Programming Exercise 7 - Answer Key

Crime Analytics (CJUS 6106)

Instructions

For this exercise, you will be tasked with estimating linear regression models, creating figures describing bivariate and multivariate relationships between select variables, and evaluating some of the important assumptions of linear regression. You will be using the state_data file I have used in prior lectures. Note that this file must be knitted as a PDF. I strongly encourage you to use the .rmd file I provide for the exercise to begin your assignment.

Question 1

Create a variable called **total_crime** within the **state_data** data frame that is the sum of the seven individual crime rate variables. Provide a summary of this new variable using the **summary()** function and interpret its mean. Note that all of these variables are measured as crime rates per 100,000.

```
state_data$total_crime <- with(state_data, rmurd + rrape + rrobb +
                                  raggr + rburg + rlarc + rauto)
summary(state_data$total_crime)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1946    2819    3470    3545    4191    6404
```

Interpretation: The average total crime rate value of 3545 indicates that, within this sample, the total rate of these different types of crime is 3545 per 100,000.

Question 2

Estimate an intercept-only linear regression model with the **total_crime** variable as the outcome. Summarize this model and explain why the intercept value is equivalent to the mean of the **total_crime** rate variable.

```
summary(lm(total_crime~1, data=state_data))
```

```
##
## Call:
## lm(formula = total_crime ~ 1, data = state_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1599.14  -726.84   -75.24   645.76  2858.86
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3545.44      35.24   100.6   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 889.3 on 636 degrees of freedom
```

Interpretation: The intercept value of 3545.44 is very close to the sample average total crime rate - just reported to two decimal places in the regression model. This is because the linear regression model reduces the sum of the squared deviations from a regression line and, without any independent variables in the model, it simply returns the average value for the outcome.

Next, add the following independent variables to the model and provide an interpretation for each coefficient: the percent under the poverty line (**poverty**), the average state resident income in dollars (**avwage**), the percent income share for the top 0.1% of the population (**top_1**), the percent of the state population that is unemployed (**urate**), and the incarceration rate per 1,000 population (**inc_rate**). Also be sure to indicate when a coefficient is significant and at what probability value (e.g., p<.05, p<.01, p<.001). You should also store this regression model as an object because you will need to use these results in another question.

```
lm1 <- lm(total_crime~poverty+avwage+top_1+urate+inc_rate,
          data=state_data)
summary(lm1)
```

```
##
## Call:
## lm(formula = total_crime ~ poverty + avwage + top_1 + urate +
##     inc_rate, data = state_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2005.43  -553.72   -37.43   487.10  2419.08
##
## Coefficients:
##                Estimate  Std. Error t value      Pr(>|t|)
## (Intercept) 3578.224727  220.217273  16.249       < 2e-16 ***
## poverty       15.787046   12.107378   1.304         0.193
## avwage        -0.032498    0.005564  -5.841 0.00000000831 ***
## top_1         14.060111   10.682634   1.316         0.189    
## urate         -1.177123   20.603138  -0.057         0.954    
## inc_rate     227.058558   20.380226  11.141       < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 767.4 on 631 degrees of freedom
## Multiple R-squared:  0.2613, Adjusted R-squared:  0.2554
## F-statistic: 44.63 on 5 and 631 DF,  p-value: < 2.2e-16
```

