Week Eleven: Linear Regression

CS 217

Correlation

- Say we want to measure the price of stamps over time and see if there is a relationship between the year, and the price of a stamp
- Specifically we want to see the relationship between the **number of years since 1960** and the price of a stamp

Correlation

- Say we want to measure the price of stamps over time and see if there is a relationship between the year, and the price of a stamp
- Specifically we want to see the relationship between the **number of years since 1960** and the price of a stamp
- Using what we learned last week, we can find the **covariance** and **correlation** of the relationship

Years since 1960	Price of Stamp
3	0.05
8	0.06
11	0.08
14	0.10

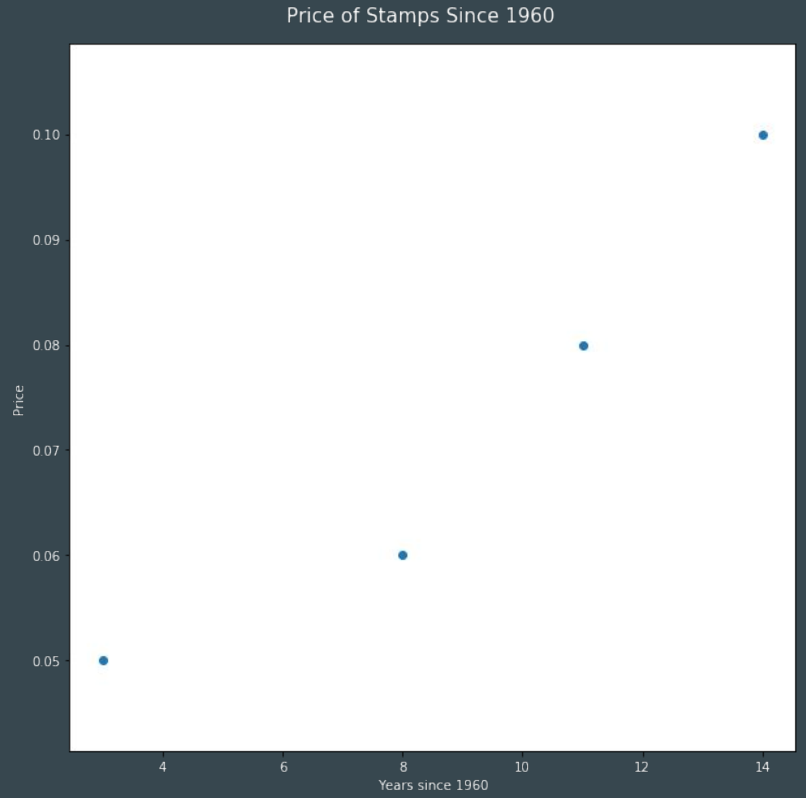
Correlation

- The covariance is 0.3 / 4, or 0.075

Years since 1960	Price of Stamp	Years - E(Years)	Price - E(Price)	Year Diff * Price Diff
3	0.05	-6	-0.0225	0.135
8	0.06	-1	-0.0125	0.0125
11	0.08	2	0.0075	0.015
14	0.10	5	0.0275	0.1375
Expected Value: 9	**Expected Value: 0.0725**			Sum: 0.3

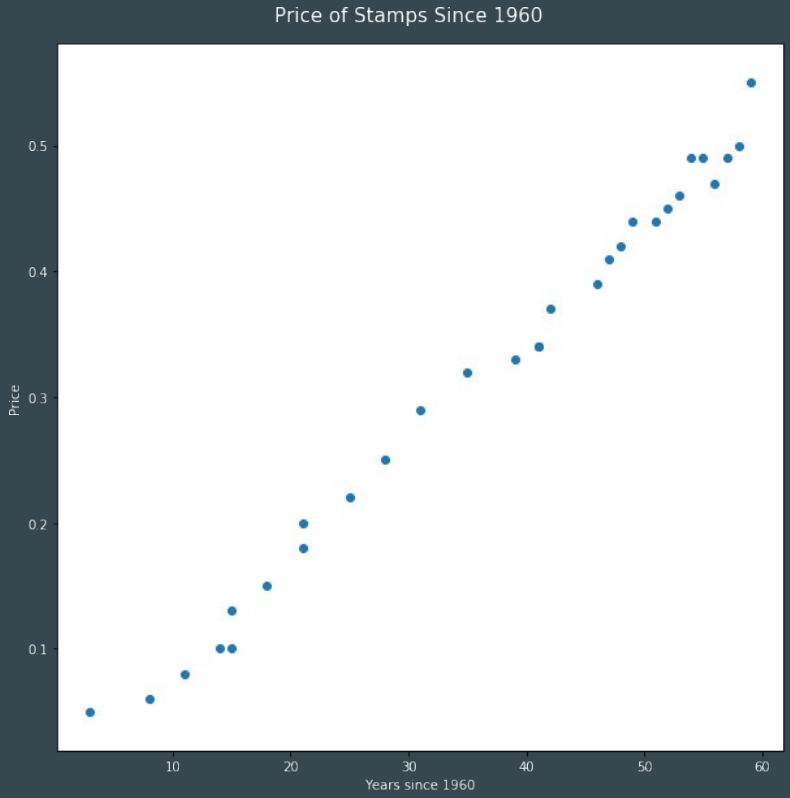
Scatterplots

- We can also create a scatterplot to visualize the relationship between the two variables



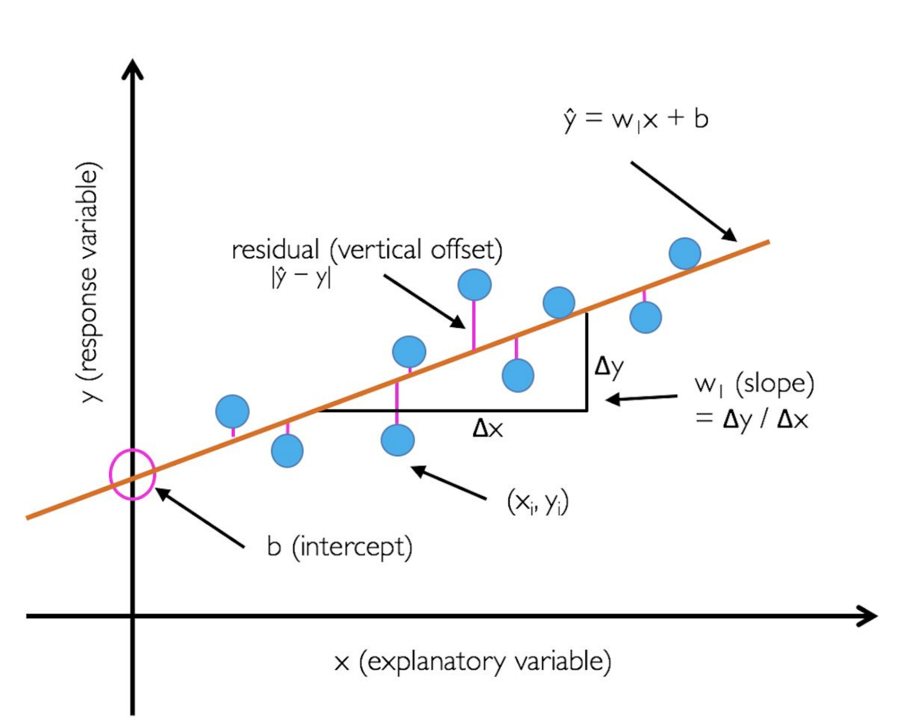
Scatterplots

- We can also create a scatterplot to visualize the relationship between the two variables
- Clearly there’s a positive linear relationship here, but how do we determine what the actual slope of the line is?



