EC315 Assignment:

Statistical Regression Analysis in Gretl

Lewis Britton {201724452}

EC315: Topics in Microeconomics With Cross Section Econometrics

November 2019

Word Count: {N/A}

Group Assignment: Results

1: Background About Issue

The goal here is to make an attempt to find a best fitting model based on a series of linear, polynomial and logarithmic variables and dummy variables, best explaining the variation in a citizen's level of worry of being a victim of crime – the dependent variable. We observe low scores as low levels of worry and high scores as high levels of worry. The survey carried out uses a sample size of 35,371 with a variable amount being disregarded conditionally on a negligibility basis.

- Sample Information:
  - Location: England & Wales
  - Size of Observation: 35,371 (Variable)
- Age of Data:
  - 2013-14 (5-6 Years)
- Source of Data:
  - UK Data Service: Crime Survey for England & Wales

2: Dataset Details

$$worryx = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + u$$

Intuition was first used to determine which variables were likely to affect degree of crime worry. This number totalled roughly fifty. It was then important to identify statistically significant variables in the context of explanation of the dependent variable. This was possible through the observation of associated p-values when regressing these in the context of the dependent variable, individually. To determine statistical significance, a p-value of less than 0.01 must be achieved for significance at the 1% level, less than 0.05 for the 5% level and 0.1 for the 10% level. Anything out-with these observations was disregarded. These variables also produced the highest R^2 values and combined to show stronger comparisons (higher values) of the Schwartz Criterion. Our criteria of choice.

We observe from the results, in the subsequent document, seventeen variables which are statistically significant, out of a possible fifty, and are as follows:

- sex
- genhealt
- agegrp7
- rural2
- confx

Note that the specific "worry" variables were omitted from analysis due to the fact that our dependent variable (*worryx*) was derived from these variables. Therefore, when regressing *worryx* against all of the "worry" explanatory variables, there is a perfect R^2 value of 1 (100%) produced. These were clearly also highly correlated and we know this ideally leads to omission. This is therefore a weak and zero-explanatory method of explaining *worryx*.

We must also have a basic understanding of the intuition behind the variables. The 'sex' variable can effect crime worry simply due to the difference in the nature of crime between males and females. For example, females are more prone to rape etc. and males may have more of an ignorant mindset involving violence as a response to attack etc. The 'health' variable is pretty self-explanatory where you would see less healthy and less physically able people worry more as physical crime or simply response to crime could impact them more. Expect a negative correlation with worry. The 'age group' variable is also quite self-explanatory where you would expect older people to worry more about crime as their response would be limited. Expect a positive correlation with worry. Within the 'rural/urban' variable, we would expect to see urban areas worry less due to more advanced security, regulation and policing. Despite the increased population and general crime in these areas. Finally, the 'confidence in police' variable is the most self-explanatory of them all. If people believe less in their police, they don't believe they have good protection from crime. Expect a negative correlation with worry.

3: Descriptive Analysis of Dependent Variable

Here we are observing *worryx* as the dependent or 'explained' variable. This means that we are seeking variables in attempt to explain the degree to which citizens worry about being a victim of crime. Therefore, as the dependent variable relying on any linear or polynomial explanatory variables, the dependent variable *worryx* is displayed on the *y* axis.

Simply for means of ease, not statistical advantage, we eliminate any negligible observations where people have not answered or answered incorrectly etc. by stating a restriction where the variable *worryx* is equal to any **real number** ($\mathbb{R}$). This reduced the number of observations from 35,371 to 8,183 and displayed the (1,N) matrix in a more simplistic form.

Note: $\{-\infty \leq worryx \leq \infty\}$

We observe the summary statistics of *worryx* as:

Mean	-2.1330e-10
Median	0.22631
Minimum	-2.9024
Maximum	1.3885
Standard deviation	1.0000
C.V.	4.6882e+09
Skewness	-0.91221
Ex. kurtosis	0.63565
5% percentile	-2.0699
95% percentile	1.3885
Interquartile range	0.99016
Missing obs.	27188

As this *worryx* variable was derived from various other "worry" variables, the scale is slightly off-norm however, an increase in value still represents an increase in worry. This is shown with minimum and maximum "worry" values of -2.9024 and 1.3885, respectively.

We observe the Mean to be extremely close to 0 at -2×10^{-10} and the Median at 0.22631.

We observe the Skewness at -0.91221 which is less than 0 meaning there is **Left Skewness**, as observed from the likelihood of this, implied through the maximum and minimum values. This means there is a tail of the distribution to the left. Thus, implies the outliers worry less than the mean but the majority worry more. This is reinforced where the median is shown to be greater than the mean, hence the **Positive Distribution** in relation to skewness.

