NYC DATA SCIENCE ACADEMY

Simple Linear Regression

Data Science Bootcamp

Outline

- Part 1: Simple Linear Regression
- Part 2: Diagnostics
- Part 3: Transformations
- Part 4: The Coefficient of Determination R^2
- Part 5: Review

PART 1

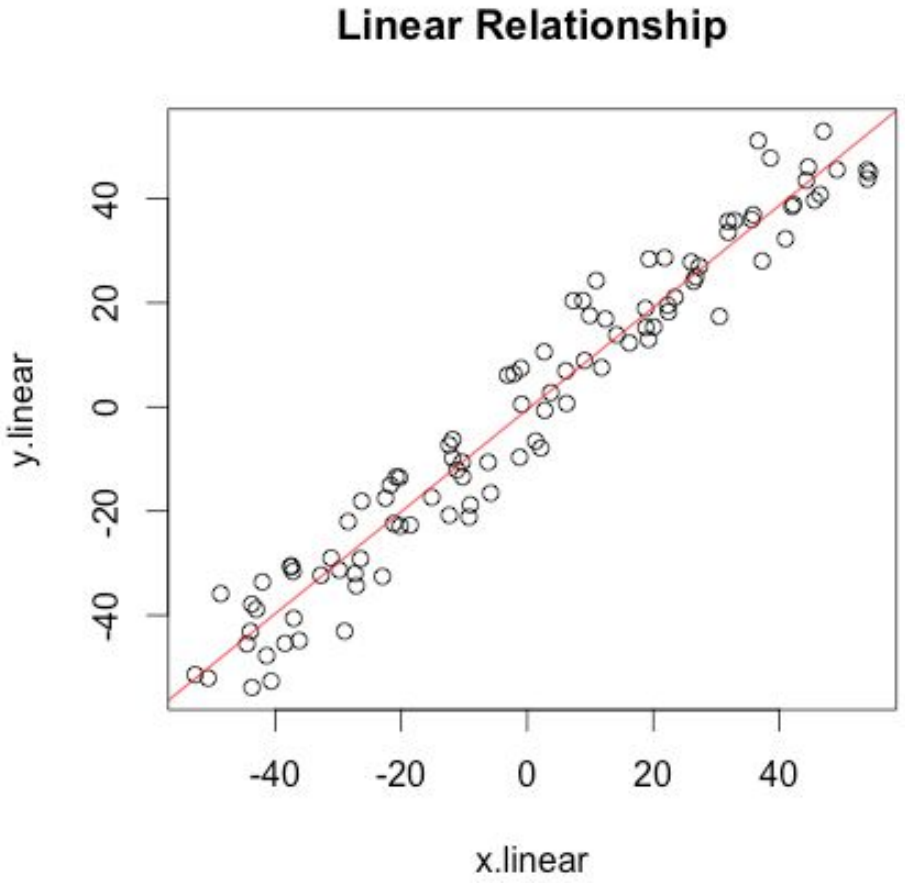
Simple Linear Regression

What is Simple Linear Regression?

- Simple linear regression is a supervised machine learning method that aims to uncover a linear relationship between two continuous variables:
  - The explanatory/independent/input variable X.
  - The response/dependent/output variable Y.

- The ultimate goal is to use this relationship to make predictions about observations not within our dataset. We answer the question:
  - If I have the value of X, what should my best guess for Y be?

What is Simple Linear Regression?



Simple Linear Regression: Mathematically

- Ultimately, we wish to quantify this relationship as follows:

$$Y \approx \beta_0 + \beta_1 X$$

- Note that this equation is linear in form.

- In order to regress Y onto X, we need to estimate two coefficients/parameters:
  - β_0: The intercept of the line; the expected value of Y when X is 0.
  - β_1: The slope of the line; the expected change in Y when X shifts by one unit.

- Once we have estimates for β_0 and β_1, we can use the equation above to estimate the value of Y given any value of X.

Simple Linear Regression: Mathematically

- The prediction for Y based on the i^{th} value of X is as follows:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- We call the difference between the i^{th} observed response (the actual value) and the i^{th} response prediction (the estimated value) the residual or error e_i:

$$e_i = y_i - \hat{y}_i$$

- Of course, we would like the residual to be as small as possible for all observations in our dataset.
  - In other words, this is essentially an optimization problem where we would like to minimize error as much as possible.
  - How do we do this?

Simple Linear Regression: Mathematically

- Consider the sum of the squared error terms for each observation in our dataset. We call this the residual sum of squares (RSS):

$$RSS = \sum_{i=1}^{n} e_i^2$$

- Equivalently:

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$RSS = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Simple Linear Regression: Mathematically

- Task: find the estimates of β_0 and β_1 that reduce the sum of the squared vertical distances from the observations to the regression line (i.e., the RSS) as much as possible.

- Procedure: derive formulas for these estimates using basic calculus.

Simple Linear Regression: Mathematically

- Procedure for the intercept coefficient estimate:

$$RSS = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$\frac{\partial RSS}{\partial \hat{\beta}_0} = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(2)(-1) \stackrel{!}{=} 0$$

$$\Rightarrow n\hat{\beta}_0 = \sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_i)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Simple Linear Regression: Mathematically

- Procedure for the intercept coefficient estimate (continued):

$$\frac{\partial^2 RSS}{\partial \hat{\beta}_0^2} = \sum_{i=1}^{n}(-1)(2)(-1)$$
$$= 2n$$

- Because the second derivative is necessarily positive, we know that this estimate is a minimum.

Simple Linear Regression: Mathematically

- Procedure for the slope coefficient estimate:

$$
\begin{aligned}
RSS &= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\
&= \sum_{i=1}^{n}(y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i)^2 \\
&= \sum_{i=1}^{n}(y_i - \bar{y} - \hat{\beta}_1 (x_i - \bar{x}))^2 \\
\frac{\partial RSS}{\partial \hat{\beta}_1} &= \sum_{i=1}^{n}(y_i - \bar{y} - \hat{\beta}_1 (x_i - \bar{x}))(2)(-1)(x_i - \bar{x}) \stackrel{!}{=} 0 \\
\Rightarrow \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}
\end{aligned}
$$

Simple Linear Regression: Mathematically

- Procedure for the slope coefficient estimate (continued):

$$\frac{\partial RSS}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x}))(x_i - \bar{x})$$

$$\frac{\partial^2 RSS}{\partial \hat{\beta}_1^2} = -2\sum_{i=1}^{n}(x_i - \bar{x})(-1(x_i - \bar{x}))$$

$$= 2\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Because the second derivative is necessarily positive, we know that this estimate is a minimum.

Simple Linear Regression: Mathematically

- Thus, the least squares coefficient estimates for simple linear regression are:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Given our data, these are the best estimates for β_0 and β_1 as they ensure the sum of the squared vertical distances from the observations to the regression line (i. e., the RSS) is at a minimum.

Accuracy of the Coefficient Estimates

- Under the simple linear regression model, β_0 and β_1 exist in the universe as theoretical, true parameter values; however, we can only calculate an estimate based on our data. How can we quantify the accuracy of our estimates?

- We investigate their standard errors, which yield an approximation of how much our estimates vary from the true parameter values:

$$\widehat{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right]$$

$$\widehat{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Notice that the estimate of the standard error for the slope gets smaller as the spread of our observations increases; this allows us to better gauge its leverage.

Accuracy of the Coefficient Estimates

- But wait...we don't know the value of σ^2! Use the residual standard error, which serves as our best estimate for σ:

$$\hat{\sigma} = RSE = \sqrt{\frac{RSS}{n-2}}$$

- Now that we can calculate the estimates of the standard errors, we can use them to assess the accuracy of our coefficient estimates by:
  - Performing hypothesis tests.
  - Constructing confidence intervals.

