**Topic 5 HW**

Alexander Bonnet

Grand Canyon University

DSC - 510

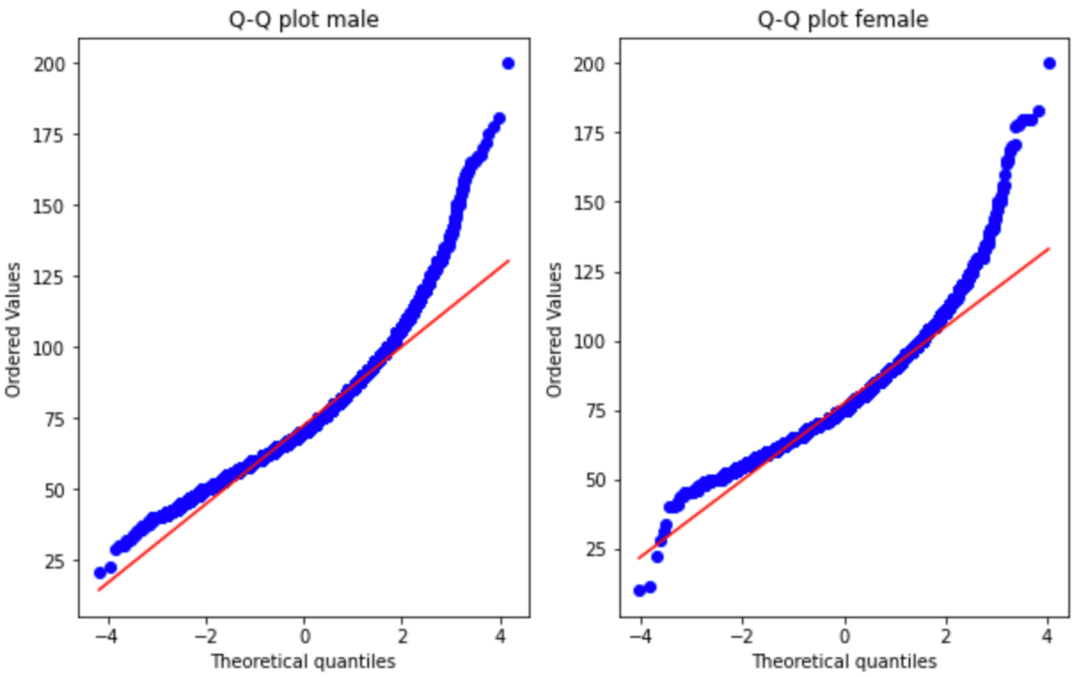
Edward Ofori

3/20/2024

## Part 1: Hypothesis

I found that there is a significant difference in means in my data set. I found this by conducting a t-test for two independent samples, and it was found to have a p-value of basically 0 and a test statistic of -41.62. The test was most likely so definitive because of size of my two groups. There are 45530 males and 24470 females in my data set, so any difference is going to show major significance.

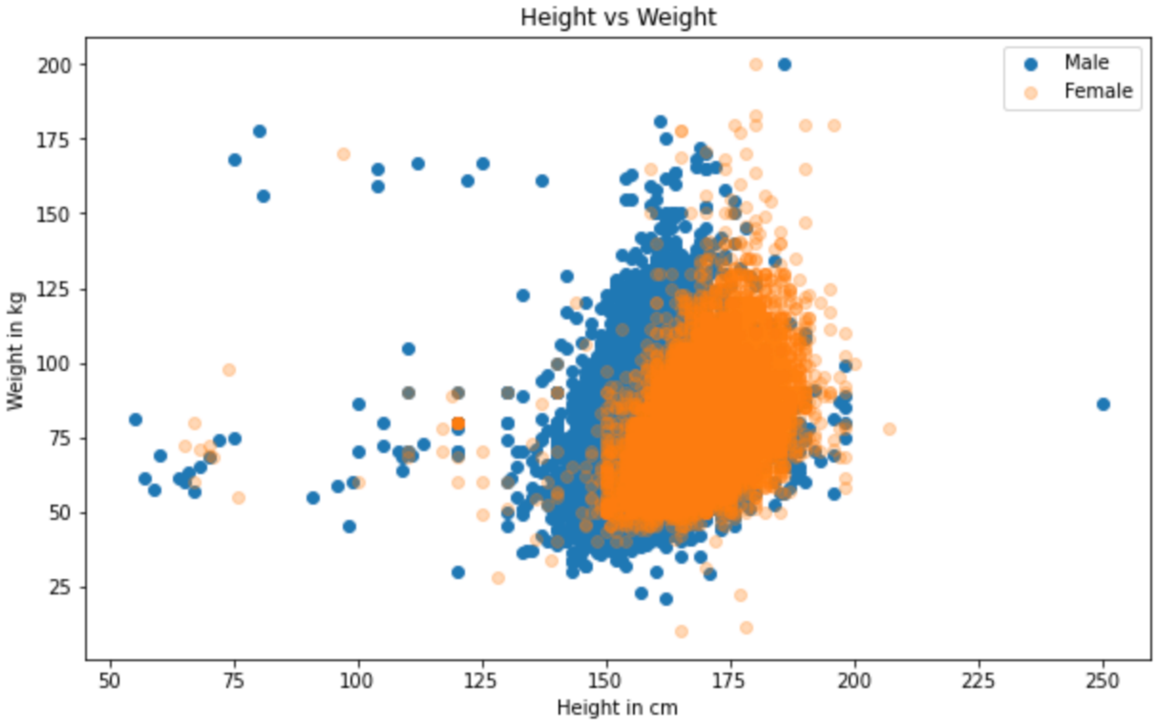
The below plots are the normal quantile plots that I created to see if the two sample followed a normal distribution. We can see that the tail that represents heavier individuals both data sets taper off being normal. For this reason, I chose to use a t-test, because it is more robust. The datasets are large enough where it could be estimated to a normal distribution though.