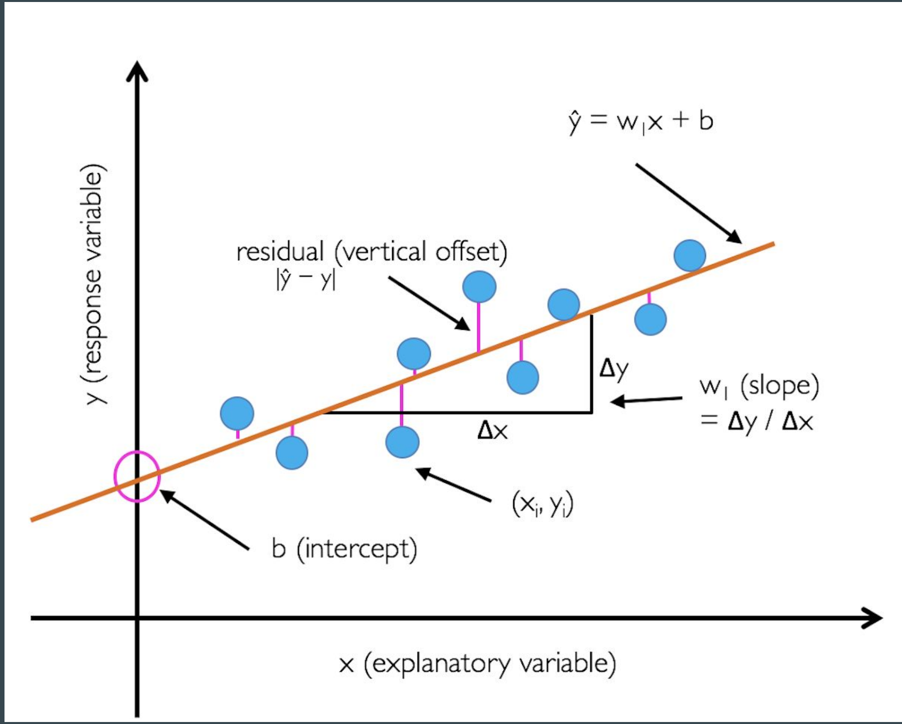
Slope

- The equation y = mx + b can approximate the linear relationship between two variables
- It insinuates that a given y value is equal to a given x value multiplied by m (the slope) + the intercept (b)
- The slope answers the question - ‘if I change X by one unit, how much does Y change?’



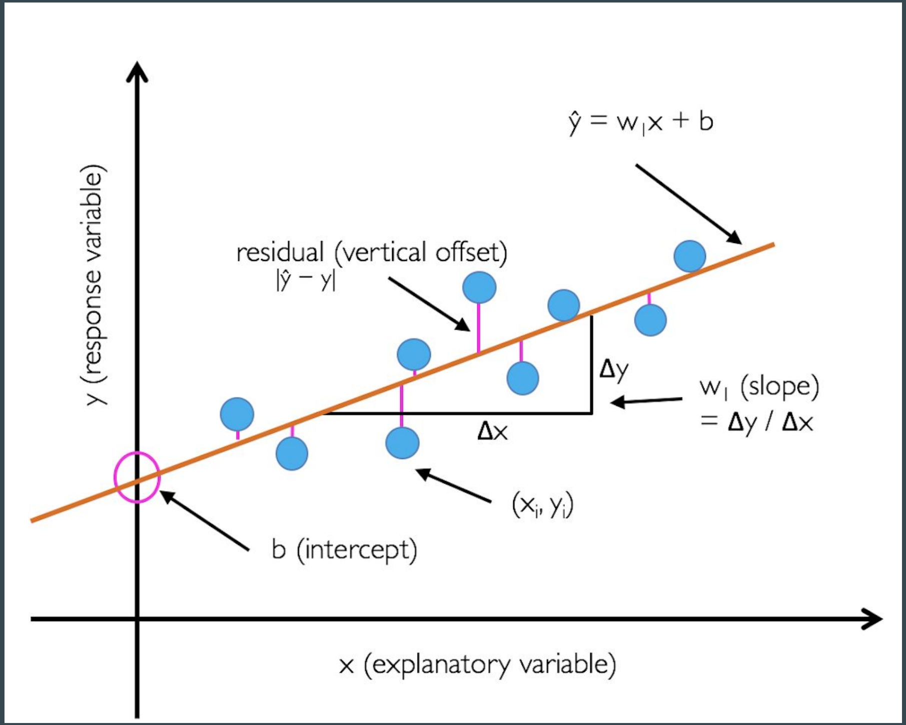
Least Squares Fit

- For a least squares fit, the **slope** is equal to the covariance of X and Y over the variance of X
- The **intercept** is equal to the mean of Y minus the **slope** times the **mean of x**



Least Squares Fit

- The vertical deviation between a given data point and the line approximating the linear relationship is called a **residual**
- We want to minimize the residual values so that we ensure our line is as accurate as possible in mapping the relationship between our two variables
- Specifically we want to minimize the sum of squared residual values. This is called a **linear least squares fit**



Least Squares Fit

- **Covariance**: 0.075
- **Mean of X**: 9
- **Mean of Y**: 0.0725
- **Variance of X**: 16.5

Given these variables, what are the **slope** and **intercept** of the linear least squares fit?

Years since 1960	Price of Stamp
3	0.05
8	0.06
11	0.08
14	0.10

Least Squares Fit

- **Covariance**: 0.075
- **Mean of X**: 9
- **Mean of Y**: 0.0725
- **Variance of X**: 16.5

Given these variables, what are the **slope** and **intercept** of the linear least squares fit?