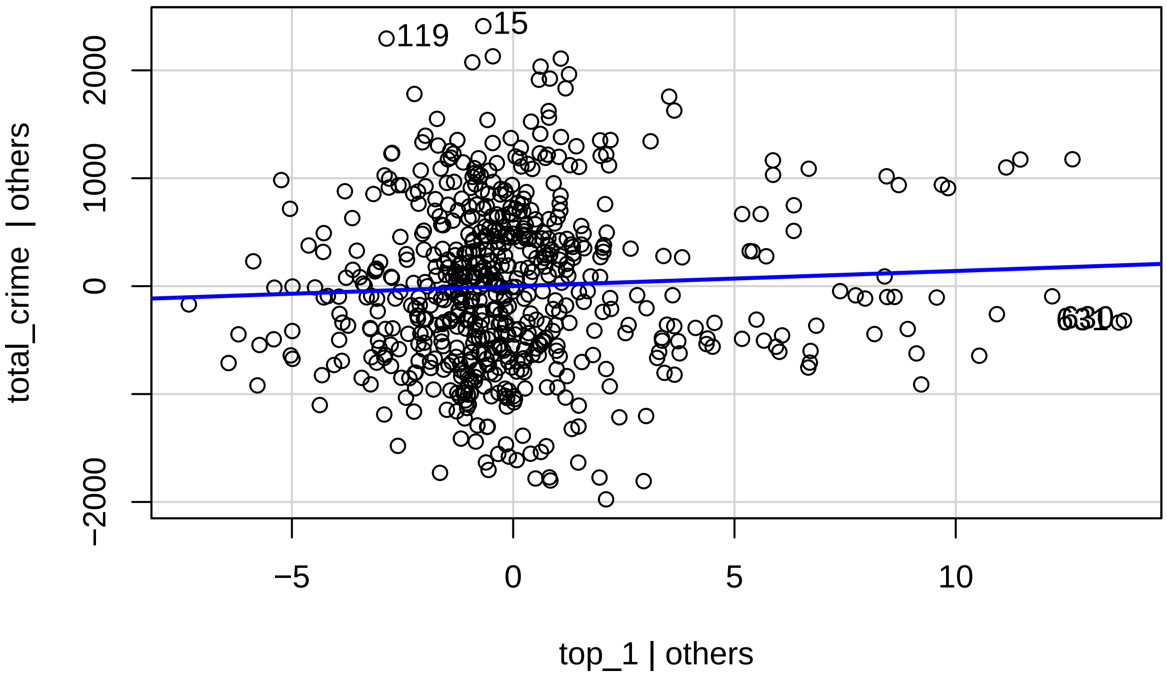
Interpretations:

- poverty: For every one percentage point increase in the state population under the poverty line, we expect there to be a 15.787 crimes per 100,000 increase to the total crime rate, on average and controlling for all other variables. This coefficient is significant at p<.001.
- avwage: For every dollar increase in the average wage for state residents, we expect there to be a 0.032 crimes per 100,000 decrease to the total crime rate, on average and controlling for all other variables in this model. This coefficient is significant at p<.001.
- top_1: For every one percentage point increase in that income share for the top 0.1% of earners, we expect there to be a 14.06 crimes per 100,000 increase to the total crime rate, on average and controlling for all other variables in this model. This coefficient is not significant.
- urate: For every one percentage point increase in the unemployment rate, we expect there to be a 1.177 crimes per 100,000 decrease to the total crime rate, on average and controlling for all other variables in this model. This coefficient is not significant.
- inc_rate: For every one unit increase in the number of state residents incarcerated per 1,000 population, we expect there to be a 227.059 crimes per 100,000 increase to the total crime rate, on average and controlling for all other variables in this model. This coefficient is significant at p<.001.

Question 3

Create an added variable plot using the stored regression model from the prior question and limit the graph output to just the **top_1** variable as I did in lecture.