Q-Q plot male

Q-Q plot female

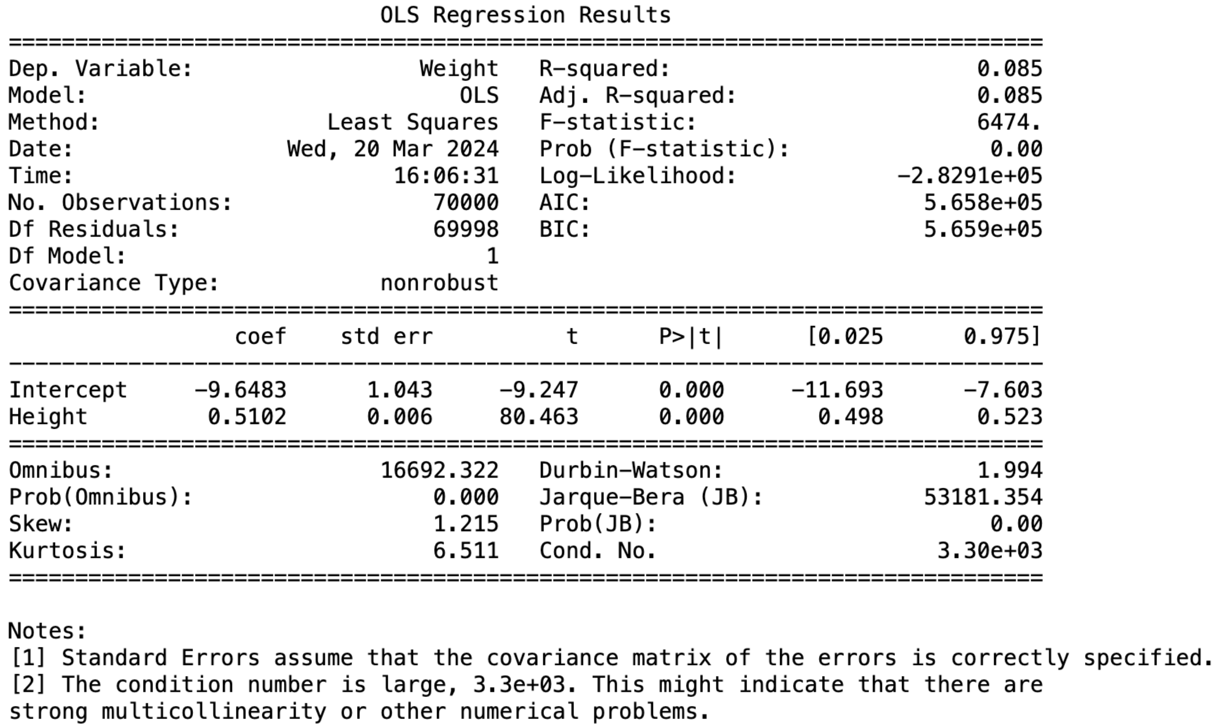
**Part 2: Correlation Coefficient**

To visualize the difference in the two groups I created a plot that shows the height and weight of the two separate groups. In this we can see that in this dataset the woman has a much smaller spread then males, and they do tend to be on the taller side in this sample.



Using the pearsonr test I was able to determine that the correlation coefficient is .29 and is statistically significant with a p-value of basically 0. By looking back at the plot above we can see that the data does not appear to be very linear, as most of the data is sitting in the middle of the plot. This would explain the low correlation.