Slope: 0.00454

Intercept: 0.03159

Y = 0.00454 * X + 0.03159

Years since 1960	Price of Stamp
3	0.05
8	0.06
11	0.08
14	0.10

Least Squares Fit

- **Covariance**: 0.075
- **Mean of X**: 9
- **Mean of Y**: 0.0725
- **Variance of X**: 16.5

Given these variables, what are the **slope** and **intercept** of the linear least squares fit?

Slope: 0.00454

Intercept: 0.03159

Y = 0.00454 * X + 0.03159

Years since 1960	Price of Stamp	Predicted Price
3	0.05	0.0452
8	0.06	0.0679
11	0.08	0.0816
14	0.10	0.0952

Least Squares Fit

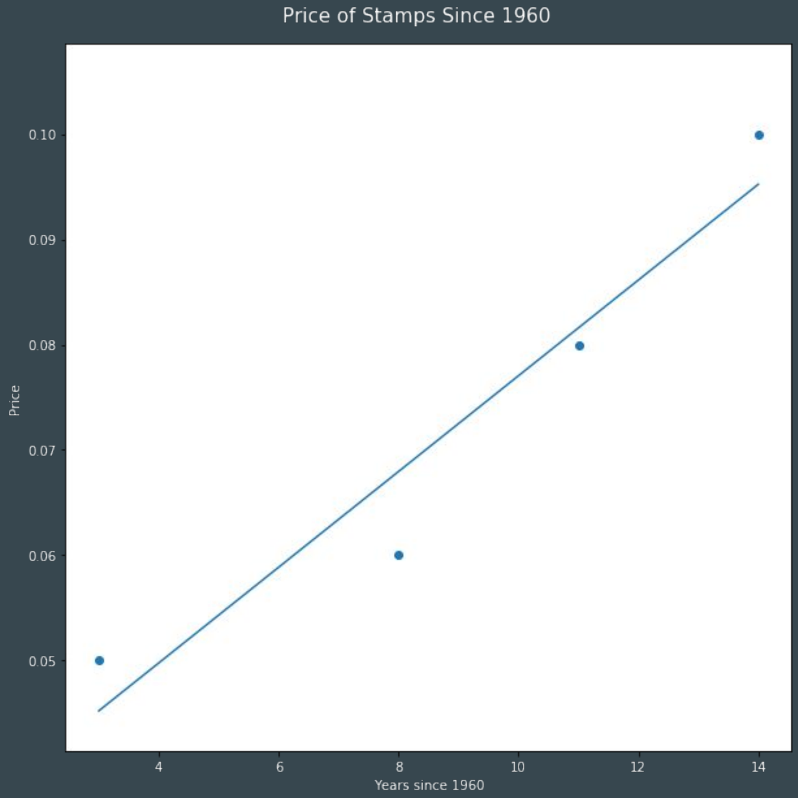
- **Covariance**: 0.075
- **Mean of X**: 9
- **Mean of Y**: 0.0725
- **Variance of X**: 16.5

Given these variables, what are the **slope** and **intercept** of the linear least squares fit?

Slope: 0.00454

Intercept: 0.03159

Y = 0.00454 * X + 0.03159



Least Squares Fit

- Now that we have a least squares regression, we can use it to predict **new values** for our data that we didn't previously have.
- Y = 0.00454 * X + 0.3159
- Y = 0.00454 * 18 + 0.3159
- 0.1134 = 0.00454 * 18 + 0.3159

Years since 1960	Price of Stamp	Predicted Value
3	0.05	0.0452
8	0.06	0.0679
11	0.08	0.0816
14	0.10	0.0952
18		0.1134

Least Squares Fit

- Now that we have a least squares regression, we can use it to predict **new values** for our data that we didn’t previously have.
- Y = 0.00454 * X + 0.3159
- Y = 0.00454 * 18 + 0.3159
- 0.1134 = 0.00454 * 18 + 0.3159
- Obviously our model will be better with more than four data points.

Years since 1960	Price of Stamp	Predicted Value
3	0.05	0.0452
8	0.06	0.0679
11	0.08	0.0816
14	0.10	0.0952
18	0.15	0.1134

Least Squares Fit

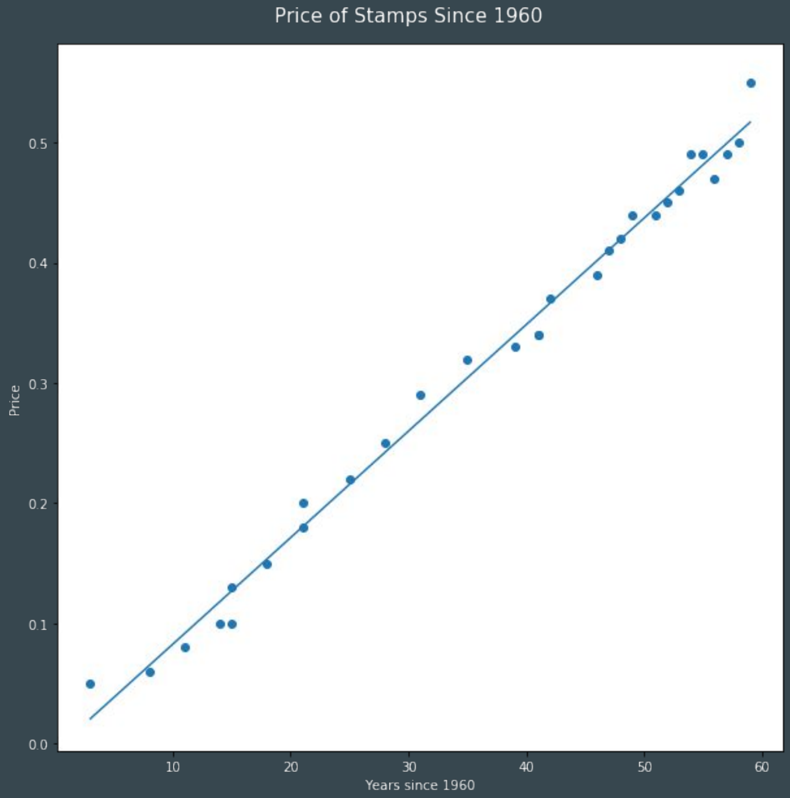
- **Covariance**: 2.682
- **Mean of X**: 36.43
- **Mean of Y**: 0.317
- **Variance of X**: 302.80

Given these variables, what are the **slope** and **intercept** of the linear least squares fit?

