater than α=.05 therefore we cannot reject the null hypothesis and instead accept it. B is roughly equivalent to 0.

In the terms of the question asked, we can draw the conclusion that Gender does not demonstrate a significant impact upon salary. All in all the pay in this workplace appears to be fair between genders. The summary function generates several useful values for us. It also runs two basic t-tests to determine if the α value is 0, and if the slope of the regression line, β, is 0. It also generates an r-squared value, the coefficient of correlation.

For these t-tests we follow a standard decision rule that if p is less than the standard type I error rate α=.05 then we reject the null hypothesis, otherwise we accept the null hypothesis

*The third method of analysis is to use the anova function.

```
>anova(LM)

Analysis of Variance Table
```

```
Response: Y
          Df Sum Sq Mean Sq F value Pr(>F)
X          1   11.0   11.01  0.0916 0.7644
Residuals 28 3366.4  120.23
```

Interpretation:

This generates an ANOVA table along with running an ANOVA F-test for β=0. For this test we obtained an F-value of .0916, and a p value of .7644. Note that the p-value between the t-test in the summary table is effectively equivalent to the p-value from this test. These tests ask the same question and generate the same result.

All the same conclusions drawn from the t-test above can also be generated from this anova table. Both the summary function and the linear model function are highly useful in analyzing multiple regressions as well.

Multiple Regression:

In this dataset there are a multitude of other variables besides Gender that can be used to help predict the values of Salary. To analyze the data in terms of all of these variables it is possible to generate a linear model using all of these variables as independent variables.

Before we can generate these variables, however, there is one problem to consider. While a linear model can use a factor as an independent variable with no special considerations if it has only two levels which are reduced to 1 and 0, some consideration is required for factors with more than two levels. Fortunately, R has a very powerful means of doing all the dummy coding required to properly analyze these factors, the factor() function. This function specifies that a variable is a factor and in doing so covers all the required considerations. As long as you are careful to specify variables as factors when necessary this will not be a problem.

Now with the factor() function in mind we can generate a linear model for a multiple regression:

```
>Y=Salary
>X1=Gender
>X2=factor(Rank)
>X3=factor(Dept)
>X4=factor(Years)
>X5=Merit
>LM=lm(Y~X1+X2+X3+X4+X5)   ←——   to specify multiple x's in the lm function we just
put a + for each additional X
>LM
Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5)

Coefficients:
(Intercept)       X1       X22       X23       X32       X33
   24.172      6.273    10.596    18.877     8.607    15.021
      X41       X42       X43       X44       X45       X46
    1.385     4.436     1.496    -3.281     5.234     6.527
```

```
     X47        X48        X49        X5
   -2.227      5.126      13.797      1.099
```

The first and broadest question to ask in multiple regressions is if any of the variables are having any effect on y. This is the overall F-test for all β=0:
Ho: all slopes, all β = 0
H1: at least one β for any variable =/= 0
To run this test in R we once again use the summary function
>summary(LM)

```
Residuals:
      Min        1Q    Median        3Q       Max
-1.287e+01 -2.149e+00 -7.070e-16  1.854e+00  1.287e+01

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   24.172      7.241   3.338 0.004878 **
X1             6.273      3.444   1.821 0.089969 .
X22           10.596      3.513   3.016 0.009248 **
X23           18.877      3.255   5.799 4.61e-05 ***
X32            8.607      3.706   2.323 0.035782 *
X33           15.021      3.448   4.356 0.000658 ***
X41            1.385      4.695   0.295 0.772291
X42            4.436      4.002   1.108 0.286330
X43            1.496      4.515   0.331 0.745355
X44           -3.281      8.638  -0.380 0.709769
X45            5.234      7.851   0.667 0.515810
X46            6.527      8.722   0.748 0.466649
X47           -2.227      5.532  -0.403 0.693282
X48            5.126      5.891   0.870 0.398896
X49           13.797      6.684   2.064 0.058034 .
X5             1.099      2.025   0.543 0.595766
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.164 on 14 degrees of freedom
Multiple R-squared: 0.8425,     Adjusted R-squared: 0.6738
F-statistic: 4.993 on 15 and 14 DF,  p-value: 0.002261
```

Interpretation:

In the case of multiple regressions the summary function reports far more information than in simple regression. For this particular test we are only interested in the very last line, in which an F statistic of 4.933 and a p-value of .002261 are reported. Those values are the values specifically for the over-all test. In this case the p-value is less than .05, meaning we reject the null hypothesis that all β=0.