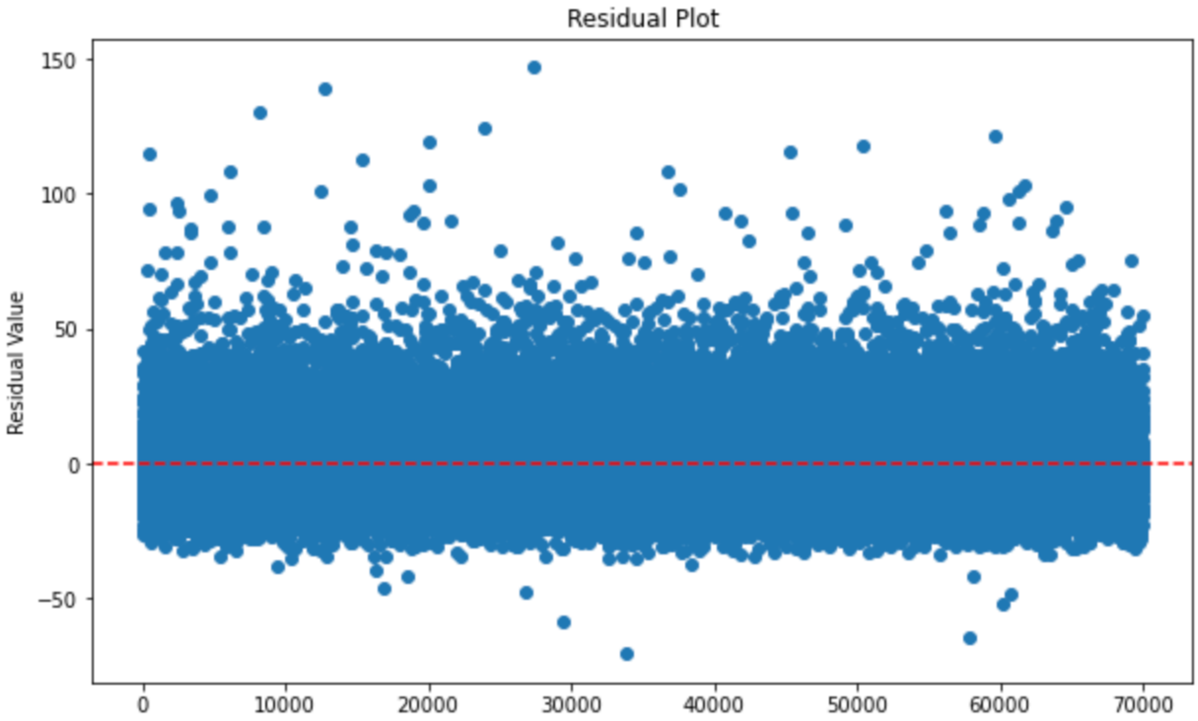
## Part 3: Linear Regression



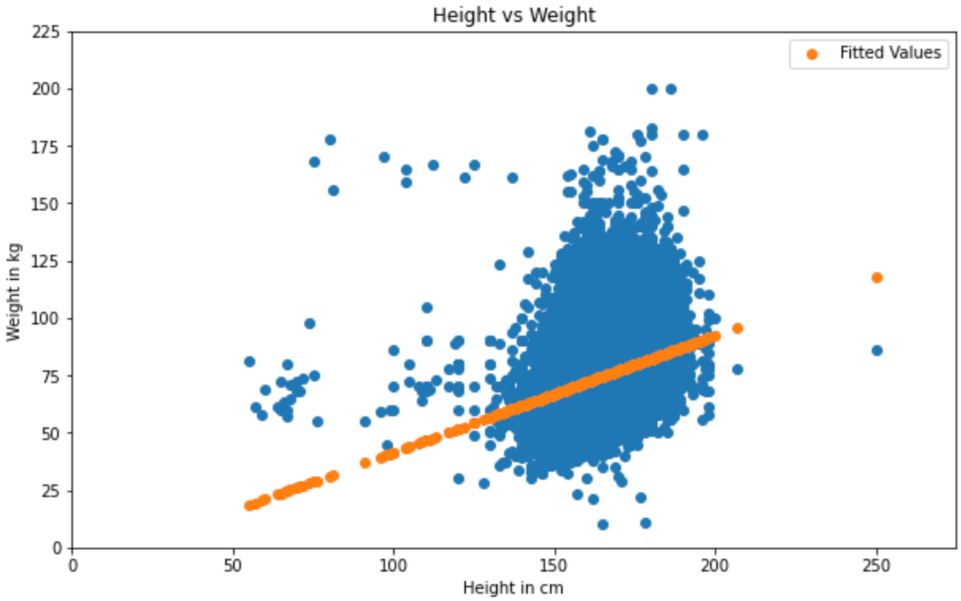
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Weight   R-squared:                       0.085
Model:                            OLS   Adj. R-squared:                  0.085
Method:                 Least Squares   F-statistic:                     6474.
Date:                Wed, 20 Mar 2024   Prob (F-statistic):               0.00
Time:                        16:06:31   Log-Likelihood:            -2.8291e+05
No. Observations:               70000   AIC:                         5.658e+05
Df Residuals:                   69998   BIC:                         5.659e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -9.6483      1.043     -9.247      0.000     -11.693      -7.603
Height         0.5102      0.006     80.463      0.000       0.498       0.523
==============================================================================
Omnibus:                    16692.322   Durbin-Watson:                   1.994
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            53181.354
Skew:                           1.215   Prob(JB):                         0.00
Kurtosis:                       6.511   Cond. No.                     3.30e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.3e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

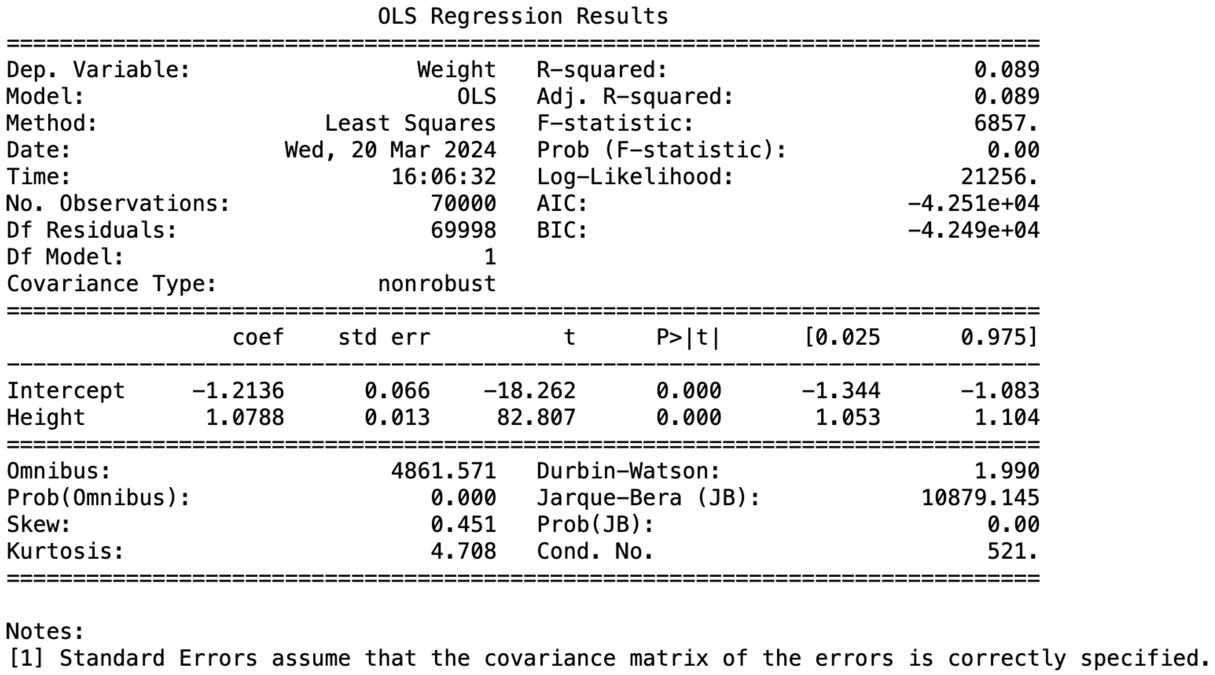
In creating a model based on the data to predict weight based on height it was found that the model was statistically significant with a F-statistic of 6474, leading to a near 0 p-value. The model is not very good though as the R-squared value is very low, being .085. Once again statistical significance is most likely because of the sheer size of the dataset. There is a flag for multicollinearity, but this is simple regression, so there is nothing to be colinear with. This is most likely do to the fact that most of the data is centered around one height, so the variance of xi is too small. I wouldn't say that this data fits the linearity assumption to even create a linear regression model, so this model is not going to perform well based off this data.