Slope: 0.00886

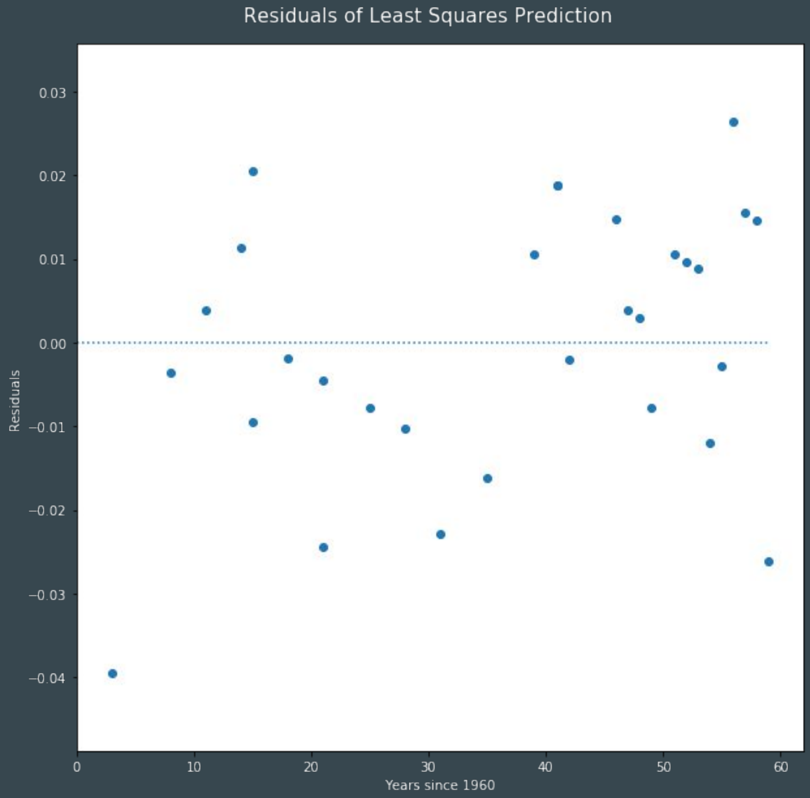
Intercept: -0.00578

Y = 0.00886 * X - 0.00578



Residuals

- The **residuals** are the difference between the **predicted** value and the actual value
- Least squares fit is a regression model that minimizes the **squared residuals**

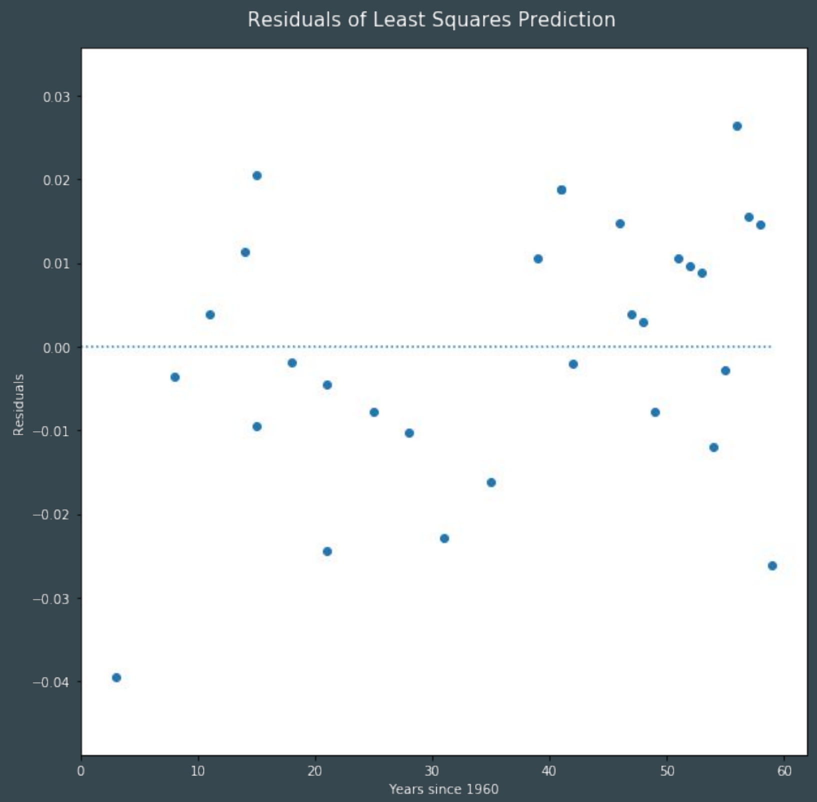


Least Squares Fit

Years since 1960	Price of Stamp	Predicted Value	Residual	Squared Residual
3	0.05	0.0452	-0.004	0.0000227
8	0.06	0.0679	0.008	0.0000632
11	0.08	0.0816	0.001	0.0000025
14	0.10	0.0952	-0.004	0.0000227

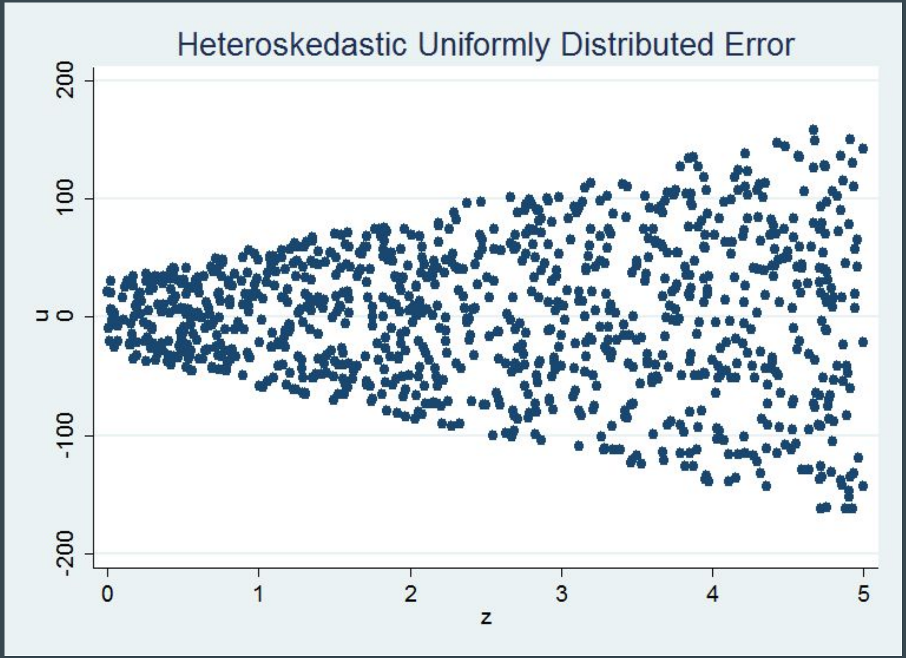
Residuals

- There are a few rules the residuals should follow in order for the assumptions of a linear regression to hold.
- The residuals should look like they do on the right, randomly scattered amongst the graph, with positive and negative values
- This is called homoskedasticity.



Residuals

- The opposite, where there is a pattern in the residuals, such as what we see on the right, is called **hetereoskedasticy**
- If the residuals are hetereoskedastic, the relationship between the variables isn't linear!!