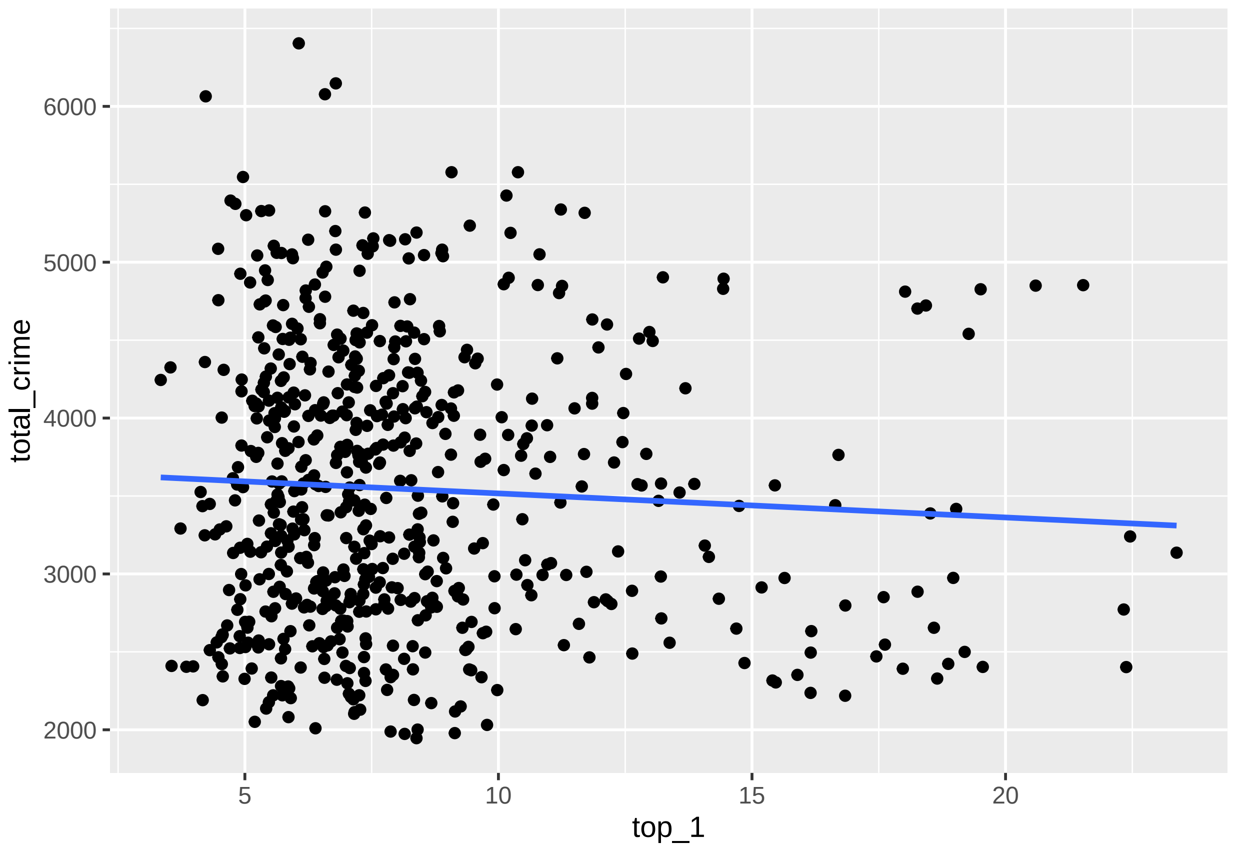
```
avPlots(lm1, terms=~top_1)
```



Now, create a plot using ggplot that depicts the bivariate relationship between the top 0.1% income share and the total crime rate. Be sure to include a best fit line as I do in lecture.

```
ggplot(state_data, aes(x=top_1, y=total_crime)) +
  geom_point() +
  geom_smooth(method="lm", se=FALSE)
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



Finally, comment on any differences you observe in these relationships. Do our inferences about the relationship between these variables change when we control for other variables? If so, how?

Interpretation: In the added variable plot the relationship between the top 0.1% income share and the total crime rate is positive, indicated by the inclining trend line. However, the simple bivariate relationship in the second graph is negative, as indicated by the declining trend line. This tells us that the parts of the relationship between the top 0.1% income share and the total crime rate that are negative overlap with variation in the other variables in the model. When we account for this overlap in the multivariate model, the positive relationship emerges.

Question 4

Create a new variable called **incrate_quint** that is the incarceration rate per 1,000 divided into five equal quintiles. You will need to use the ntiles() function to accomplish this. Be sure to also make this a factor variable.