The residual plot does show homoscedasticity of the residuals, as they are randomly distributed. There just happens to be a very large spread of residuals because the model is not very accurate.

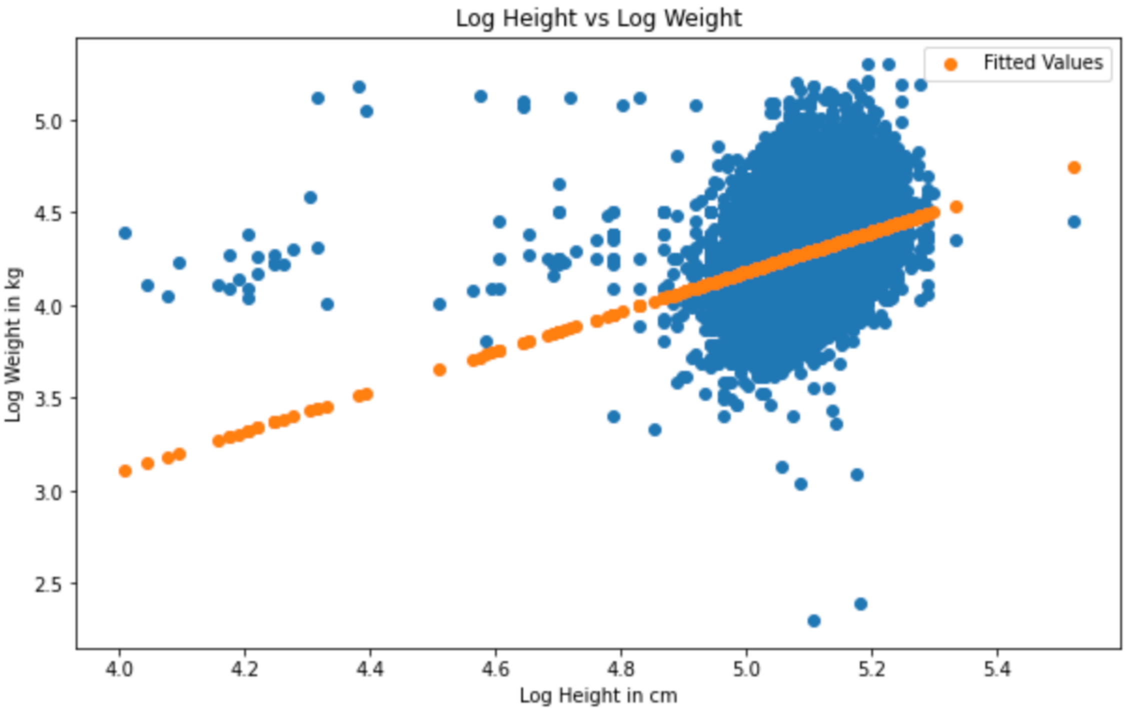


We can further see that the model is not accurate by looking at the dataset versus the fitted values, and we can see that it does not perform well.



