# MAT 303 Module Three Problem Set Report

Second Order Models

Brandon Hobbs

Brandon.Hobbs@snhu.edu

Southern New Hampshire University

## Introduction

The data set being analyzed consists of 99 rows and 6 columns. Each row contains data about a particular economic period, e.g., *wage\_growth*, *inflation*, *gdp*, etc. See Figure 1 for the first few rows for data.

The data will be used to build a second order linear regression model with the purpose of predicting wage growth (wage\_growth) from the other available data.  
  
First, the data in the csv-file will be ingested into a dataframe so the R-language may be used for the stated purpose. Next, it will be plotted to provide a sense of the data and then the regression models, and their appropriateness, will be calculated. Finally, the models will be used to make predictions.

Table

Description automatically generated

**Figure 1: First 5 Rows of Data used for Analysis**

## Data Preparation

To begin the analysis the data, all 99 rows and 6 columns, were imported into a data frame for consumption in the R-language. Of particular interest are the *wage\_growth*, *unemployment, gdp,* and *economy*. Three regression models will be created.

One model will try and predict the wage growth from unemployment.

The second model will try and predict the wage growth from *unemployment* and *gdp*. The model will use a complete second order model and contain the first- and second-order and interaction terms.

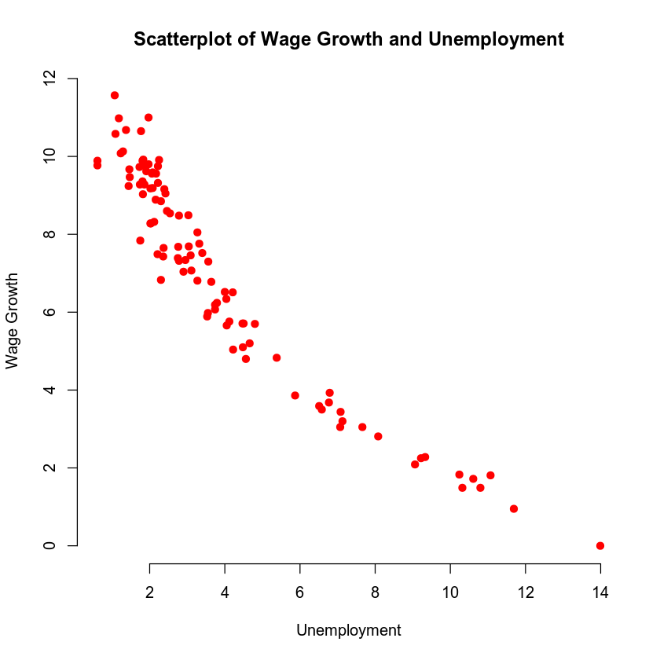
The last model will try and predict the wage growth from *unemployment* and *economy*. The variable economy is a qualitative variable. The model will use a complete second order model and contain the first- and second-order and interaction terms.

## First Second-Order Model Using Only Unemployment

This model will try and predict the wage growth from unemployment. It will be a second order polynomial.

### Correlation Analysis

To begin, *wage\_growth* was plotted versus *unemployment* to see if any trends could be spotted visually, Figure 1.



**Figure 1: Scatterplot of *wage\_growth* versus *unemployment***

From visual inspection the data appears to be non-linear, first-order would not be appropriate, and have an upwards concavity. The upwards concavity would suggest that the model should have a positive second-order term.

### Reporting Results

The previous section showed that *wage\_growth* and unemployment appear to be non-linearly related. This suggest that a higher order variable is needed. This model will be of the form:

With the final model being:

With *unemployment* as X1. The positive (+)0.67 confirms the suspicion from visual inspection and upwards concavity.

This model has a coefficient of determination (*R2*) of 0.944 – meaning that 94.4% of the variability in mpg is explained by the predictor variables. The model also has an adjusted *R2* of 0.942. The adjusted *R2* tends to only increase when a worthwhile predictor variable is added. This value should not be used in isolation but could be used if a new predictor variable, e.g., *gdp*, was added to the model. The adjusted *R2* could evaluate if it was a valuable addition.