```
state_data$incrate_quint <- as.factor(ntile(state_data$inc_rate, 5))
```

After you do so, estimate a new regression model predicting the total crime rate using only the quintile variable you just created. Confirm that the coefficients from the different quintiles reproduce the mean total crime rates for each category of this variable. Provide an interpretation for why these values should be the same.

```
summary(lm(total_crime~incrate_quint,
           data=state_data))
```

```
## 
## Call:
## lm(formula = total_crime ~ incrate_quint, data = state_data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -1753.6  -602.7  -111.2   522.2  2564.9 
## 
## Coefficients:
##                Estimate Std. Error t value        Pr(>|t|)    
## (Intercept)     3026.84      70.93  42.674          < 2e-16 ***
## incrate_quint2   295.13     100.31   2.942          0.00338 ** 
## incrate_quint3   472.61     100.51   4.702 0.0000031614616 ***
## incrate_quint4   673.07     100.51   6.697 0.0000000000471 ***
## incrate_quint5  1158.05     100.51  11.522          < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 802.5 on 632 degrees of freedom
## Multiple R-squared:  0.1909, Adjusted R-squared:  0.1858 
## F-statistic: 37.29 on 4 and 632 DF,  p-value: < 2.2e-16
```

```
with(state_data, tapply(total_crime, incrate_quint, mean))
```

```
##        1        2        3        4        5 
## 3026.838 3321.965 3499.445 3699.909 4184.891 
```

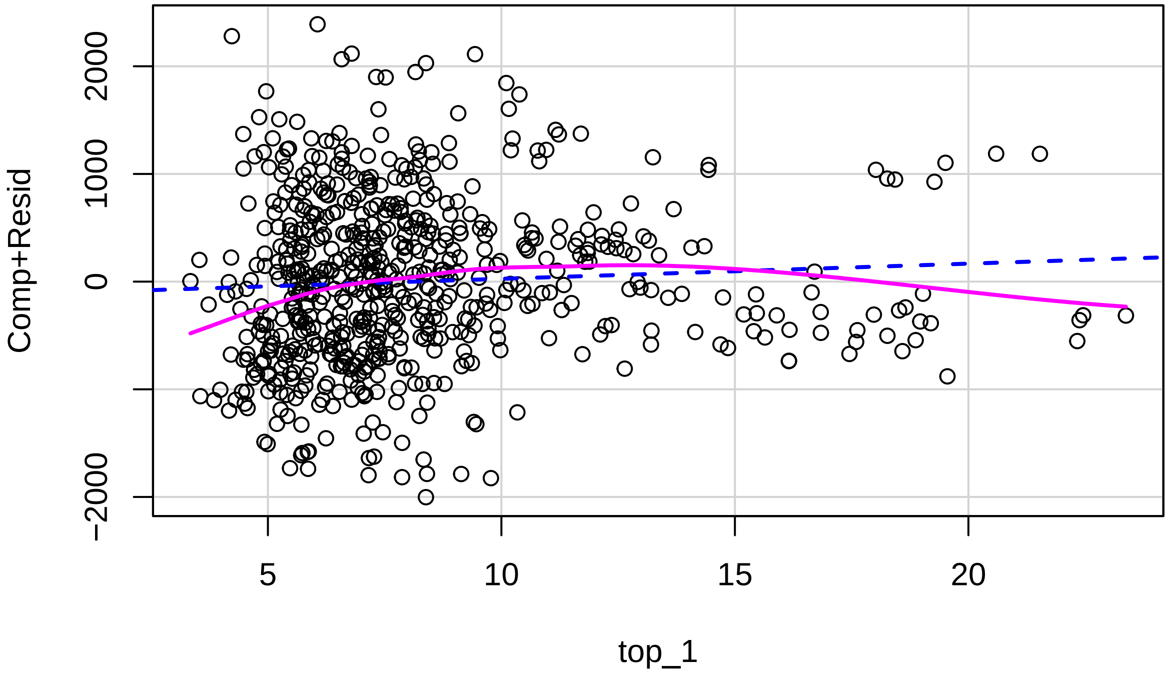
Interpretation: The regression model will return conditional means for the independent variables in the model. When these independent variables represent categories, it will return the average value for the outcome within each of these categories, leaving one category out which is then reflected in the intercept term. The intercept term (3026.84) from this model represents the average total crime rate for states in the lowest 20% of the incarceration rate distribution. Adding the coefficients for each category to this intercept then provides the average total crime rate for the subsequent quintiles of the incarceration rate distribution because the model will return the conditional average given a state belongs in that quintile.

Question 5

For this question, you will return to the multivariate regression model you estimated in Question 2 that you should have stored as an object. Using the results from this model, create a component plus residual plot for the **top_1** regressor.

Provide a comment about whether there is any evidence of nonlinearity in the relationship between **top_1** and **total_crime**.