Performing Hypothesis Tests

- For simple linear regression, the principal hypothesis test is as follows:
  - Null Hypothesis (H_0): $\beta_1 = 0$
  - Alternative Hypothesis (H_A): $\beta_1 \neq 0$

- What would it mean if the null hypothesis were true?
  - We would expect that the population mean of Y would be β_0 no matter what the value of X.
  - In other words, this would mean that X has no effect on Y!

- What would it mean if the null hypothesis were false?
  - We would expect that Y would vary with different values of X.
  - In other words, this would mean that X does have an effect on Y.

Performing Hypothesis Tests: The T-Test

- We perform this hypothesis test by calculating a t-statistic as follows:

$$t^* = \frac{\widehat{\beta}_1 - 0}{SE(\widehat{\beta}_1)} \sim t_{n-2}$$

- As we have previously seen, this will yield a t-value that we use to calculate the area under the associated theoretical distribution to assess the probability (p-value) of observing results at least as extreme as our own.

- Should the p-value be less than a certain threshold (< 0.05), we reject the null hypothesis in favor of the alternative.

Performing Hypothesis Tests: The F-Test

- For simple linear regression, this t-test is equivalent to the following F-test:

$$F^* = \frac{TSS/1}{RSS/(n-2)} \sim F_{1,n-2}$$

- Where the total sum of squares (TSS) is defined as follows:

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- **NB:** The t-test is sufficient for simple linear regression; however we will see that the F-test is more flexible and useful when we get to more advanced regression. (Theoretically, the t and F tests are related in the simple linear regression case such that $F = t^2$.)

Constructing Confidence Intervals

- It is often helpful to gauge approximate bounds of the true theoretical parameter values; to do exactly this, we construct a confidence interval.
  - Most commonly, a 95% confidence interval is a range of values that will contain the true, unknown value of the parameter with a probability of 95%.

- For the slope coefficient in simple linear regression, we can construct an approximate 95% confidence interval as follows:

$$\hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1)$$

- Note that the range of our confidence interval, and thus our uncertainty of the true parameter value, gets larger as the estimate of the standard error gets larger.

In Tandem: Hypothesis Testing & Confidence Intervals

- There is an inherent relationship between hypothesis testing and confidence intervals: they both aim to describe aspects of a theoretical parameter, but in different ways:
  - Hypothesis tests aim to describe the plausibility of a parameter taking on a specific value.
  - Confidence intervals aim to describe a range of plausible values a parameter can take on.

- If the value of the parameter specified by H_0 is contained within the 95% confidence interval, then H_0 cannot be rejected at the 0.05 p-value threshold.
- If the value of the parameter specified by H_0 is not contained within the 95% confidence interval, then H_0 can be rejected at the 0.05 p-value threshold.

PART 2

Assumptions & Diagnostics

Assumptions of Simple Linear Regression

- Recall that we wish to quantify a direct linear relationship between X and Y; however, it is likely that in reality there are other factors that tend to influence the behavior of Y, and other sources of variability (e.g., measurement error).

- To account for these unmeasured discrepancies, we say that the true linear model takes the following form:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Here, ϵ is the error term that accounts for all the shifts we might miss by just considering the relationship between X and Y; it accounts for the fact that the statistical model does not yield an exact fit to the data.

Assumptions of Simple Linear Regression

- With respect to the true linear model, it is necessary that we take into account the following assumptions:
  - Linearity
  - Constant Variance
  - Normality
  - Independent Errors

- We can succinctly write this as follows:

$$Y = \beta_0 + \beta_1 X + \epsilon, \quad \epsilon \overset{iid}{\sim} N(0, \sigma^2)$$

- The validity of our model depends on these assumptions being satisfied, most of which are attached to our error term.

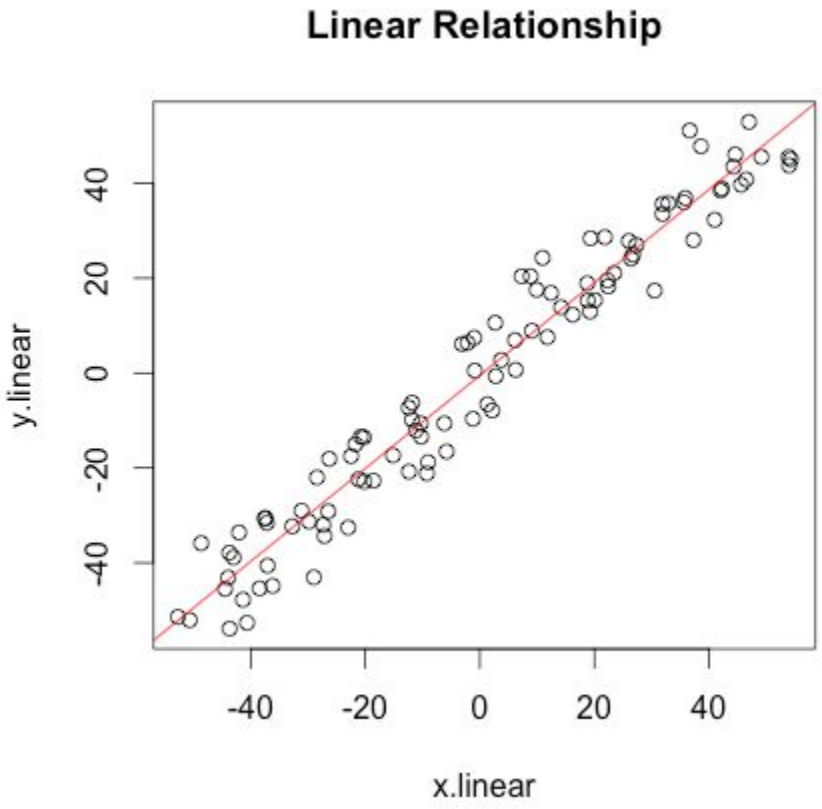
Assumptions of Simple Linear Regression: Linearity

- What does it mean?
  - The assumption of linearity implies that the underlying function connecting the independent variable to the dependent variable is indeed linear. If it were not, the model would be invalid and would not produce accurate estimates.

- How do we check?
  - Initially, the fastest and easiest way to check for linearity is to create a scatterplot of your data and assess whether a straight-line fit would roughly follow the pattern.
  - **NB:** We will delve into more sophisticated ways to determine if a linear relationship exists between our variables.

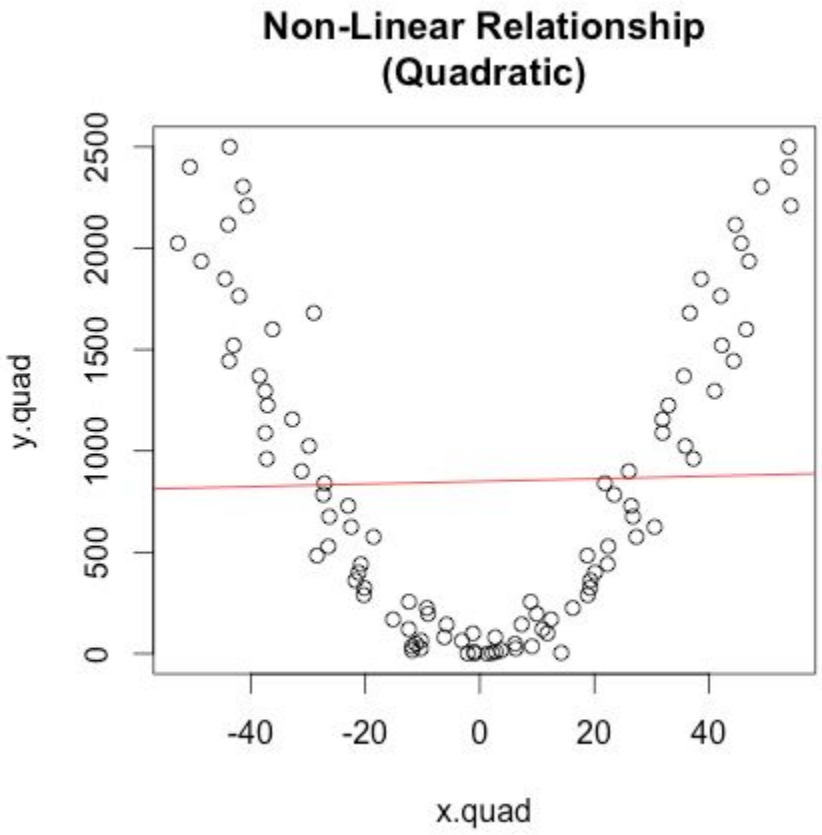
Assumptions of Simple Linear Regression: Linearity