Goodness of Fit

- R-Squared is a measure of how close the data are to your fitted regression line
- Specifically it measures the “explained variation” over “total variation”, and is on a range between 0 and 100%
- It is 1 - the **variance of the residuals** divided by the **variance of the dependent variable**
- (It is also the squared value of the correlation coefficient)

$$r^2 = 1 - \frac{\Sigma(y - y')^2}{\Sigma(y - \bar{y})^2}$$

Least Squares Fit

- Find the R-Squared value of our predictions from earlier.

Years since 1960	Price of Stamp	Predicted Value	Residual
3	0.05	0.0452	-0.004
8	0.06	0.0679	0.008
11	0.08	0.0816	0.001
14	0.10	0.0952	-0.004

Least Squares Fit

- Find the R-Squared value of our predictions from earlier.

Variance(Price of Stamp) = 0.000369

Variance(Residuals) = 0.0000241

1 - (0.0000241/0.000369) = 0.9344

R-Squared = 0.9344 or 93.44%

93.44% of the actual variance in the price of stamps is ‘accounted for’ by our prediction

Years since 1960	Price of Stamp	Predicted Value	Residual
3	0.05	0.0452	-0.004
8	0.06	0.0679	0.008
11	0.08	0.0816	0.001
14	0.10	0.0952	-0.004

Least Squares Fit

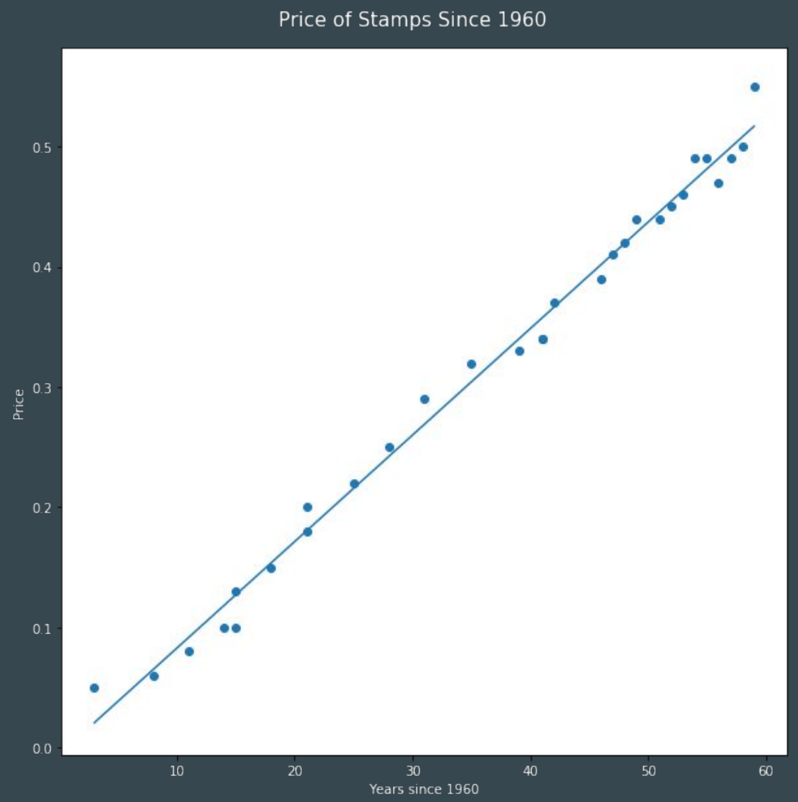
- Find the R-Squared value of our predictions from earlier.

Variance(Price of Stamp) = 0.02398

Variance(Residuals) = 0.000219

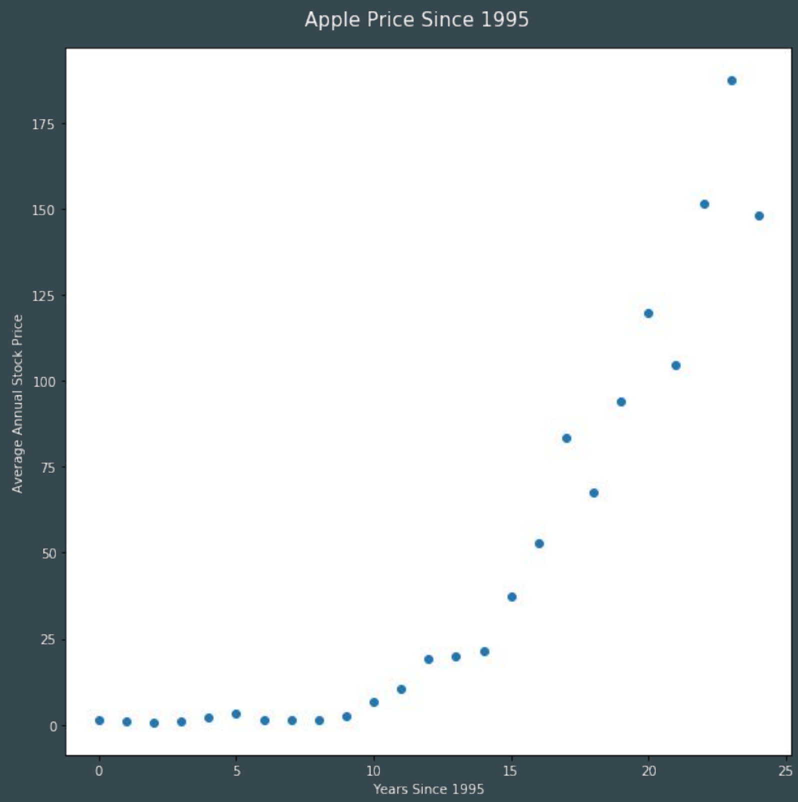
R-Squared = 0.9908, or 99.08%

99.08% of the actual variance in the price of stamps is 'accounted for' by our prediction