We observe the Excess Kurtosis at 0.6356 which is greater than 0 and therefore shows **Platykurtic (Negative) Kurtosis** where the distribution lies slightly below the normal distribution and therefore has slightly fatter tails meaning outliers, as has been previously implied. This further enhances the statement of negative outliers, worrying less than the mean value.

4: OLS Regressions – Logic & Hypothesis Testing

1) Dummy Variables
2) Polynomials & Logs
3) Individual t-test Values
    - p-values
    - t-ratios
    - Coefficients
4) Integrity of Model
    - Adjusted R^2 Value
    - Collective f-test Values
    - RESET Test
5) Are the Assumptions Met?

We now want to attempt to best explain the statements seen in the descriptive analysis.

4.1: Dummy Variables

The variable *sex* requires 2 – one real and one hypothetical of binary values 0 and 1, male and female [*sex_male* where female is hypothetical]. As here is no clear description of what gender 1 and 2 represent in in the *sex* variable, it will be assumed: 1 – male, 2 – female. The variable *rural2* also requires a real and hypothetical, rural and urban [*rural2_urban* where rural is hypothetical]. These dummy variables will now be used in the multiple regression in order to compare male worry to female worry and rural worry to urban worry.

4.2: Polynomials & Logs

A number of RESET tests and p-value comparisons took place in order to define which variables required polynomial and logarithmic values. The *sex_male* variable required none due to exact collinearity. The *genhealt* variable responded well to its logarithmic values being used (*genhealt_log*) where it increased the RESET p-value. Using the *agegrp7_poly2* variable with exponent 2, the Ramsay's RESET test returned a p-value closer to 0.05, which we want. The *rural2_urban* variable required none due to exact collinearity. And finally, the *confx_log* variable was used with logarithmic values of *confx* as it also returned a higher p-value, closer to 0.05, in the RESET test. Note that higher R^2 and lower t-test p-values are possible on a small level however, it sacrifices a lot of the RESET p-value.

This produced final model:

```
ols worryx 0 sex_male genhealt_log agegrp7_poly2 rural2_urban confx_log
```

4.3: Integrity of Model

4.3.1: Adjusted R^2 Value

As this is a multiple regression, we must observe the adjusted R^2 value rather than the regular R^2 value. We see from the results that this value was equal to 0.0898 therefore, the collection of explanatory variables selected were responsible for and can statistically explain 8.98% of the variation in the dependent variable, *worryx*. This is a mediocre explanatory rate however, is a strong result in comparison to other combinations as seen previously.

4.3.2: Collective f-test Values

We observe the f-test as a hypothesis test, like a t-test, but for the model as a whole. Therefore, we interpret the associated p-value in the same manner. Here we see an f-test p-value of 4.17×10^{-91}, which is less than 0.01, meaning the model is statistically significant at explaining the dependent variable at the 1% level (99% certainty). And, we can reject the Null Hypothesis that $R^2 = 0$, in favour of the Alternative Hypothesis showing a strong R^2 value.

4.3.3: RESET Test

Using the Ramsay's RESET test in Gretl, we are able to observe how well-specified the regression model is. As discussed previously, RESET values were compared to choose a combination of explanatory variables however now we can experiment with polynomials and logs in search of an associated p-value of greater than 0.05. A p-value of less than 0.05 suggests there's still a need for logarithmic and polynomial values. The model was altered until the closest possible value to 0.05 was met. This is observed at 0.109 in the optimal case here, which is far greater than 0.05. Therefore, we fail to reject the null hypothesis and observe a well-specified model after polynomial and log experimentation.

4.3.4: Heteroscedasticity (White) Test

For this we observe an H_0 that there is homoscedasticity meaning there is constant error variance. Alternatively, we observe an H_A that there is heteroscedasticity meaning there is non-constant error variance. We seek homoscedasticity and therefore want to fail to reject the null hypothesis with a p-value of greater than 0.05. In this case, we observe an associated p-value of 0.0917 which is greater than 0.05 and therefore homoscedasticity is present and we have constant error variance.