To further determine if the model was relevant an F-test is used. An F-test is run to determine if there is indeed an association between the predictor variables and the response variable. First, the null hypothesis (*H0*) and alternative hypothesis (*Ha*) are created:

*H0: β1 = β2 =0*

*Ha: At least one βi ≠ 0 for i = 1 to 2*

The null hypothesis states that *β1* through *β2* are zero; meaning there is no correlation between *unemployment* and *wage\_growth*. The alternative states at least one beta term, *β1* or *β2,*are not zero; meaning there is a correlation between *unemployment* and *wage\_growth*. This will be evaluated against an α of 5% or a 95% confidence interval. Table 2 shows the F-Test statistic and its associated P-value:

**Table 2: Hypothesis Test for the Overall F-Test**

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 803 |
| P-value | 2.2E-16 |

The P-value confirms that the null value may be rejected, 2.2E-16 << 0.05; thus, at least one variable is linearly correlated to wage growth. Moreover, this further confirms that the model shown above is valid at the 95% confidence level.

What the F-test does not reveal is how many of the predictor variables are relevant or which ones. To determine which variables are relevant an individual t-test is conducted on each variable. Each t-test will have a similar null hypothesis and alternative hypothesis. The null hypothesis and alternative hypothesis will be of this form:

*H0: βi =0*

*Ha: βi ≠ 0 for i = 1…n*

As before, the null hypothesis states that *βi* is zero; meaning there is no correlation between its predictor variable and *wage\_growth*. The alternative states that *βi* is not zero; meaning there is a correlation between its predictor variable and *wage\_growth*. Based on these hypotheses the P-values can be used to determine statistical relevance, see Table 3.

**Table 3: T-test for Individual Predictor Variables**

| **Variable** | **P-Value** |
| --- | --- |
| *unemployment* | 2E-16 |
| *unemployment2* | 6.07E-15 |

All P-values are less than the 5% significance level, i.e., P-value << 0.05. Therefore, all variables, are shown to be statistically relevant.

### Making Predictions Using the Model

With the new model created and confirmed as relevant it is useable for predictions. As an example, the expected wage growth when the unemployment rate is 2.54 is 8.241.

However, the value of 8.241 is the mean and someone can be 95% confidant that the actual average growth exists in the range of 8.094 to 8.39 for all periods with that unemployment figure.

Due to the uncertainty in estimating the mean value and the random variation in what was already observed someone could be certain, to 95%, that a single period with the same unemployment figure will have a growth between 6.907 and 9.576. This wider range is known as the prediction interval. The prediction interval is wider because it considers the variability of the individual points around the predicted mean in addition to the uncertainty in sampling.

## Second-Order Regression Model Using *unemployment* and *gdp*

The second model will try and predict the wage growth from *unemployment* and *gdp*. The model will use a complete second order model and contain the first- and second-order and interaction terms.

### Reporting Results

As before, this model is expected to need higher order terms and will, therefore, use the complete, second-order form for the model. The model will be of this form:

With the final model being:

With *unemployment* as X1 and *gdp* as X2.

This model has a coefficient of determination (*R2*) of 0.959 – meaning that 95.9% of the variability in mpg is explained by the predictor variables. The model also has an adjusted *R2* of 0.957. The adjusted *R2* tends to only increase when a worthwhile predictor variable is added. This value should not be used in isolation but could be used if a new predictor variable, e.g., *economy*, was added to the model. The adjusted *R2* could evaluate if it was a valuable addition.