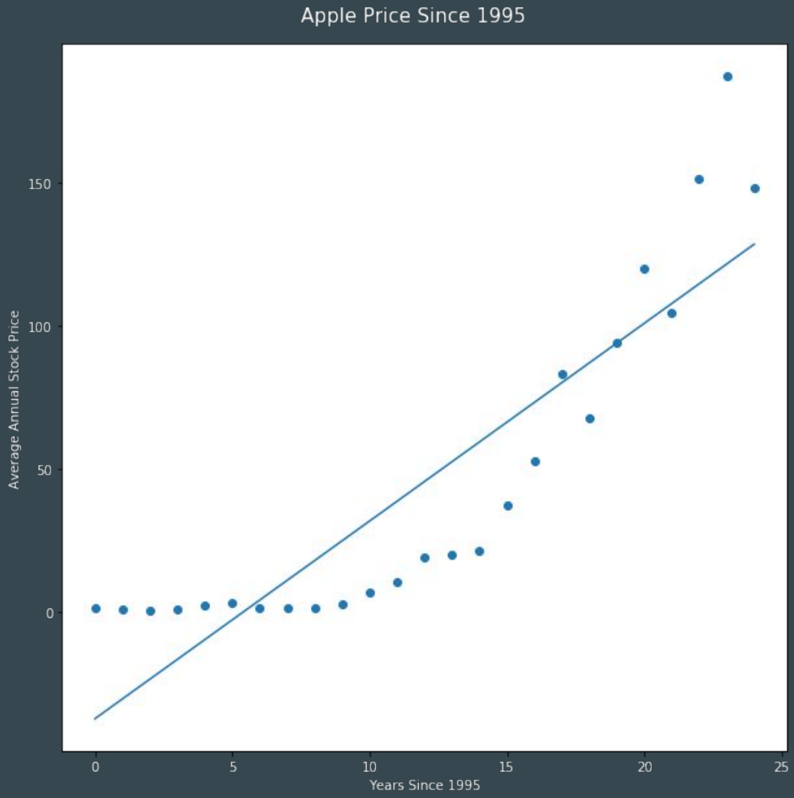
Non-Linear Regression

- Let's look at the stock price of Apple since 1990 for an example of a non-linear relationship.



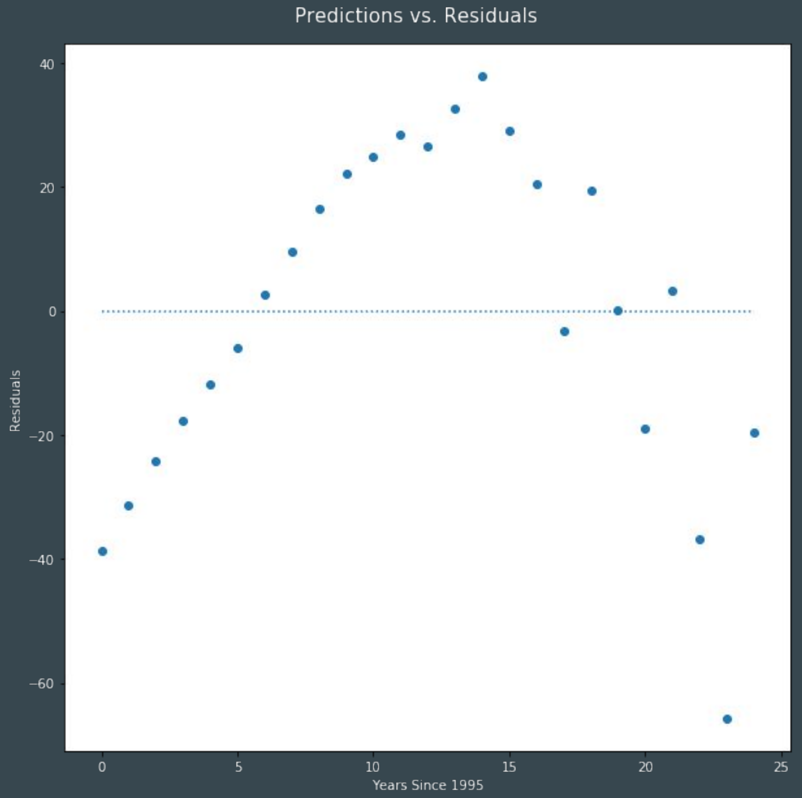
Non-Linear Regression

- Let's look at the stock price of Apple since 1995 for an example of a non-linear relationship.
- While a linear fit can capture the general trend of the data, it leaves a lot to be desired.



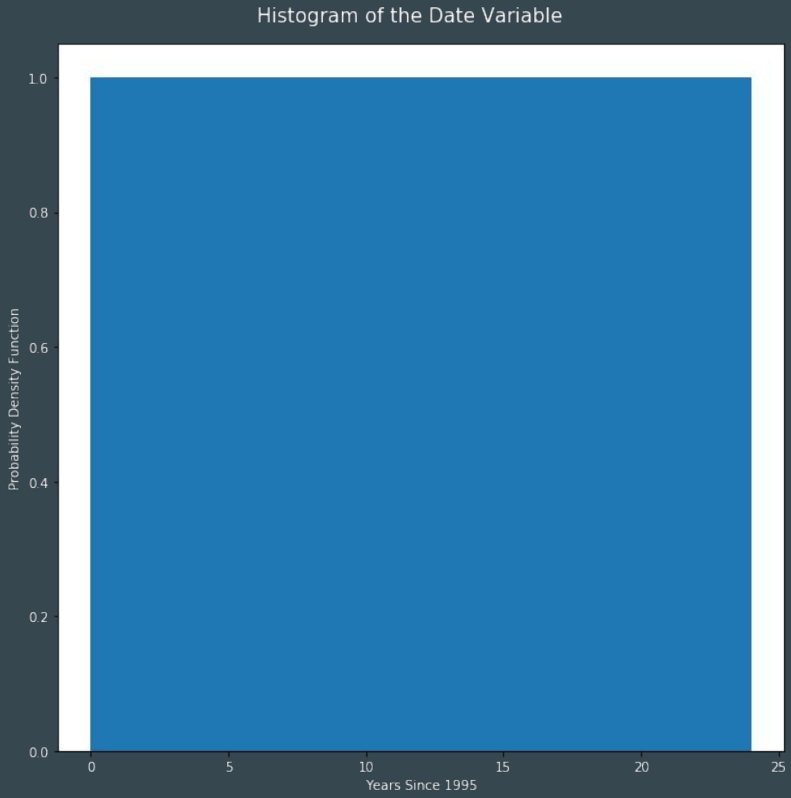
Non-Linear Regression

- Additionally the residuals graph clearly has a pattern, suggesting that the assumptions for linear regression do not hold.
- The R-squared value for this relationship, however, is still 0.78, which is relatively high.
- R-squared on it's own isn't enough to evaluate the strength of a linear model - you must confirm that the linear assumptions hold!