Now that we have determined that at least one variable has an effect we can now try to determine which of the specific variables is having an effect.
The summary function has reported a complex summary due to the fact that we have specified many of the variables as factors. As such it has tried to separate out the factors into a separate factor for each of the levels. For each of these variables, a partial t-test is run. This partial t-test is roughly equivalent to the t-test for β=0 for simple regression seen in the simple regression. However, in multiple regressions it does this for each of the variables in a marginal manner, meaning it does it for one independent of the others, then for the next, and so on. The null hypothesis is also identical for the test in the summary of the simple regression.

In this manner we can look at the summary table and determine which variables we think are significant based on which marginal tests generated a p-value less than .05. From our summary table it looks as though the most important variables are Rank and Dept. Within these variables the differences in the treatment classes are also significant.

The data for a multiple regression can also be analyzed via the anova function.
>anova(LM)

```
Analysis of Variance Table
Response: Y
         Df  Sum Sq Mean Sq F value    Pr(>F)
X1        1   11.01   11.01  0.2898 0.5988122
X2        2 1357.93  678.97 17.8714 0.0001399 ***
X3        2  935.73  467.86 12.3148 0.0008212 ***
X4        9  529.62   58.85  1.5489 0.2232873
X5        1   11.20   11.20  0.2947 0.5957665
Residuals 14  531.89   37.99
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Description of multiple regressions in R:

In multiple regressions R reports a very different set of results than in simple regression. The anova function in multiple regressions runs a serial test, which refers to a test in which one variable is added and then the linear models with and without that variable is compared. A serial test does this for each variable, adding each variable one at a time. In this manner the serial test compares the linear model of Y to Y~X1, then Y~X1 to Y~X1+X2 then Y~X1+X2 to Y~X1+X2+X3 etc. until it reaches the last variable. The anova function runs the serial test in the order that you specified when you created your linear model. The F and p-values generated reflect whether the addition of that variable affected the linear model in a significant manner.

From this anova we can determine that Years and Merit may be unimportant to the prediction of Salary.

There also exists a marginal alternative to the serial test of the anova function. This is called the Anova function. This function is not found pre-loaded in R. It is necessary to install and load the car package to use this function.

```
>Anova(lm)
Analysis of Variance Table

Response: Y
         Df  Sum Sq Mean Sq F value    Pr(>F)
X1        1   11.01   11.01  0.2898 0.5988122
X2        2 1357.93  678.97 17.8714 0.0001399 ***
X3        2  935.73  467.86 12.3148 0.0008212 ***
X4        9  529.62   58.85  1.5489 0.2232873
X5        1   11.20   11.20  0.2947 0.5957665
Residuals 14  531.89   37.99
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpretation:

The marginal test compares the full linear model we specified to reduced versions of the model. The first row labeled X1 is the comparison of the full model to the same model without the variable X1. Each other row is the comparison for when that variable is removed and all the others remain. From this table we can determine similar results as from the anova table; Years and Merit appear not to be useful in the model.

Choosing an Optimal Model:

The anova function and the Anova function are examples of methods by which the full linear model specified is compared to reduced versions to try and determine if some variables in the model may be unnecessary. The optimal linear model is one in which only the variables which are deemed necessary remain. It is the simplest model possible that still has the most predictive power.

For this example the goal is to generate an optimal model that predicts Salary without unnecessary variables. As outlined above Years and Merit look as though they may be unnecessary in the model. There exists other ways to determine more quantitatively which model is more optimal.

One of these is the Akaike's Information Criterion (AIC). By measuring certain changes between a full model and a reduced model and generating an AIC value for each it is possible to say which of the models is more optimal. AIC values can be calculated via a long-hand method in R; however R has three very useful functions for generating AIC values quickly. When interpreting

AIC values, the best model is the model with the lowest AIC value. At times this is the AIC value that is most negative.

The AIC value can be calculated explicitly for any model via the extract AIC function

```
>extractAIC(LM)
[1]  16.0000 118.2569
```

We can compare this value to AIC values for any reduced model that we wish to specify. For example let us look at two reduced models:

```
>RM1=lm(Y~X1+X2+X3+X4)
>extractAIC(RM1)
[1]  15.0000 116.8818
```

```
>RM2=lm(Y~X1+X2+X3)
>extractAIC(RM2)
[1]   6.0000 119.3021
```

Interpretation:

Extract AIC reports both a degrees of freedom and the AIC value, the AIC value is the value on the right. From this comparison we can see that removing Merit, in comparison to the full model, gives a lower AIC value. However, if both Merit and Years are removed together, the AIC value is higher than that of the full model. In this case we might say that either Years or Merit are necessary variables but not both.