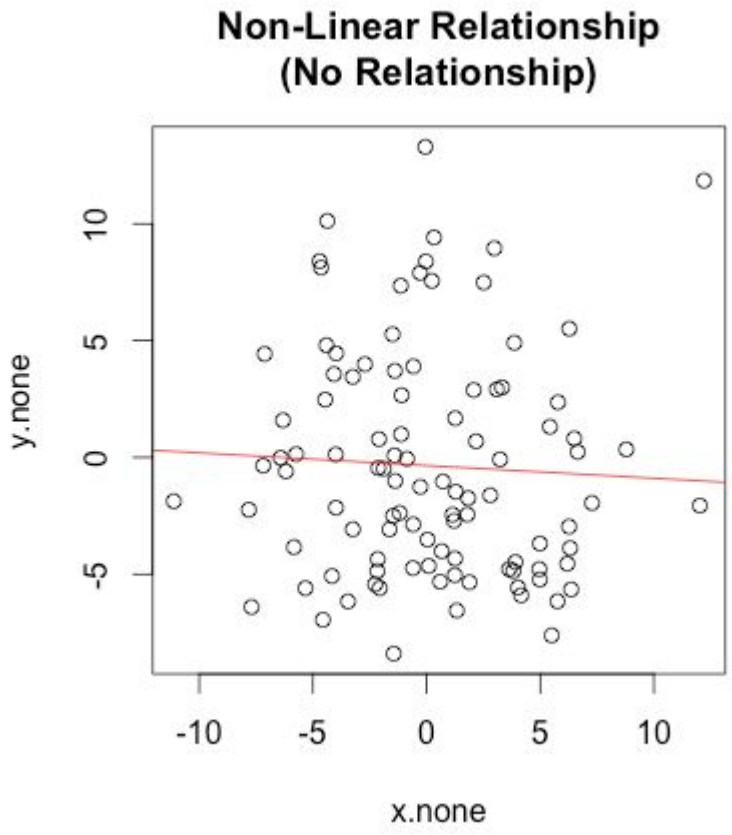
- What might linearity look like?



Assumptions of Simple Linear Regression: Linearity

- What might a violation in linearity look like?





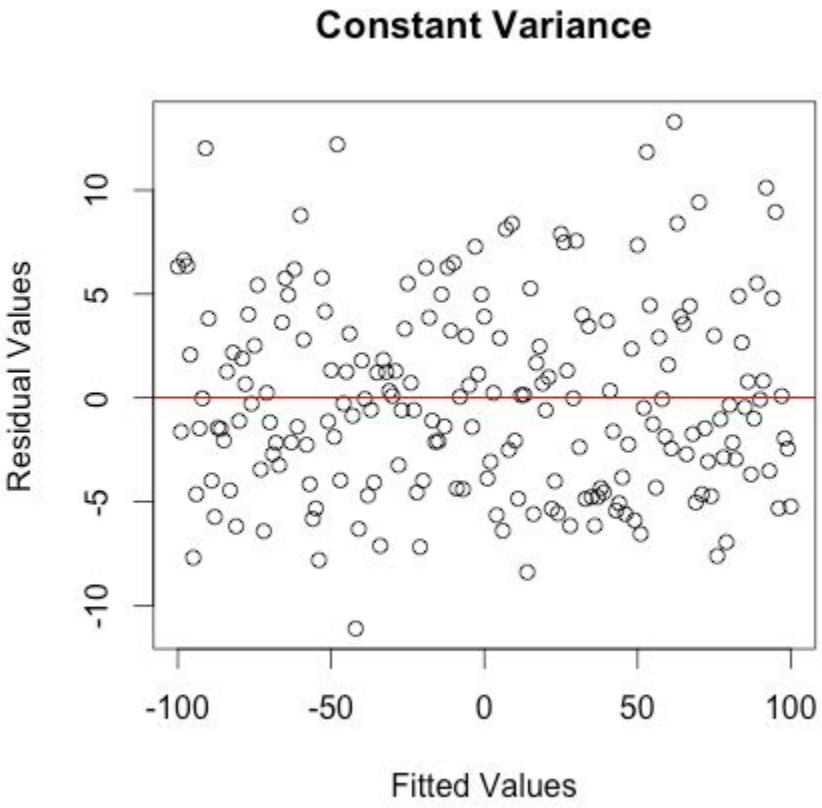
Assumptions of Simple Linear Regression: Constant Variance

- What does it mean?
  - The assumption of constant variance (aka homoscedasticity) implies that the error terms have the same variance no matter where they appear along the regression line.

- How do we check?
  - After fitting the regression, inspect the residual plot (a scatterplot of the residual values versus the fitted values).
  - Ideally, there would be no inherent pattern of varying ranges of residual values.

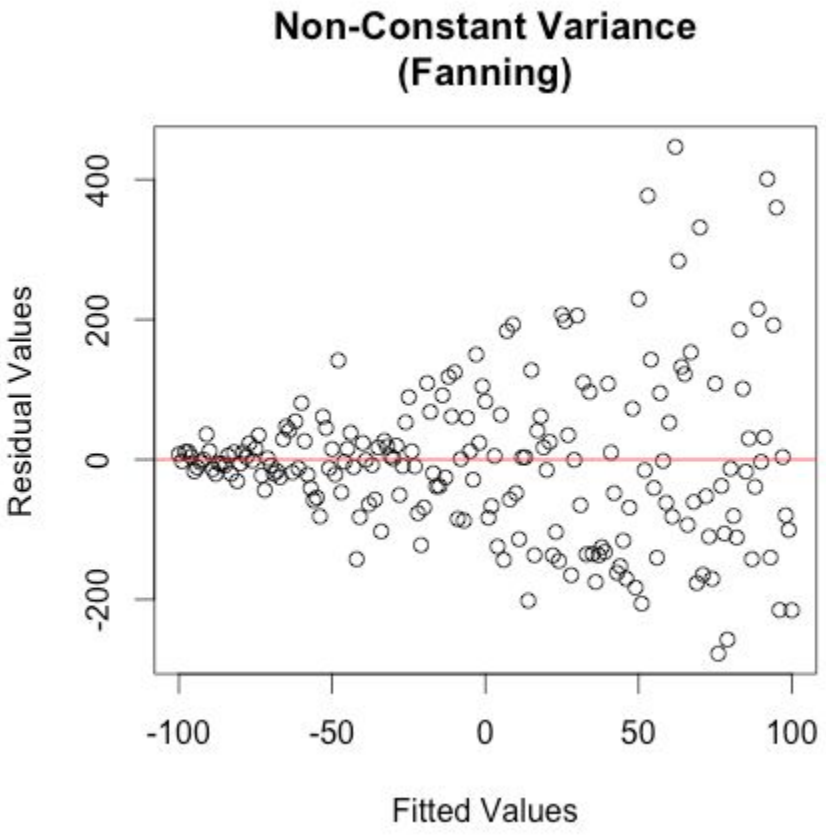
Assumptions of Simple Linear Regression: Constant Variance