To further determine if the model was relevant an F-test is used. An F-test is run to determine if there is indeed an association between the predictor variables and the response variable. First, the null hypothesis (*H0*) and alternative hypothesis (*Ha*) are created:

*H0: β1 = β2 =…=0*

*Ha: At least one βi ≠ 0 for i = 1 to 5*

The null hypothesis states that *β1* through *β5* are zero; meaning there is no correlation between *unemployment* and *gdp* and *wage\_growth*. The alternative states at least one beta term, *β1* through *β5,*are not zero; meaning there is a correlation between at least one predictor variable and *wage\_growth*. This will be evaluated against an α of 5% or a 95% confidence interval. Table 2 shows the F-Test statistic and its associated P-value:

**Table 2: Hypothesis Test for the Overall F-Test**

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 432 |
| P-value | 2.2E-16 |

The P-value confirms that the null value may be rejected, 2.2E-16 << 0.05; thus, at least one variable is linearly correlated to wage growth. Moreover, this further confirms that the model shown above is valid at the 95% confidence level.

What the F-test does not reveal is how many of the predictor variables are relevant or which ones. To determine which variables are relevant an individual t-test is conducted on each variable. Each t-test will have a similar null hypothesis and alternative hypothesis. The null hypothesis and alternative hypothesis will be of this form:

*H0: βi =0*

*Ha: βi ≠ 0 for i = 1…n*

As before, the null hypothesis states that *βi* is zero; meaning there is no correlation between its predictor variable and *wage\_growth*. The alternative states that *βi* is not zero; meaning there is a correlation between its predictor variable and *wage\_growth*. Based on these hypotheses the P-values can be used to determine statistical relevance, see Table 3.

**Table 3: T-test for Individual Predictor Variables**

| **Variable** | **P-Value** |
| --- | --- |
| *unemployment* | 8.26E-06 |
| *unemployment2* | 0.00489 |
| *gdp* | 0.04682 |
| *gdp2* | 0.12815 |
| *unemployment/gdp* | 0.76678 |

All P-values are less than the 5% significance level, i.e., P-value << 0.05, except the square of *gdp* and the interaction term. A third model, like this one, but without the square of gdp and the interaction term (all P-values > 0.05) should be created and tested.

### Making Predictions Using the Model

With the new model created and confirmed as relevant it is useable for predictions. As an example, the expected wage growth when the unemployment rate is 2.5 and gdp is 6.5 is 7.806.

However, the value of 7.806 is the mean and someone can be 95% confidant that the actual average growth exists in the range of 7.583 to 8.029 for all periods with those gdp and unemployment figures.

Due to the uncertainty in estimating the mean value and the random variation in what was already observed someone could be certain, to 95%, that a single period with the same unemployment and gdp figures will have a growth between 6.632 and 8.981. This wider range is known as the prediction interval. The prediction interval is wider because it considers the variability of the individual points around the predicted mean in addition to the uncertainty in sampling.

## Second-Order Regression Model Using *unemployment* and *economy*

The last model will try and predict the wage growth from *unemployment* and *economy*. The variable economy is a qualitative variable. The model will use a complete second order model and contain the first- and second-order and interaction terms.

### Reporting Results

The first step in building the model is to relevel the qualitative predictor, *economy*. In the data there are two possible values of only a single variable is needed to represent the data – a 1 indicates that a quality exists. Table 5 shows these values:

**Table 5: Qualitative Predictors with ‘no\_recession’ as the reference**

|  |  |
| --- | --- |
|  | **recession** |
| **no\_recession** | 0 |
| **recession** | 1 |

A linear regression model was created with *unemployment and economy* as the predictor variables and *wage\_growth* as the response variable. This model will be of the form:

With the final model being:

With *unemployment* as X1 and qualitative variable [0,1] *economy* as X2. This model also includes the interaction term *economy/unemployment*. The term is not present in the calculated model since X2 represents a qualitative variable and squaring 1 is not relevant.

This model has a coefficient of determination (*R2*) of 0.945 – meaning that 94.5% of the variability in wage growth is explained by the predictor variables. The model also has an adjusted *R2* of 0.943. The adjusted *R2* tends to only increase when a worthwhile predictor variable is added. Since the adjusted *R2*of this model is lower than the previous model it suggests the new variables were not valuable additions.