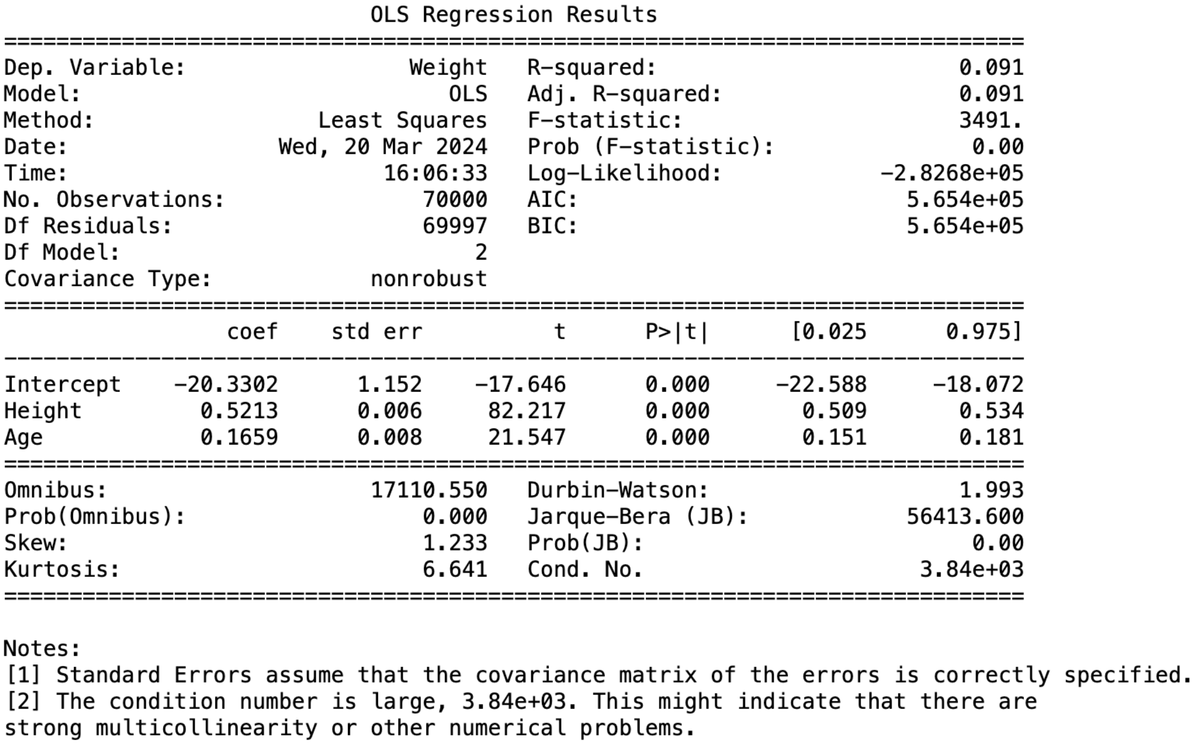
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Weight   R-squared:                       0.089
Model:                            OLS   Adj. R-squared:                  0.089
Method:                 Least Squares   F-statistic:                     6857.
Date:                Wed, 20 Mar 2024   Prob (F-statistic):               0.00
Time:                        16:06:32   Log-Likelihood:                 21256.
No. Observations:               70000   AIC:                        -4.251e+04
Df Residuals:                   69998   BIC:                        -4.249e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -1.2136      0.066    -18.262      0.000      -1.344      -1.083
Height         1.0788      0.013     82.807      0.000       1.053       1.104
==============================================================================
Omnibus:                     4861.571   Durbin-Watson:                   1.990
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            10879.145
Skew:                           0.451   Prob(JB):                         0.00
Kurtosis:                       4.708   Cond. No.                         521.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```



To make the model perform better I took the logarithm of all the numerical data to tighten the spread to make it more linear. This did very little improve the model though. From the results above we can see that the spread of the main cluster is too larger to estimate effectively.

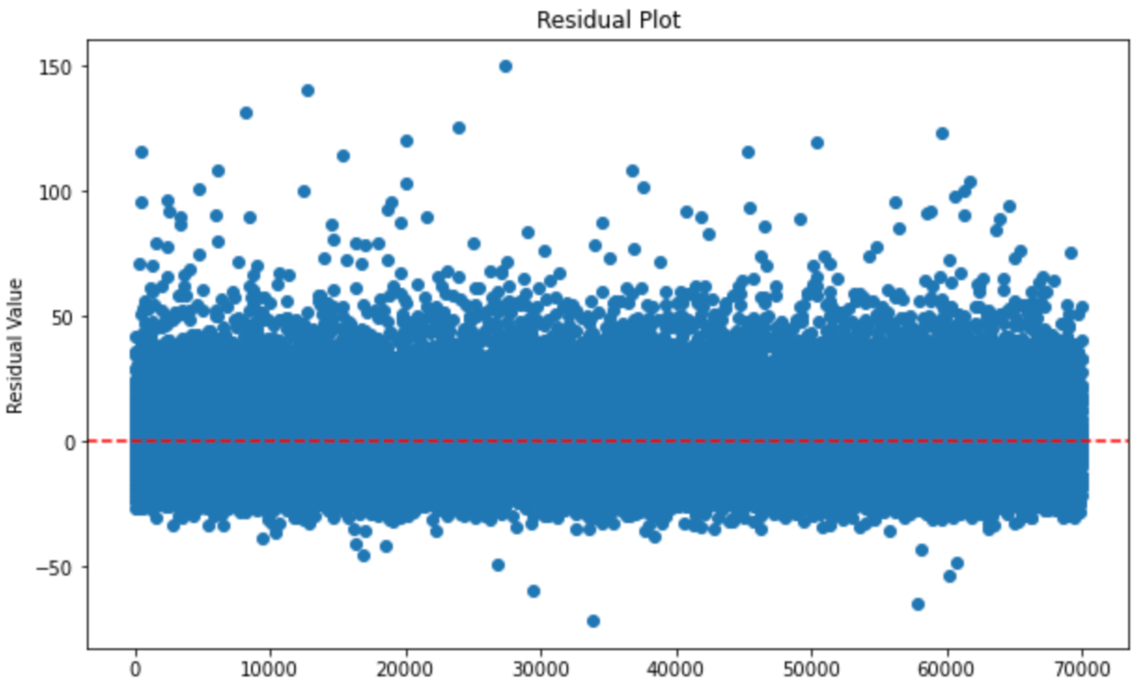
## Part 4 Multiple Regression



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Weight   R-squared:                       0.091
Model:                            OLS   Adj. R-squared:                  0.091
Method:                 Least Squares   F-statistic:                     3491.
Date:                Wed, 20 Mar 2024   Prob (F-statistic):               0.00
Time:                        16:06:33   Log-Likelihood:            -2.8268e+05
No. Observations:               70000   AIC:                         5.654e+05
Df Residuals:                   69997   BIC:                         5.654e+05
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -20.3302      1.152    -17.646      0.000     -22.588     -18.072
Height         0.5213      0.006     82.217      0.000       0.509       0.534
Age            0.1659      0.008     21.547      0.000       0.151       0.181
==============================================================================
Omnibus:                    17110.550   Durbin-Watson:                   1.993
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            56413.600
Skew:                           1.233   Prob(JB):                         0.00
Kurtosis:                       6.641   Cond. No.                     3.84e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.84e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