The drop1 function is a marginal test that removes the variables one at a time, generates and then generates an AIC value for the reduced model of losing that one variable.

```
>drop1(LM)

Single term deletions

Model:
Y ~ X1 + X2 + X3 + X4 + X5
       Df Sum of Sq     RSS    AIC
<none>               531.89 118.26
X1      1    126.05  657.93 122.64
X2      2   1479.11 2010.99 154.16
X3      2    723.24 1255.12 140.01
X4      9    325.03  856.91 114.56
X5      1     11.20  543.08 116.88
```

Interpretation:

In this case we can see that the AIC values that are lowest are for removing Years and Merit. This agrees with the Anova function tests earlier.

There is also an automated version of the drop1 function called the step function.

```
>step(LM)
Start:  AIC=118.26
Y ~ X1 + X2 + X3 + X4 + X5

       Df Sum of Sq     RSS    AIC
- X4    9    325.03  856.91 114.56
- X5    1     11.20  543.08 116.88
<none>               531.89 118.26
- X1    1    126.05  657.93 122.64
- X3    2    723.24 1255.12 140.01
- X2    2   1479.11 2010.99 154.16

Step:  AIC=114.56
Y ~ X1 + X2 + X3 + X5

       Df Sum of Sq     RSS    AIC
<none>               856.91 114.56
- X5    1    215.79 1072.70 119.30
- X1    1    398.46 1255.37 124.02
- X3    2   1137.29 1994.20 135.90
- X2    2   1730.96 2587.87 143.72

Call:
lm(formula = Y ~ X1 + X2 + X3 + X5)

Coefficients:
(Intercept)           X1          X22          X23          X32          X33
     18.069        9.186        9.975       19.340        9.501       16.229
         X5
      3.035
```

Interpretation:

The step function runs a drop1, chooses the reduced model that has the lowest AIC value, and then runs a drop1 on that reduced model, treating it as though it were a new full model. The step function continues doing this until it finds that the 'full' model for that particular step has the lowest AIC value. In this example the step function removed X4 from the model based on the first drop1, and then in the second step the 'full model', a reduced model of Y~X1+X2+X3+X5, a removal of the variable years, gives the lowest AIC value and is thus the most optimal model. This agrees with the conclusions drawn from explicitly calculating AIC values above.

GLM F-test

Another method that can be used to compare a full model to a reduced model without using AIC values is the GLM F-test using the anova function. This test compares a full model to a reduced model under the following null hypotheses:

Ho: coefficients, or β, in the full model but NOT INCLUDED in the reduced model = 0
H1: at least one coefficient, or β, in the full model but not in the reduced model $\neq 0$

In simpler terms, the null hypothesis is that all the variables that were excluded from the reduced model have no significant effect on the prediction of the independent variable, as defined by the β being 0. Acceptance of the null hypothesis is effectively acceptance of the reduced model as a more optimal model.

To run this test we must specify a full model and a reduced model, and then run an anova function.

```
>FullModel=LM
>ReducedModel=lm(Y~X1+X2+X3+X5)
>anova(ReducedModel,FullModel)
```

For this Analysis of Variance Table

```
Model 1: Y ~ X1 + X2 + X3 + X5
Model 2: Y ~ X1 + X2 + X3 + X4 + X5
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     23 856.91
2     14 531.89  9    325.03 0.9506  0.515
```

Intepretation:

For this test a p-value greater than .05 was generated, signifying acceptance of the null hypothesis. The reduced model is a more optimal model.

Conclusion:

Using a multitude of different tests to determine which is a more optimal model we were able to prove that a reduced model using only the variables Gender, Rank, Dept, and Merit has effectively the same predictive power as the full dataset given at the start that also included the variable Years. Multiple regressions is a very powerful tool to take a large dataset and simplify it into a more parsimonious set. Moreover, we were also able to demonstrate that Multiple regression is able to demonstrate significance in variables that when taken alone do not appear significant but when put in the context of a larger set of data have an impact on the response variable. In the simple regression we were not able to demonstrate a significant impact of gender alone on Salary, however, in the multiple regression Gender consistently proved to be a variable that was necessary to the analysis of the data. Variables that may appear unimportant can prove to have a great effect in the context of more data and as such it is always more useful to use a multiple regression to determine an optimal model rather than trying to use many simple regressions to throw out certain variables.