To further determine if the model was relevant an F-test is used. An F-test is run to determine if there is indeed an association between the predictor variables and the response variable. First, the null hypothesis (*H0*) and alternative hypothesis (*Ha*) are created:

*H0: β1 = β2 =…= βn = 0*

*Ha: At least one βi ≠ 0 for i = 1 to n*

From before, the null hypothesis states that *β1* through *β5* are zero; meaning there is no correlation between the predictor variables and *mpg*. The alternative states at least one beta term, *β1* through *β5,*are not zero; meaning there is a correlation between at least one predictor variable and *mpg*. This will be evaluated against an α of 5% or a 95% confidence interval. Table 6 shows the F-Test statistic and its associated P-value:

**Table 6: Hypothesis Test for the Overall F-Test**

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 405.8 |
| P-value | 2.2E-16 |

The P-value confirms that the null value may be rejected, 2.2E-16 << 0.05; thus, at least one variable is linearly correlated to *wage\_growth*. Moreover, this further confirms that the model shown above is valid at the 95% confidence level.

What the F-test does not reveal is how many of the predictor variables are relevant or which ones. To determine which variables are relevant an individual t-test is conducted on each variable. Each t-test will have a similar null hypothesis and alternative hypothesis. The null hypothesis and alternative hypothesis will be of this form:

*H0: βi =0*

*Ha: βi ≠ 0 for i = 1…n*

As before, the null hypothesis states that *βi* is zero; meaning there is no correlation between its predictor variable and *wage\_growth*. The alternative states that *βi* is not zero; meaning there is a correlation between its predictor variable and *wage\_growth*. Based on these hypotheses the P-values can be used to determine statistical relevance, see Table 7.

**Table 7: T-test for Individual Predictor Variables**

| **Variable** | **P-Value** |
| --- | --- |
| *unemployment* | 2E-16 |
| *unemployment2* | 1.77E-6 |
| *economy* | 0.135 |
| *unemployment/ economy* | 0.263 |

Two of the P-values are greater than the 5% significance level, i.e., P-value >> 0.05, therefore, this model should be evaluated carefully. A third model, similar to this one, but without the *economy* and interaction terms (all P-values > 0.05) should be created and tested – this is left for a future date.

### Making Predictions Using the Model

With the new model created and confirmed as relevant it is useable for predictions. As an example, the wage growth with the economy not in recession and an unemployment of 2.50 is predicted to have a growth of 8.340 units.

However, as before, the value of 8.340 is the mean and someone can be 95% confidant that the actual average growth exists in the range of 8.184 to 8.50 units for all periods with thoseunemployment rates and no recession.

Due to the uncertainty in estimating the mean value and the random variation in what was already observed someone could be certain, to 95%, that a single period with the same specs will have a growth between 7.010 and 9.670 units. Again, this wider range is known as the prediction interval, see above for explanation.

## Conclusion

This report details the creation of three second-order linear regressions. All models attempt to predict the growth of wages from different variables. To determine if the model was statistically relevant an F-test and individual t-tests were conducted on the models.

Both the F- and t-tests showed that the beta values predicted were relevant within a 95% confidence range. Because both tests confirmed the statistical relevance, the model is recommended for usage. However, even though all models were shown to be relevant one model may be most appropriate.

The model made using unemployment rate and gdp had the largest *R2* and adjusted *R2* values, see Table 8, suggesting that it accounts of the greatest variability in wage growth and that the added terms are valuable.

**Table 8: R2 and Adjusted R2 Values for all Models**

|  |  |  |
| --- | --- | --- |
| **Model** | ***R2*** | ***Adjusted R2*** |
|  | 0.944 | 0.942 |
|  | 0.959 | 0.957 |
|  | 0.945 | 0.943 |

Having created and tested the models they could be used for anyone reviewing an economic period to see if its growth in wages is under or outperforming prior years.

## Citations

Hobbs, B. (2022). *MAT 303 module one summary report*. [Unpublished report]. SNHU.

Hobbs, B. (2022). *MAT 303 module two summary report*. [Unpublished report]. SNHU.