4.4: t-test Values

ols worryx **0** sex_male genhealt_log agegrp7_poly2 rural2_urban confx_log

4.4.1: Variables

1) *sex_male*: standard dummy variable for male in comparison to female
2) *genhealt_log*: logarithmic values of a general health rating
3) *agegrp7_poly2*: polynomial values of age group change
4) *rural2_urban*: standard dummy variable for urban compared to rural
5) *confx_log*: logarithmic values of the rating of people's confidence in the police

4.4.2: p-values

The goal here is to confirm statistical significance of each variable by observing the relevant individual p-values. For this we are seeking a p-value of, ideally, less than 0.01 or 0.05.

sex_male, genhealt_log, agegrp7_poly2, rural2_urban, confx_log:

- All < 0.01
- Statistically significant at the 1% level (99% certainty)
- Reject the Null Hypothesis in favour of the Alternative Hypothesis stating that this variable is significant in explaining the dependent variable, with a coefficient ≠ 0
- This reasoning remains the same throughout all variables, as observed…

4.4.3: t-ratios

We see from the results that all of the associated t-ratios are either less than -1.96 or are greater than 1.96. As we have rejected all of the Null Hypotheses at the 1% or 5% significance level, we see that at least 95% of the normal distribution of results lies within 1.96 standard deviations of the mean. Note that the *confx_log* variable's close proximity to -1.96 reinforces the fact that it is observed at the 5% significance level.

4.4.4: Coefficients

sex_male: $\mathbf{coefficient = 0.504}$

- Positive Correlation w/ *worryx*
- "Males worry more than females"
- Males are observed to worry 0.504 units **more** than females in this model

genhealt_log: $\mathbf{coefficient = -0.216}$

- Negative Correlation w/ *worryx*
- "As general health decreases, worry increases"
- For a 1 unit increase in general health, people worry 0.216 units **less**

agegrp7_poly2: $\mathbf{coefficient = 0.006}$

- Positive Correlation w/ *worryx*
- "As age increases, worry increases"
- For a 1 unit increase in age group category, people worry 0.006 units **more**

rural2_urban: $\mathbf{coefficient = -0.321}$

- Negative Correlation w/ *worryx*
- "Urban areas worry less than rural areas "
- Urban areas are shown to worry 0.321 units **less** than rural areas

confx_log: $\mathbf{coefficient = -0.042}$

- Negative Correlation w/ *worryx*
- "As confidence in the police decreases, worry increases"
- For a 1 unit increase in confidence in police, worry decreases by 0.042 units

4.5: Are the Assumptions Met?

- ✓ Dependent variable lies on the best fit line
- ✓ Homoscedasticity is met where observations have constant errors
- ✓ Observations are not heavily correlated (if so – omitted)
- ✓ Errors are normally distributed (not a lot of outliers)
- ✓ Explanatory variables are fixed

Bonus: Limitations

- It is manual labour intensive to find a model which can explain a great amount of the dependent variable. That is, observing a model with a very high R^2 value.
- Cases in which we aim to fail to reject the null hypothesis, such as the RESET and White tests, are very unlikely to show absolute certainty that we can accept the null hypothesis where the p-value is equal to 0.
- In summary, we can interpret as much of the statistics as we want but it does not alter the volume or degree to which the dependent variable is explained.

Group Assignment: Code & Tables

4: Dataset

4.1: Dummy Variables

1) genr sex_male = sex = 1

Where Gretl assumes sex = 2 is not sex_male and therefore in this case female

2) genr rural2_urban = rural = 1

Where Gretl assumes rural2 = 2 is not rural2_urban and therefore in this case rural

4.2: Polynomials & Logs

1) genr sex_male_poly2 = sex_male^2
2) genr sex_male_poly3 = sex_male^3 **Disregarded** Due to Exact Collinearity
3) genr sex_male_log = log(sex_male)

4) genr genhealt_poly2 = genhealt^2
5) genr genhealt_poly3 = genhealt^3 Use: **genhealt_log**
6) genr genhealt_log = log(genhealt)

7) genr agegrp7_poly2 = agegrp7^2
8) genr agegrp7_poly3 = agegrp7^3 Use: **agegrp7_poly2**
9) genr agegrp7_log = log(agegrp7)

10) genr rural2_urban_poly2 = rural2_urban^2
11) genr rural2_urban_poly3 = rural2_urban^3 **Disregarded** Due to Exact Collinearity
12) genr rural2_urban_log = log(rural2_urban)

13) genr confx_poly2 = confx^2
14) genr confx_poly3 = confx^3 Use: **confx_log**
15) genr confx_log = log(confx)

4.3: Attempts

ols worryx 0 sex_male genhealt agegrp7 rural2_urban polatt7 confx_log;
p-value (RESET) = 0.0171

ols worryx 0 sex_male genhealt agegrp7_poly2 rural2_urban confx_log
p-value (RESET) = 0.0369

ols worryx 0 sex_male genhealt_log agegrp7_poly2 rural2_urban confx_log
p-value (RESET) = 0.04