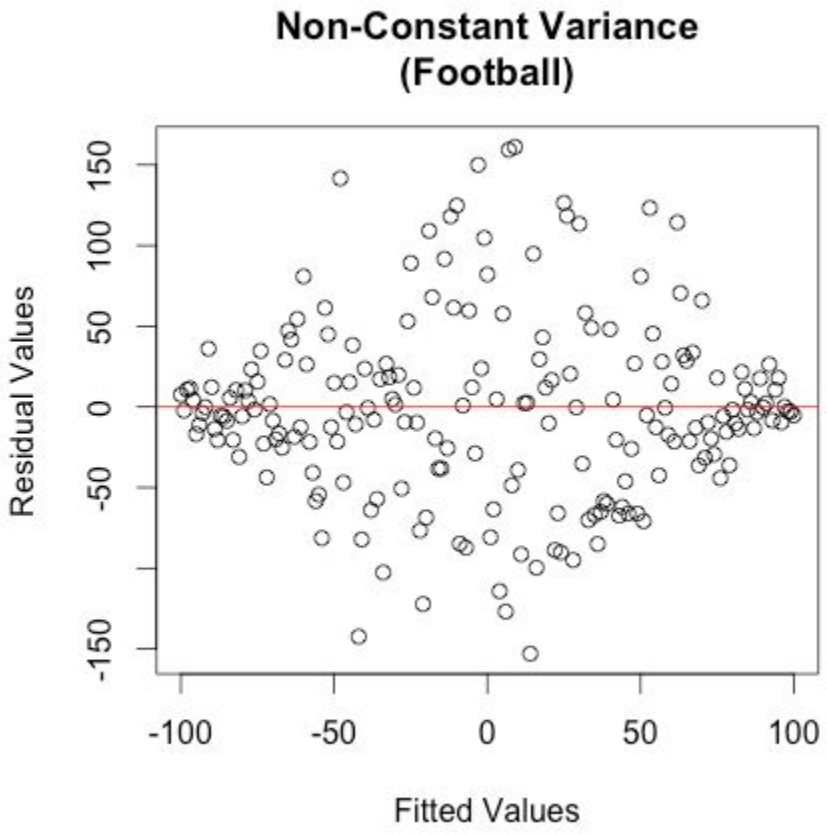
- What might constant variance look like?



Assumptions of Simple Linear Regression: Constant Variance

- What might a violation in constant variance look like?





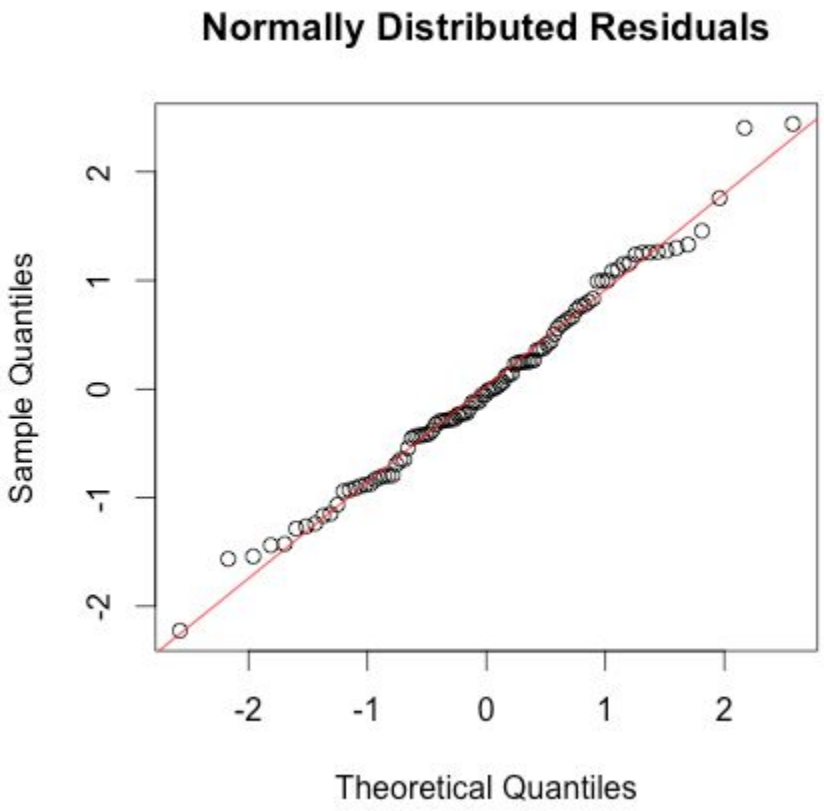
Assumptions of Simple Linear Regression: Normality

- What does it mean?
  - The assumption of normality implies that the error terms are drawn from an identical Gaussian distribution at each value of the explanatory variable. In other words, we assume a normal distribution of the dependent variable for each value of the independent variable.

- How do we check?
  - After fitting the regression, inspect the quantile-quantile plot of the residuals (a scatterplot of the residuals versus the corresponding normal theoretical value that preserves the observed quantile).
  - Ideally, this plot would display a straight-line relationship indicating that the residuals follow a pattern of normality.

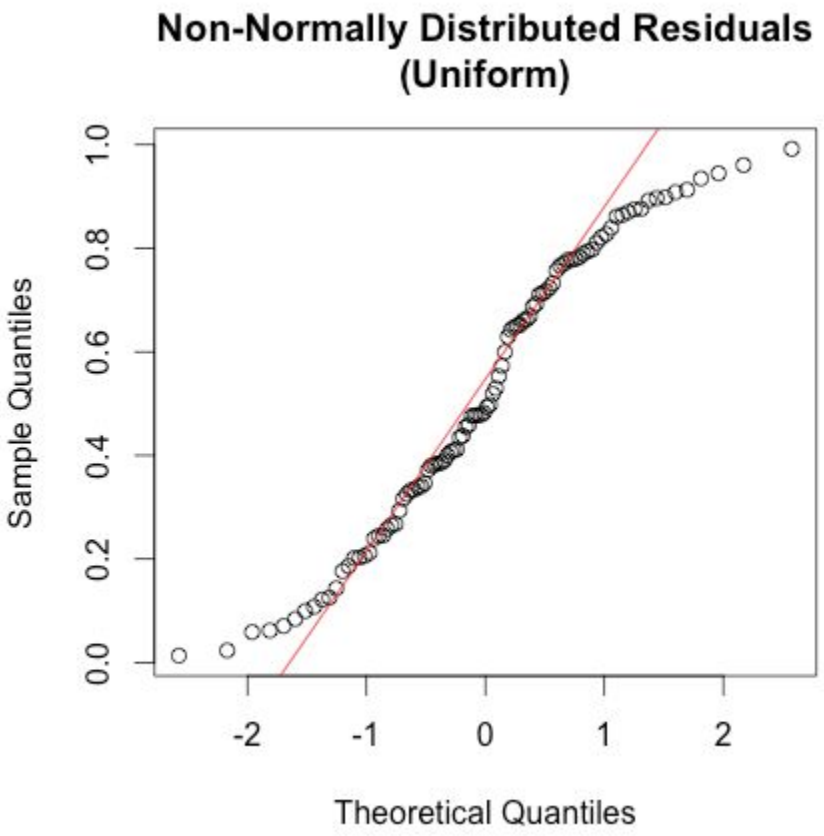
Assumptions of Simple Linear Regression: Normality