```
crPlots(lm1, terms=~top_1, line=TRUE, smooth=TRUE, ylab="Comp+Resid")
```



Next, create a squared version of the **top_1** variable within the **state_data** data frame. Re-estimate the regression model from Question 2 including the new squared version of **top_1** (note - there should be six total independent variables in your model now).

```
state_data$top_1sq <- state_data$top_1^2
lm2 <- lm(total_crime~poverty+avwage+top_1+top_1sq+urate+inc_rate,
          data=state_data)
summary(lm2)
```

```
##
## Call:
## lm(formula = total_crime ~ poverty + avwage + top_1 + top_1sq +
##     urate + inc_rate, data = state_data)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max 
## -2100.29  -550.99   -13.66   458.20  2471.70 
## 
## Coefficients:
##                Estimate  Std. Error t value       Pr(>|t|)    
## (Intercept) 3150.631063  265.393948  11.872        < 2e-16 ***
## poverty       13.091249   12.076571   1.084        0.27877    
## avwage        -0.035777    0.005651  -6.331 0.000000000462 ***
## top_1        131.503599   42.529274   3.092        0.00208 ** 
## top_1sq       -5.083792    1.782619  -2.852        0.00449 ** 
## urate          3.501613   20.553244   0.170        0.86478    
## inc_rate     227.960171   20.268465  11.247        < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 763.1 on 630 degrees of freedom
## Multiple R-squared:  0.2707, Adjusted R-squared:  0.2637 
## F-statistic: 38.97 on 6 and 630 DF,  p-value: < 2.2e-16
```