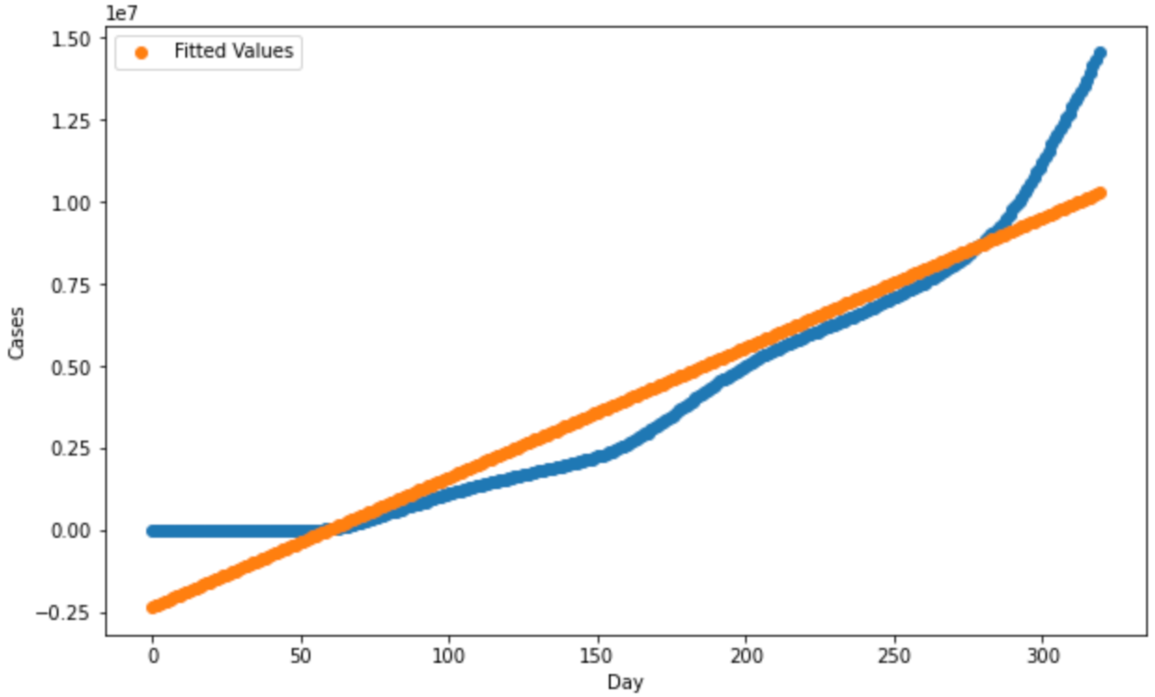
The multiple regression model suffers from many of the same detriments that the simple model does. The data is not described well by the model with a R-squared of .091, slightly better than the simple regression model, and it still suffers from being overfit by having too many data points.



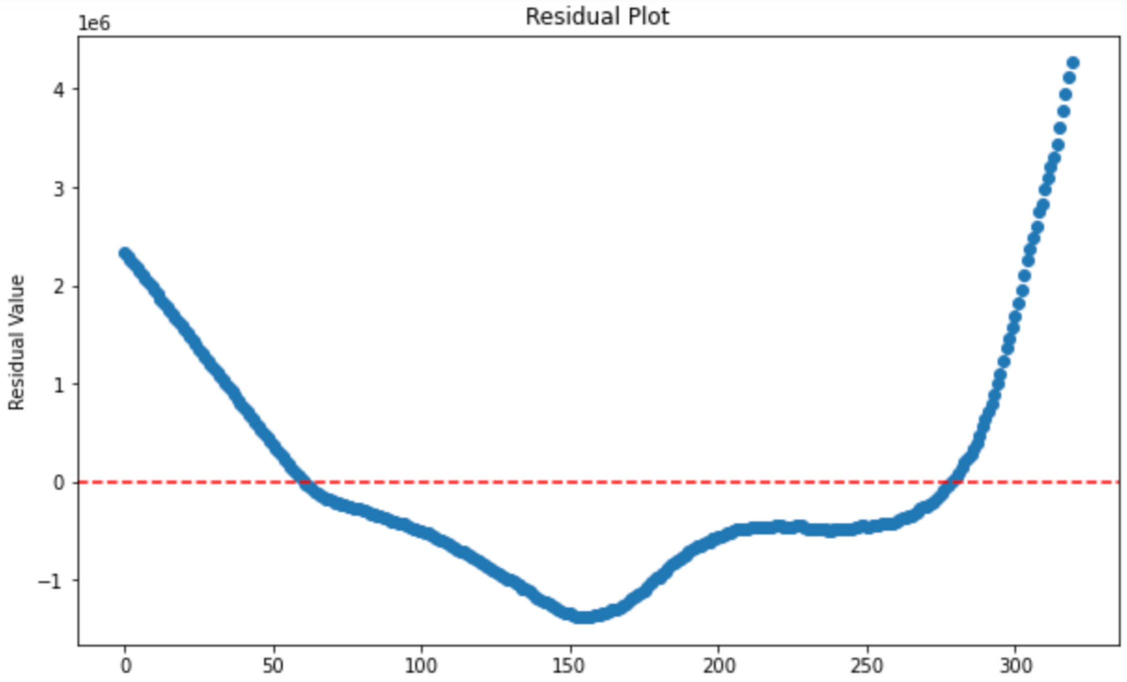
Looking at the residual plot there is very little difference from the simple model, so the heteroscedasticity condition is still met. There is a flag of multicollinearity though. To analyze this, I conduct a VIF (Variance Inflation Factor) test to analyze if there is a large amount of collinearity. It is found that there is a large amount of collinearity. This would make sense as the taller you are the heavier you are expected to be and you grow as you get older, so older people will be taller in comparison with younger people. The way to combat this would be to conduct factor reduction, but we are limited to only a couple factors so this would not do much good. As well as we have already identified that the linearity condition is not met.