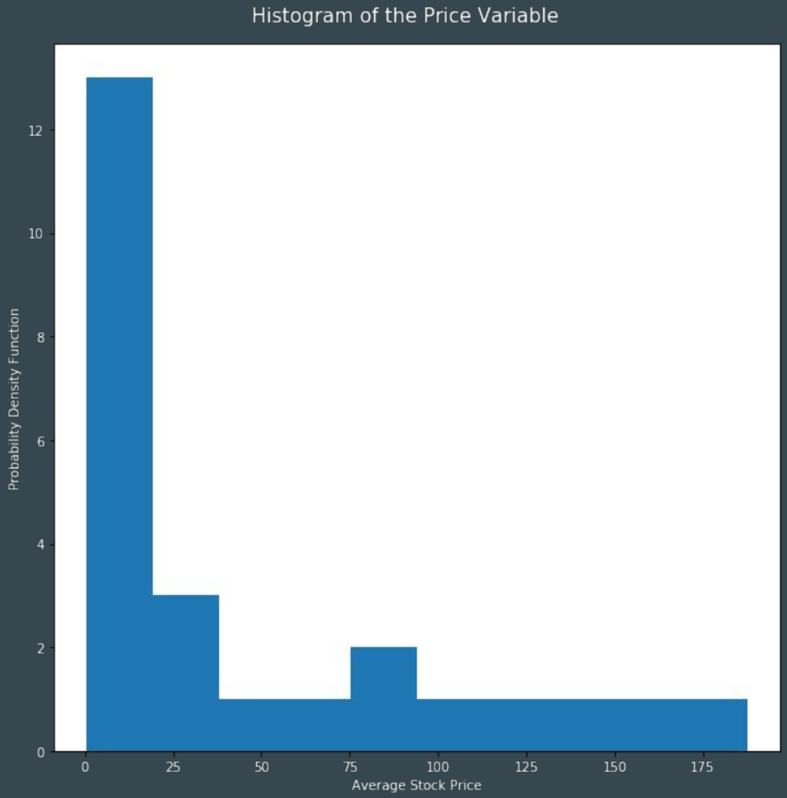
Transformation

- Like we did last week, we can see which of our variables we can transform so that we can observe a linear relationship between our variables.
- The date variable is uniform, which makes sense since we're taking one data point for every year since 1995



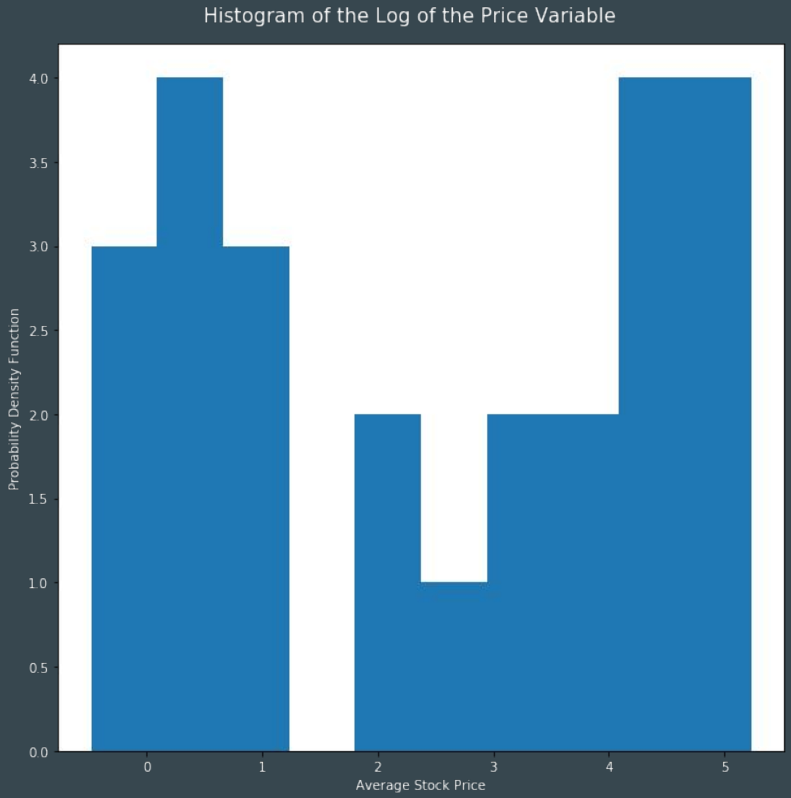
Transformation

- The price variable, on the other hand, is extremely positively skewed, as most of its values are extremely low.
- What is a good way to transform this variable?



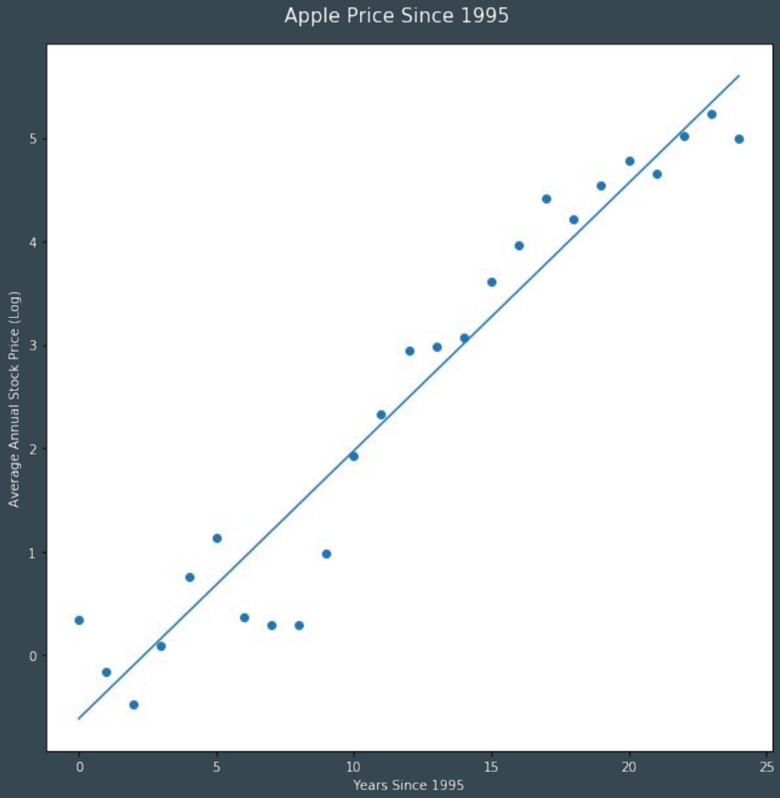
Transformation

- The price variable, on the other hand, is extremely positively skewed, as most of its values are extremely low.
- What is a good way to transform this variable?
- We can take the log of the price variable to remove the skew