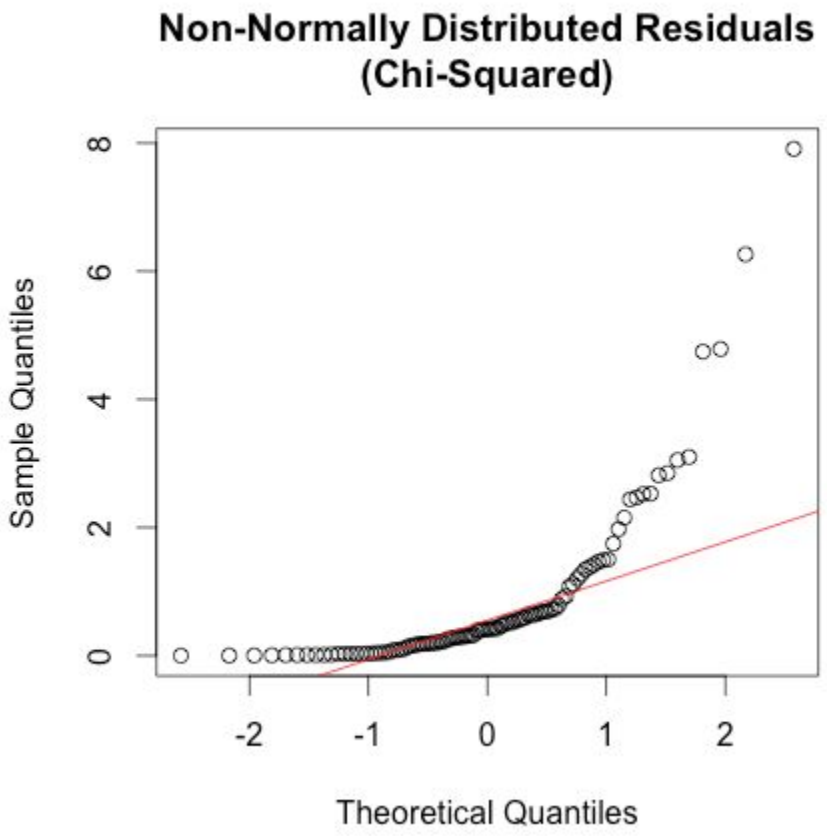
- What might normality look like?



Assumptions of Simple Linear Regression: Normality

- What might a violation in normality look like?





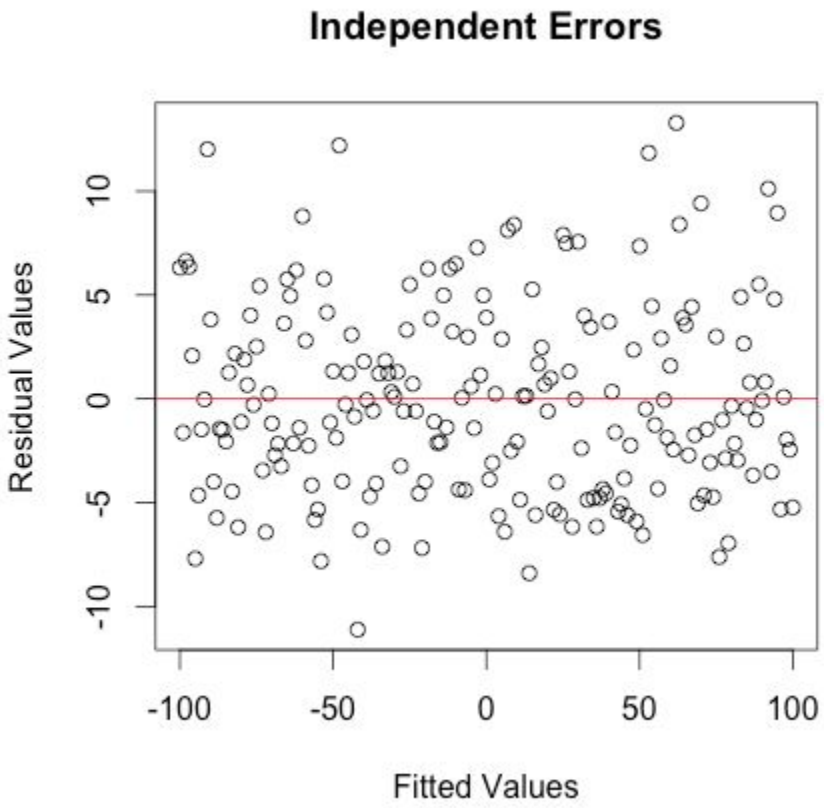
Assumptions of Simple Linear Regression: Independent Errors

- What does it mean?
  - The assumption of independent errors implies that the residual value for an arbitrary observation is not predictable from knowledge of another observation's residual value; they are uncorrelated.

- How do we check?
  - Although the assumption of independent errors usually isn't as readily visible as the other model assumptions, we can inspect the residual plot after fitting the regression. Ideally, there would be no clear pattern in the fluctuation of the error terms.
  - In general, we also judge the independent error assumption by analyzing the construct of how the data was collected in reference to the experimental design.

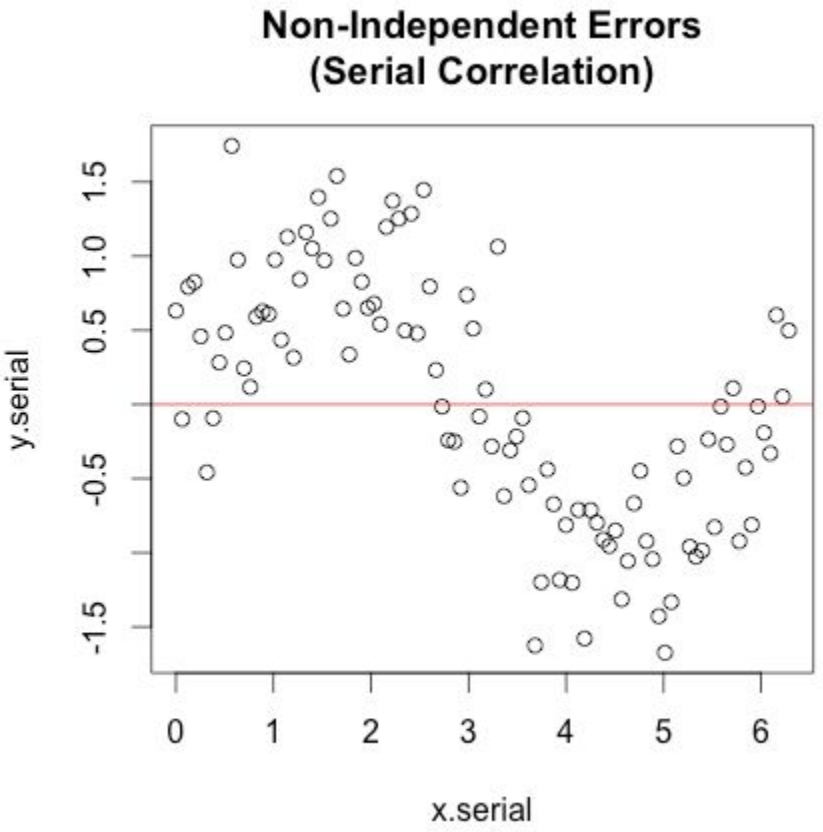
Assumptions of Simple Linear Regression: Independent Errors