**Part 5: Solutions**

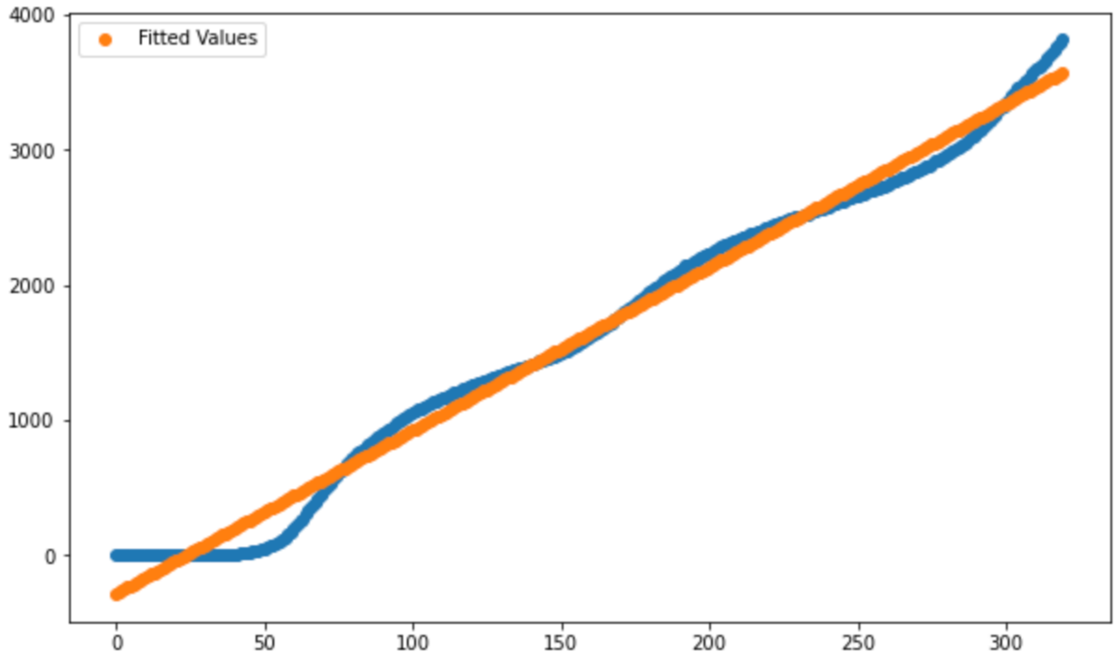
For the second part of this assignment, I took a dataset of COVID data in the United States to analyze the number of cases, as it grew exponentially. This highlighted the two main assumptions of regression linearity of the data, and homoscedasticity of the residuals.



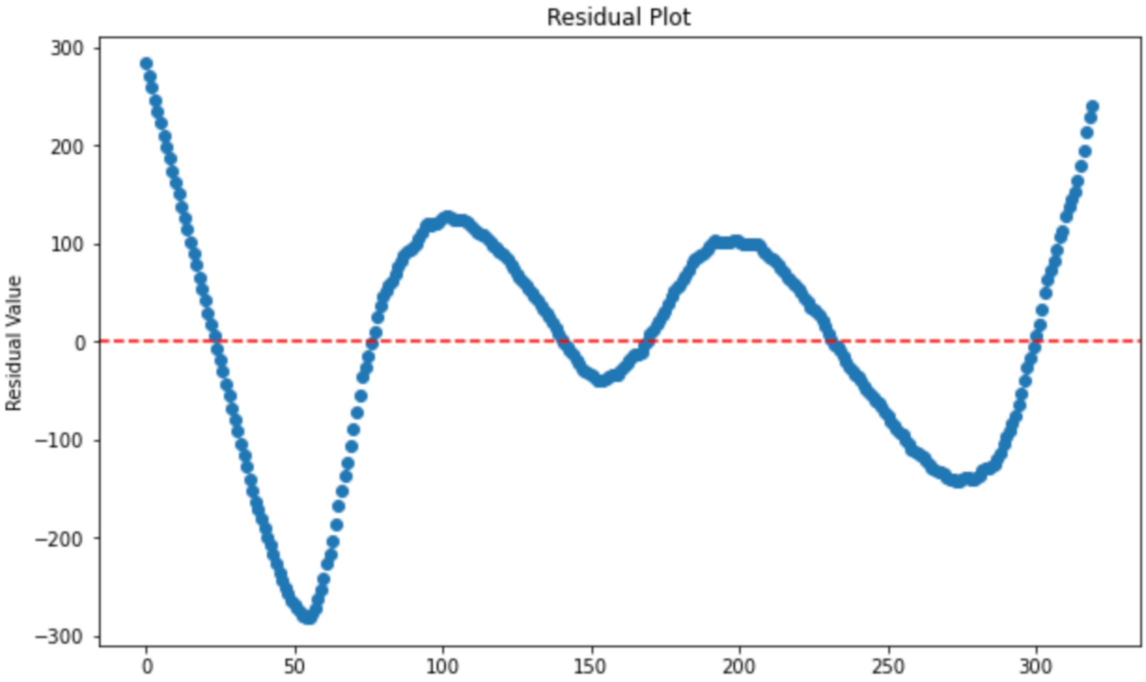
If we look at the blue line, the original data, we can see that the relationship between the number of cases and the number of days into the pandemic is exponential. This would break the linearity assumption.