Using the equation from the regression analysis lecture, compute the inflection point for the **top_1** variable where the slope turns to zero. Provide an interpretation of what this value means.

$$\text{Max}(Top1_i) = \frac{-\beta_3}{2 * \beta_4} = \frac{-131.504}{2 * (-5.084)} = 12.933$$

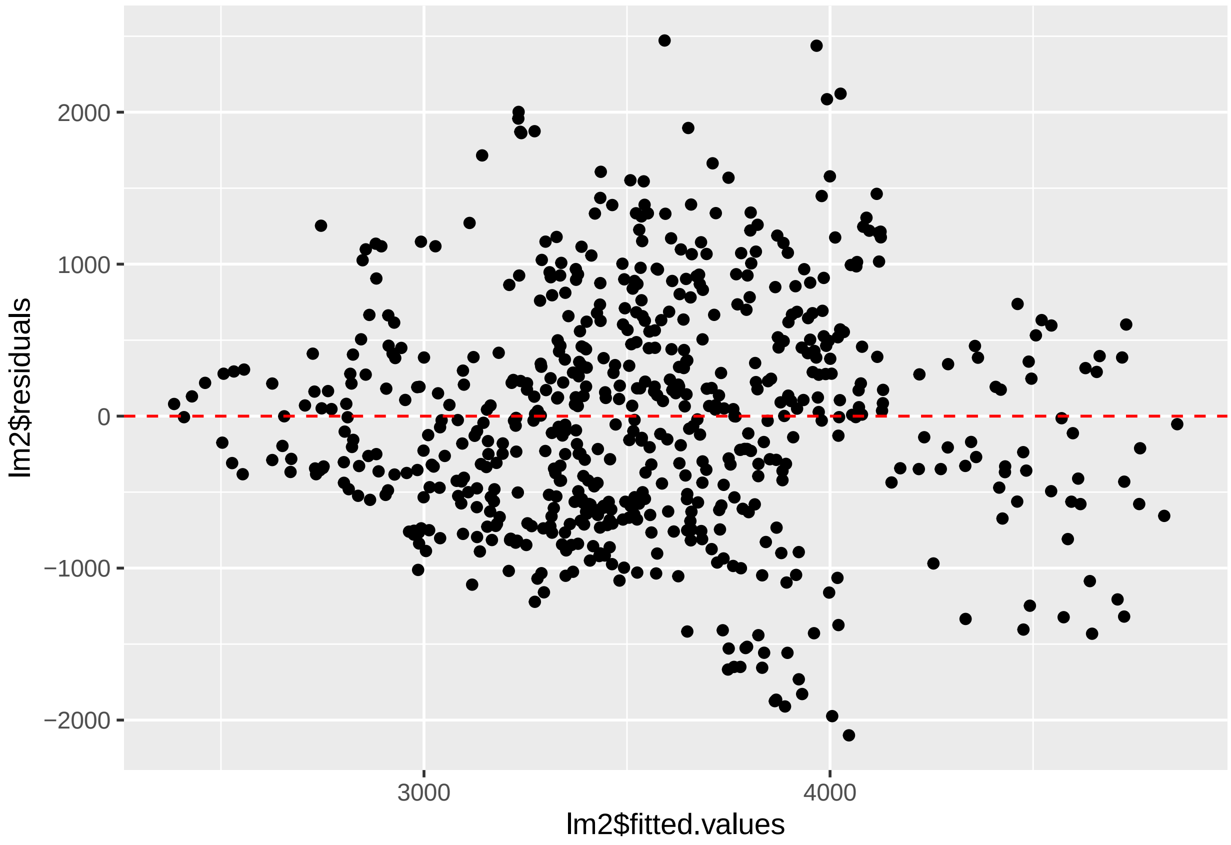
Interpretation: The inflection point value is 12.933, which means that the positive slope in the early part of the relationship between **top_1** and **total_crime** flattens out at this point, and then becomes slightly negative. We can confirm this from examining the component plus residual plot and see that at about a value of 13 on **top_1**, the smoothed line stops increasing.

Question 6

Using the appropriate graphs and statistical tests (demonstrated in the regression analysis lecture), provide an indication for whether the regression model you estimated in Question 5 (with the squared term!) violates either of the assumptions that the residuals are 1) homoscedastic or 2) normally distributed.

Homoscedasticity:

```
ggplot(, aes(x=lm2$fitted.values, y=lm2$residuals)) +
  geom_point() +
  geom_hline(yintercept=0, color='red', linetype='dashed')
```



```
bptest(lm2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lm2
## BP = 72.268, df = 6, p-value = 0.00000000000014
```