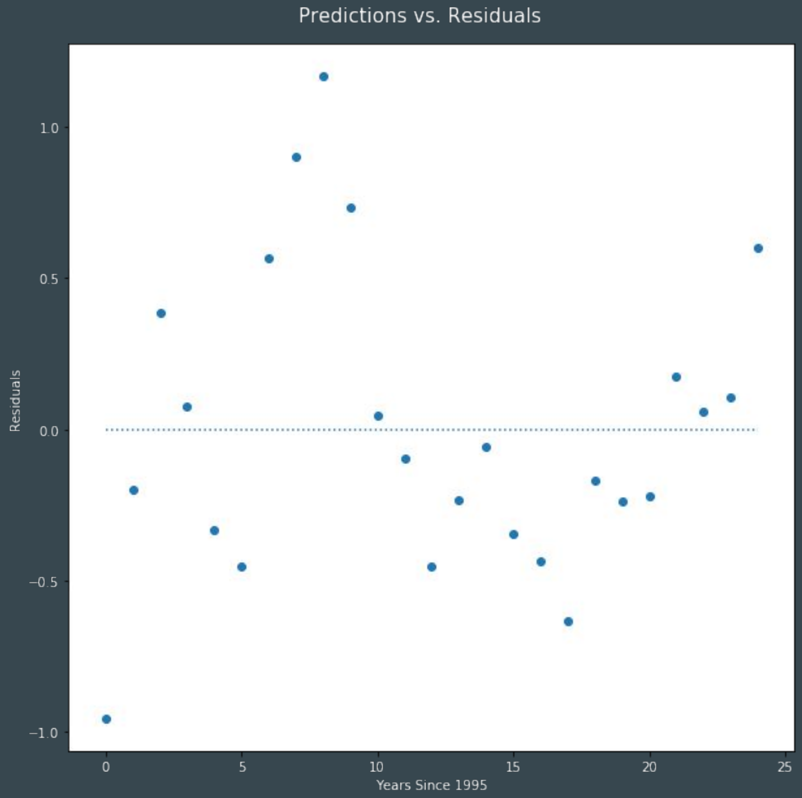
Non-Linear Regression

- Now that we've taken the log value of the response value, a linear relationship is much more appropriate.



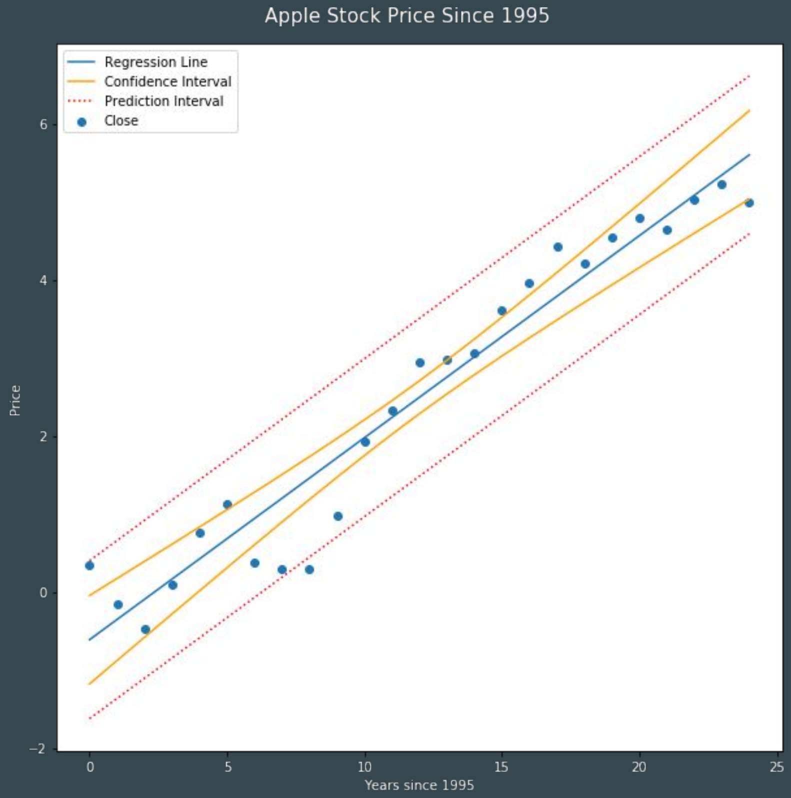
Non-Linear Regression

- The residuals graph now demonstrates heteroskedasticity, confirming that a linear regression is appropriate.



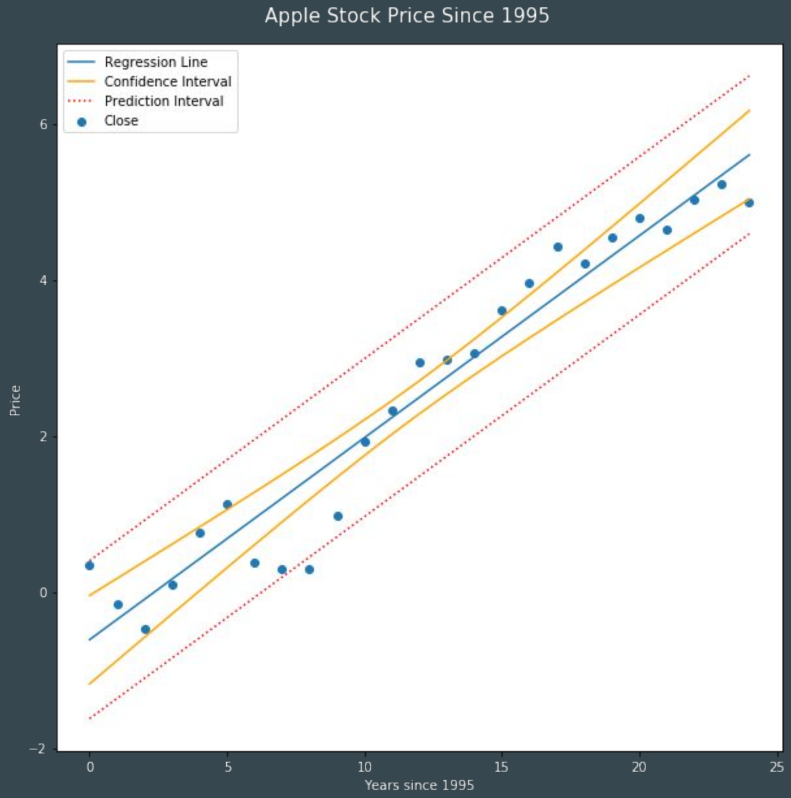
Non-Linear Regression

- We can also look at the **confidence intervals** and **prediction intervals** of the regression line
- To the right the orange lines are the **upper and lower confidence intervals** for the regression line - the **mean value** of Y for a given value of X should fall within these lines
- While the red dotted lines are the **upper and lower prediction intervals** - an **observed value** of Y for a given value of X should fall within these lines



Non-Linear Regression

- The scope of the calculation of these intervals is beyond the scope of this class, but they are a good visual tool for the power of the regression line as predictor
- A regression line is just an **estimation** - how good does it capture the relationship between two variables and allow you to make predictions on new values?
- Do you trust linear regression as a predictor given the assumptions it's making?



Non-Linear Regression

- This is the model before transformation of the Y variable - do you trust the assumptions the model makes here?