Then looking at the residual plot we can see that the data follows a very distinct pattern compared to the fitted values. Violating the other condition of linear regression. To combat this, I took the square root of the data to flatten the curve.

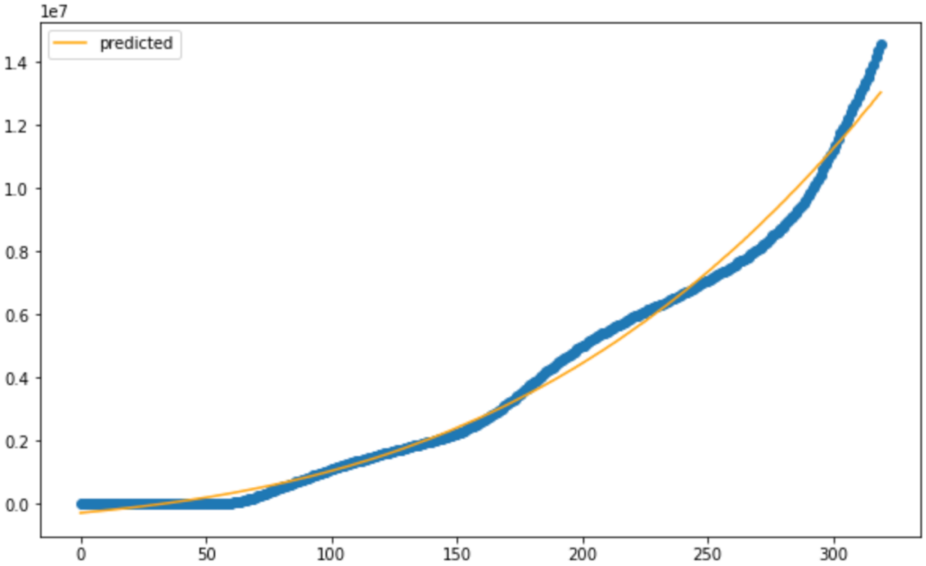


This created a much more accurate model in the middle of the data, but still struggled at the tails. This is because the tails do not follow a linear pattern. This data is also much more linear.

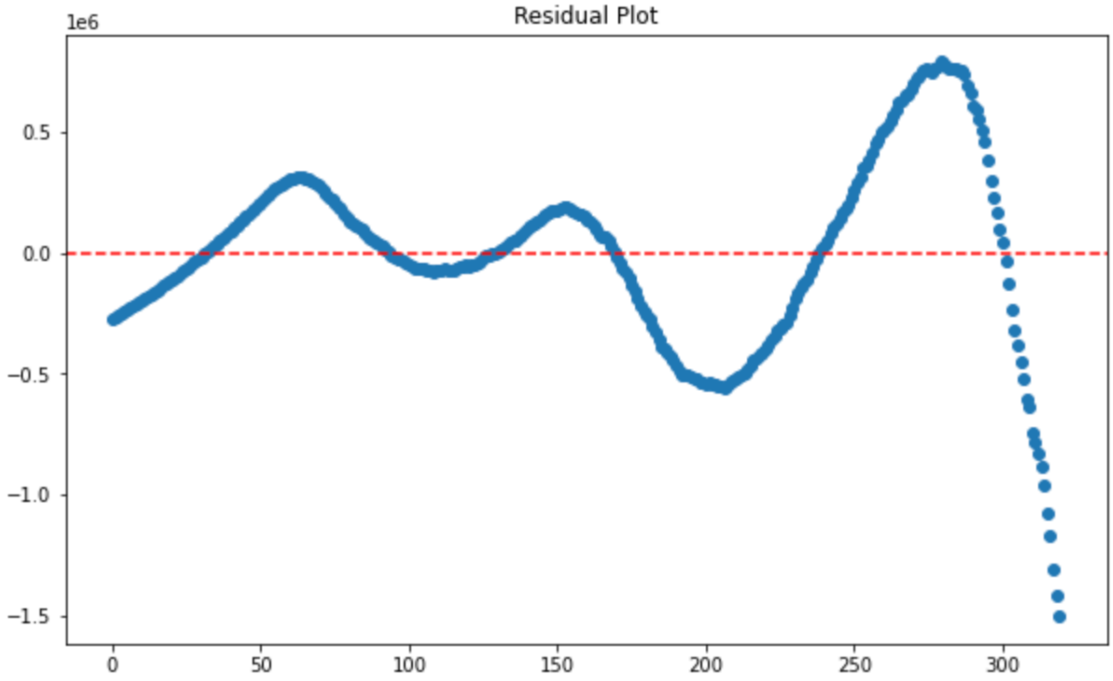


When comparing the residual plots of the original data and this data we can see that the model has become much better at predicting data in the middle of the dataset, but still holds a distinct pattern in the residuals, so the homoscedasticity check is violated.

**Part 6: Nonlinear Model**



To create a more accurate model I use polynomial regression since the data follows an exponential pattern. This allows me to be more accurate at the tails of the data, and still maintain the accuracy in the middle. We can see that it does begin to struggle towards the end of the dataset, so it will not be very reliable for future predictions. We can fix this by not overfitting the model by giving it too much data to train with.



The residual plot shows a similar story, as it is very good at the low to mid values, and then struggles towards the end. Since the data is so tightly spread, the residuals are going to appear very patterned. This is once again because the model is overfit. If we used a sample of points, it would show more of a random spread.

# References

Cardiovascular Disease dataset. (2019). Kaggle [Dataset]. https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset/data.

COVID-19 in USA. (2021). Kaggle [Dataset]. https://www.kaggle.com/datasets/sudalairajkumar/covid19-in-usa?select=us_covid19_daily.csv.

Rogel-Salazar, J. (2023). Statistics and Data Visualization with Python. CRC Press.