```
ols_test_score(lm2)
```

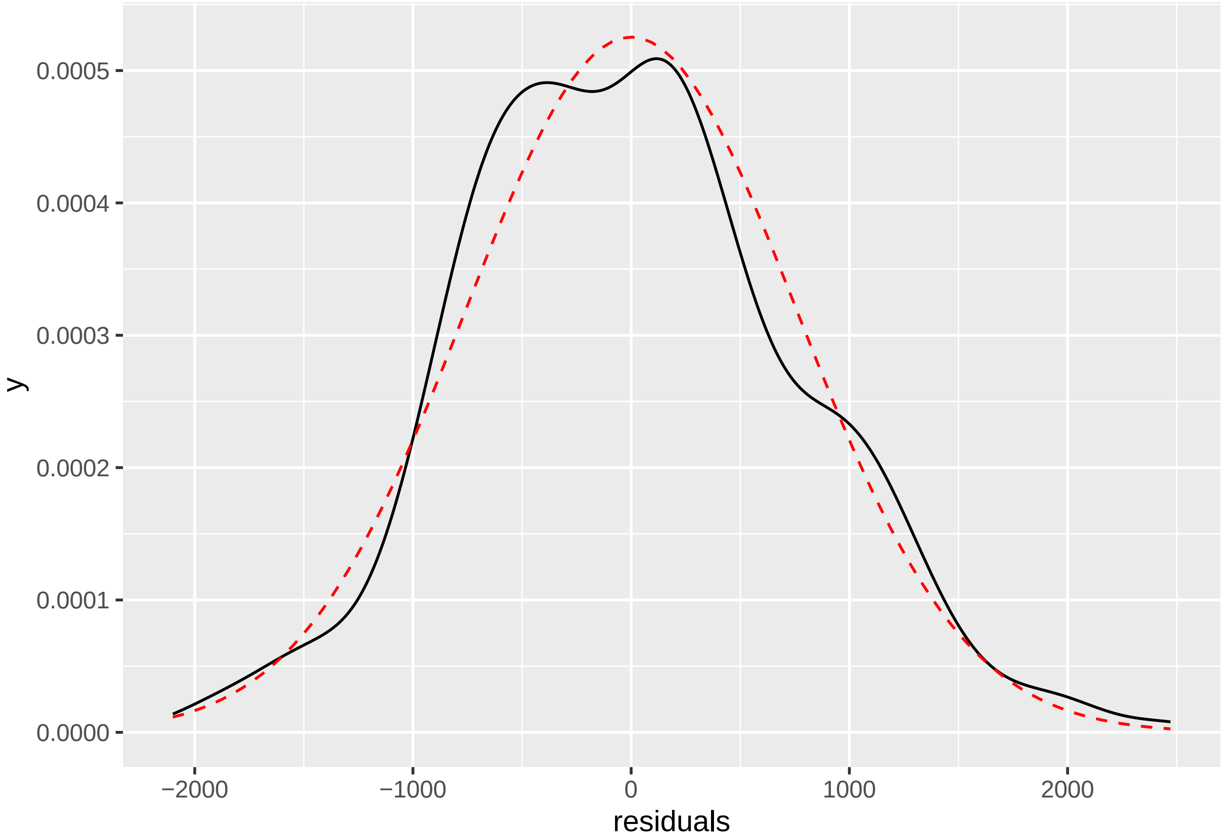
```
##
##  Score Test for Heteroskedasticity
##  ---------------------------------
##  Ho: Variance is homogenous
##  Ha: Variance is not homogenous
##
##  Variables: fitted values of total_crime
##
##          Test Summary
##  -------------------------------
##  DF            =    1
##  Chi2          =    14.75651
##  Prob > Chi2   =    0.0001223242
```

Interpretation: There is some evidence in the figure that the residuals spread out more at mid-range values but cluster tighter to the horizontal line at zero for values that are closer to the minimum and maximum for the distribution of fitted values.

Both the Breusch-Pagan and Score test confirm this. The p-values for both tests are below 0.05, indicating that we should reject the null hypothesis for each test and conclude that the residuals are not homoscedastic.

Normally Distributed:

```
ggplot(data=data.frame(residuals<-lm2$residuals),aes(x=residuals)) +
  geom_density() +
  stat_function(fun=dnorm, color='red', linetype='dashed',
                args=list(mean=mean(residuals),
                          sd=sd(residuals)))
```



```
shapiro.test(lm2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm2$residuals
## W = 0.99439, p-value = 0.01892
```

Interpretation: Judging from the plot, the errors appear to be fairly normally distributed. There are some disparities around residual values of -1000 and +1000 but these appear to be minor. The center is also slightly below the expected value of zero.

However, the Shapiro-Wilk test does indicate these residuals are not normally distributed (p<.05). So, we are forced to conclude that the errors from this model are **not** normally distributed.