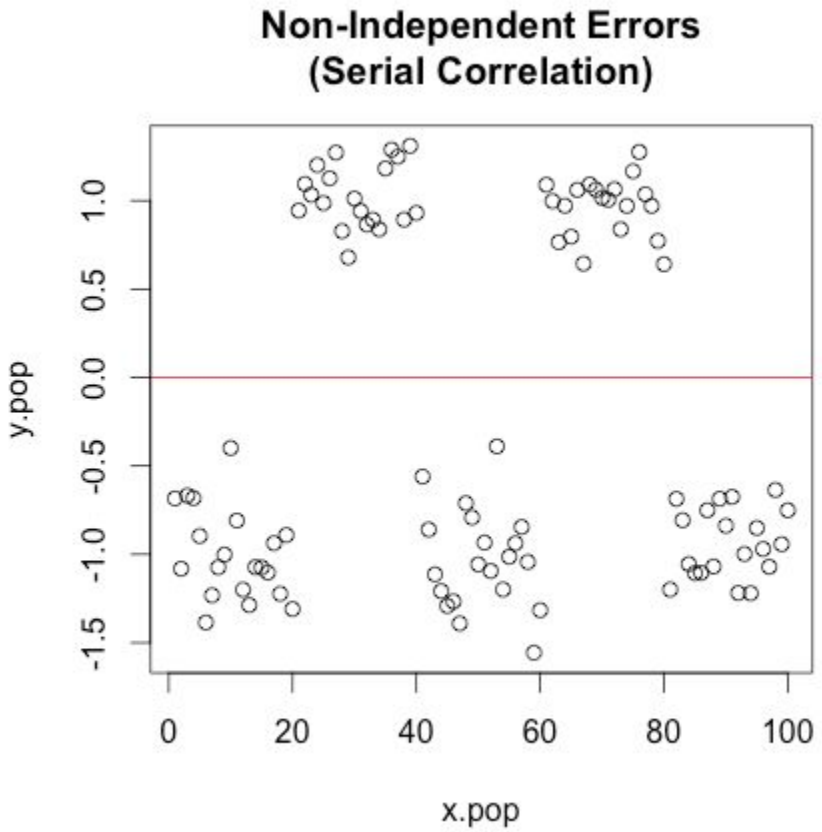
- What might independent errors look like?



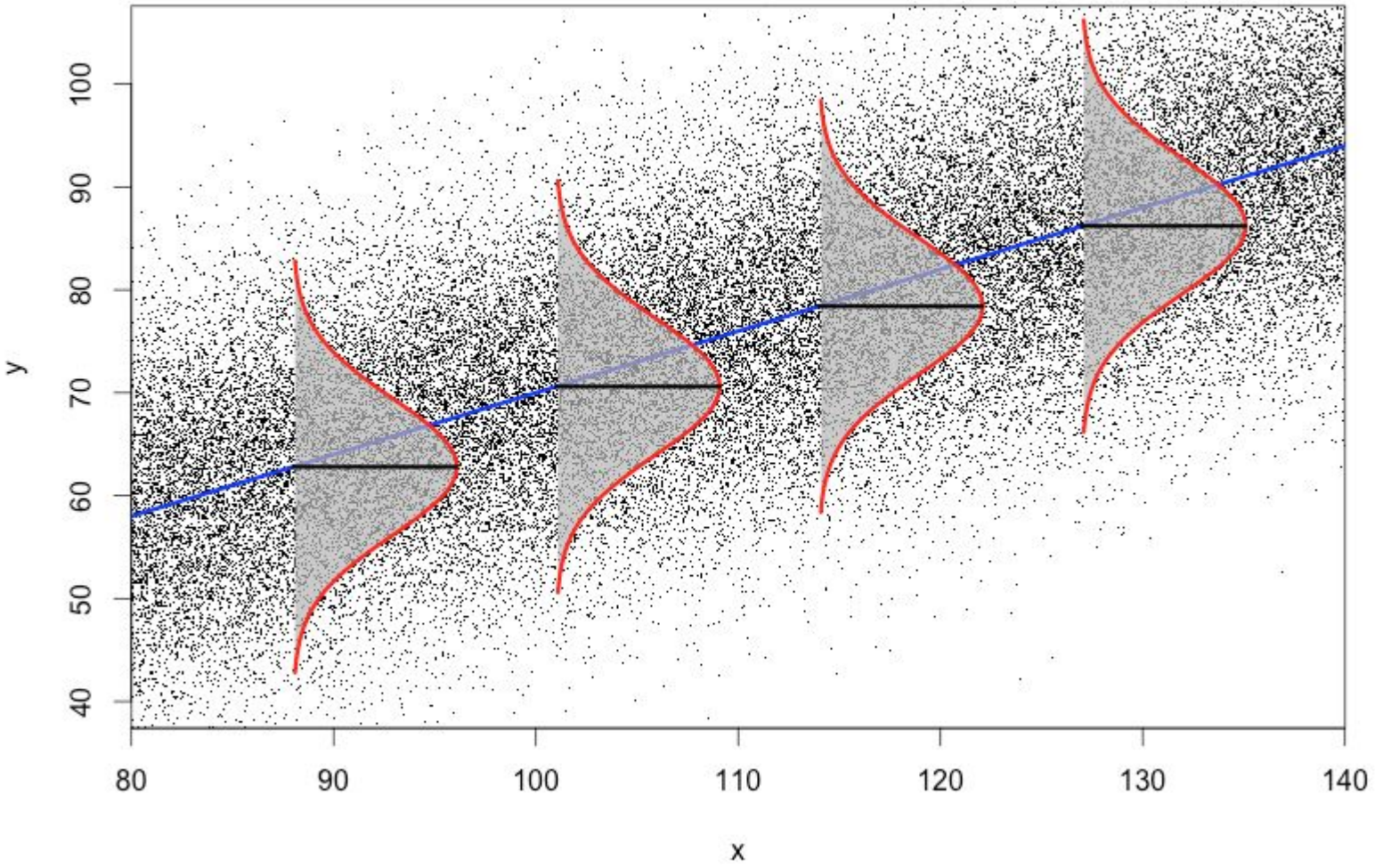
Assumptions of Simple Linear Regression: Independent Errors

- What might a violation in independent errors look like?





Visualizing All the Assumptions



PART 3

Transformations

Transforming Your Data

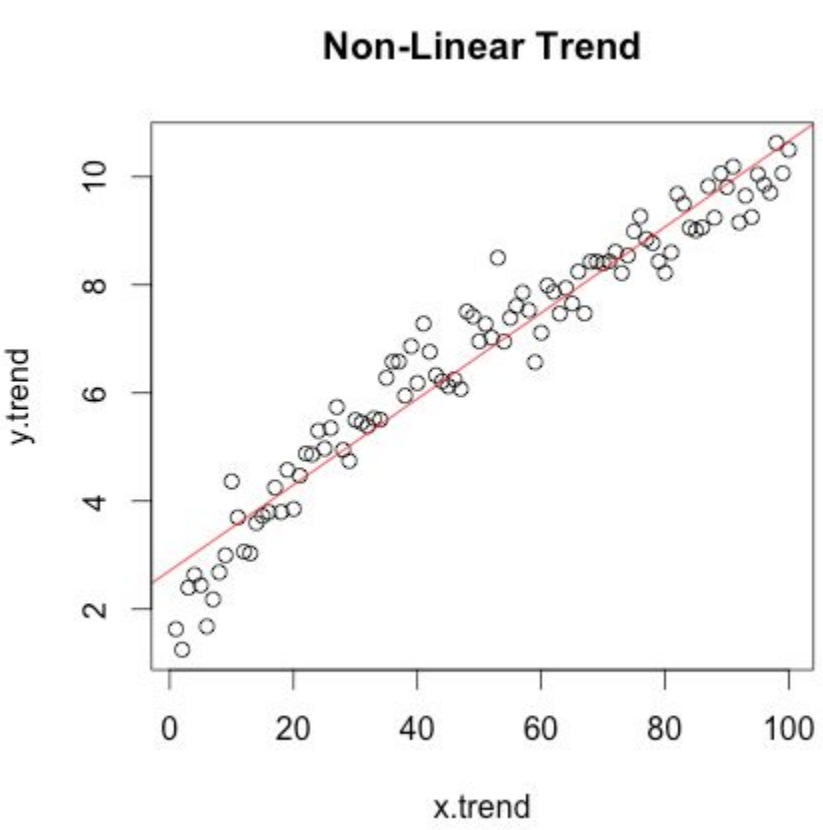
- Sometimes our raw data does not directly lend itself to a linear model and we may see that we violate some of the necessary assumptions. What can we do? Consider mathematically transforming the data.

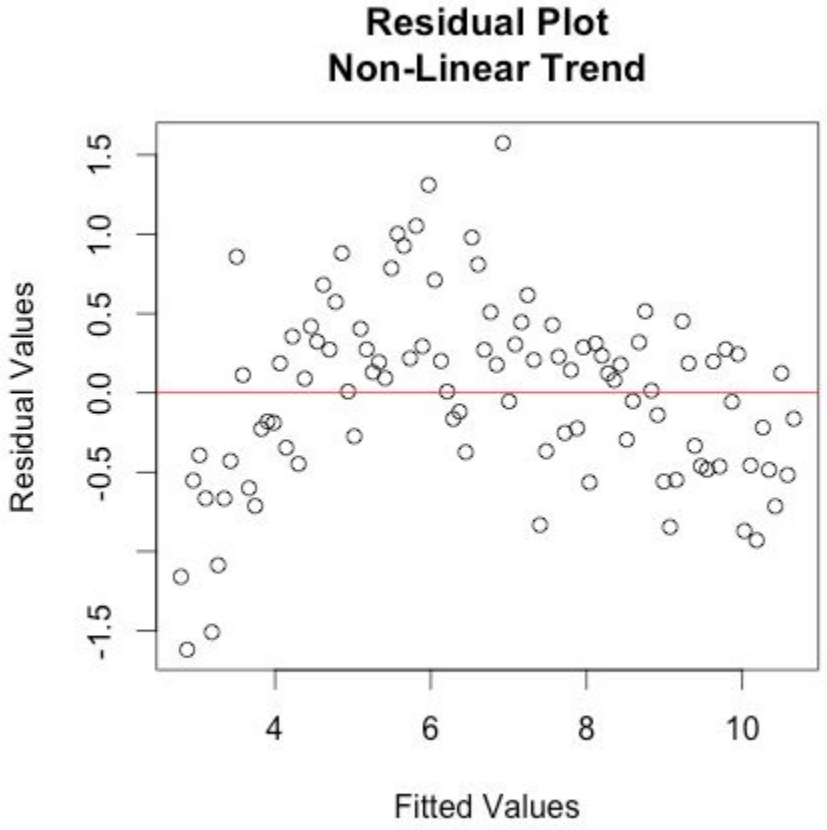
- Transformations are most helpful in some cases when the following assumptions are violated:
  - Linearity
  - Constant Variance
  - Normality

- Depending on the way in which a violation manifests, we may want to transform the explanatory or response variable.

Transforming Your Data

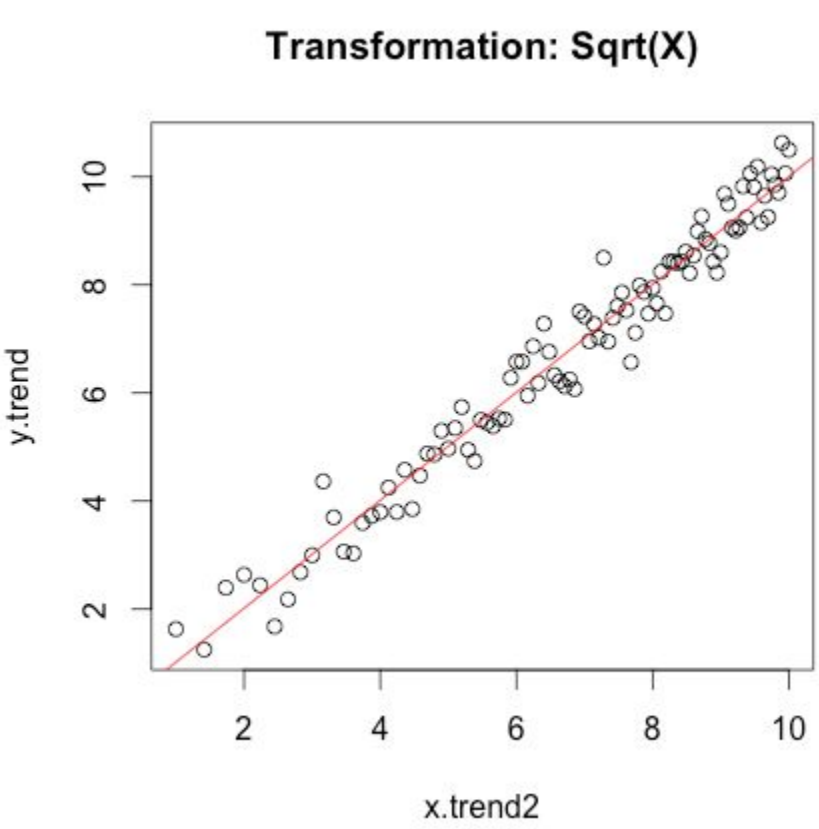
- Consider a scenario in which the linearity assumption is violated:

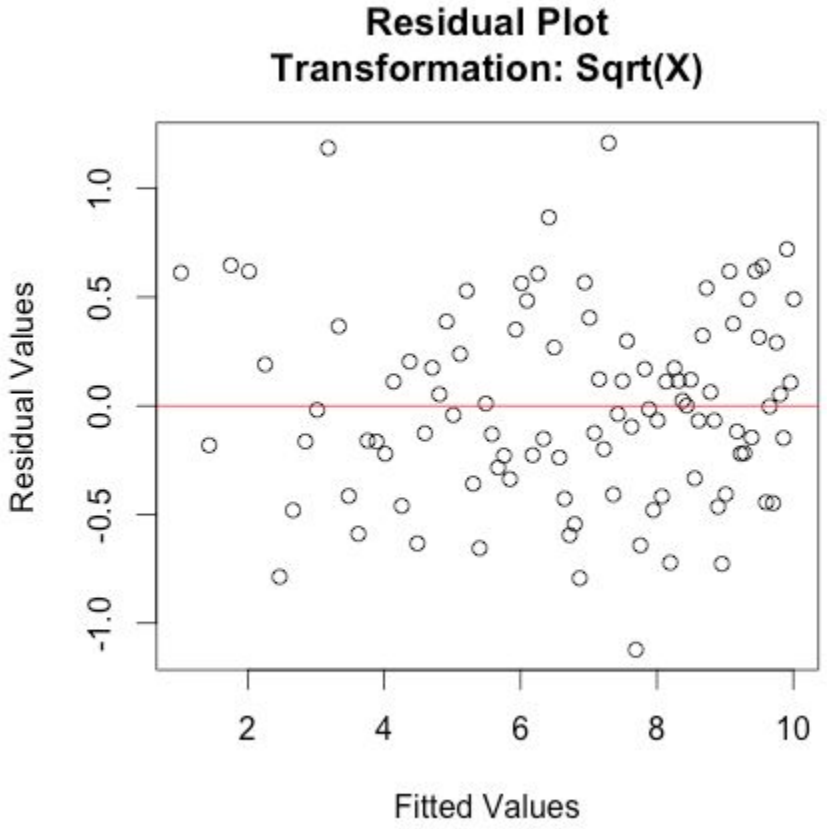




Transforming Your Data

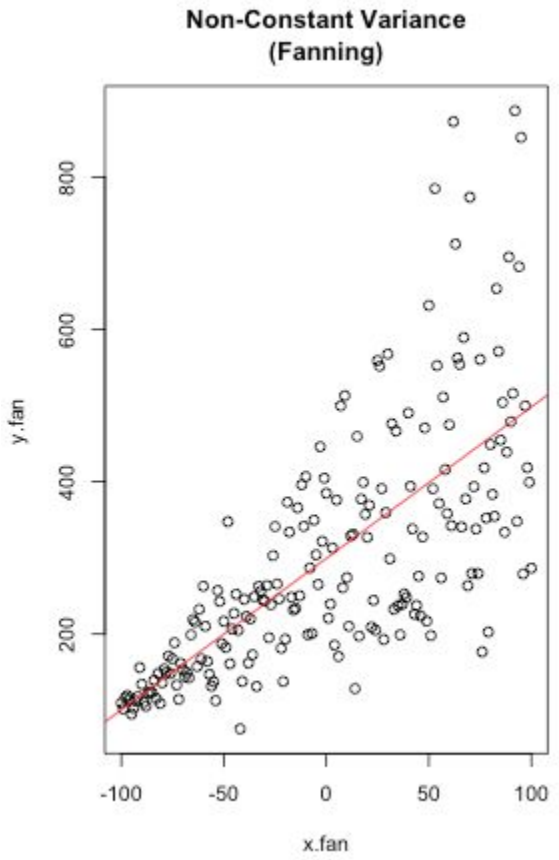
- In this case, it appears as though there is a non-linear, curved relationship. How can we "undo" this relationship? Consider a square root correction on X:

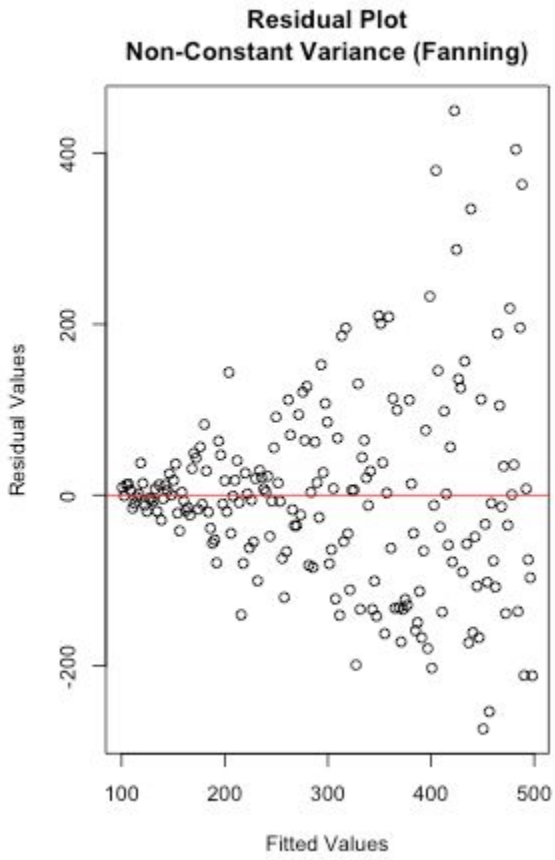


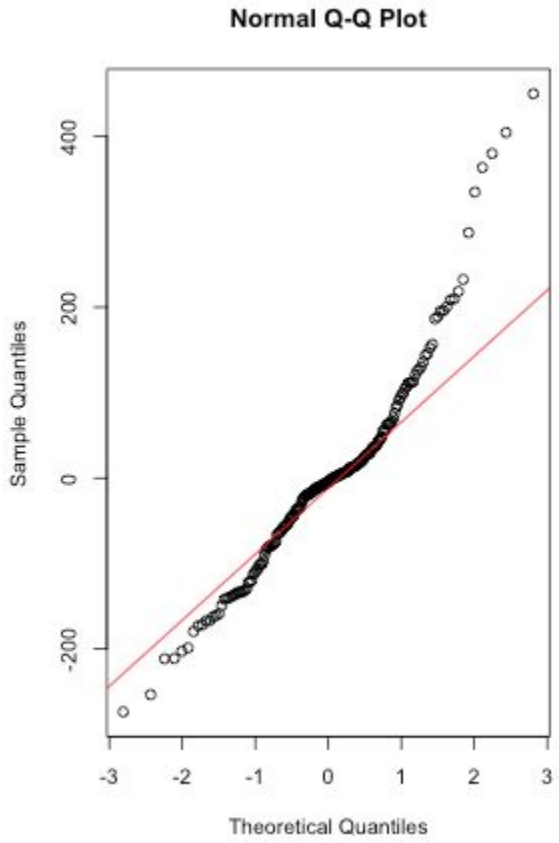


Transforming Your Data

- Consider a scenario in which the constant variance assumption is violated:

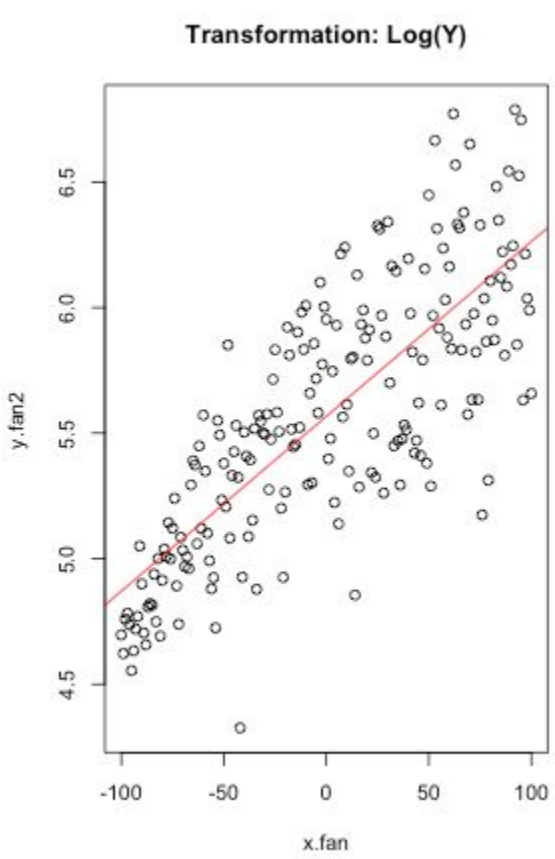


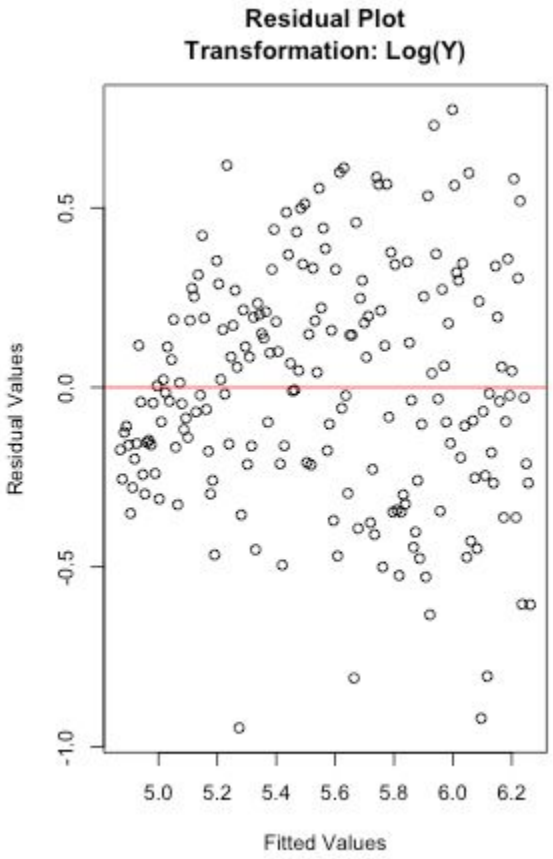


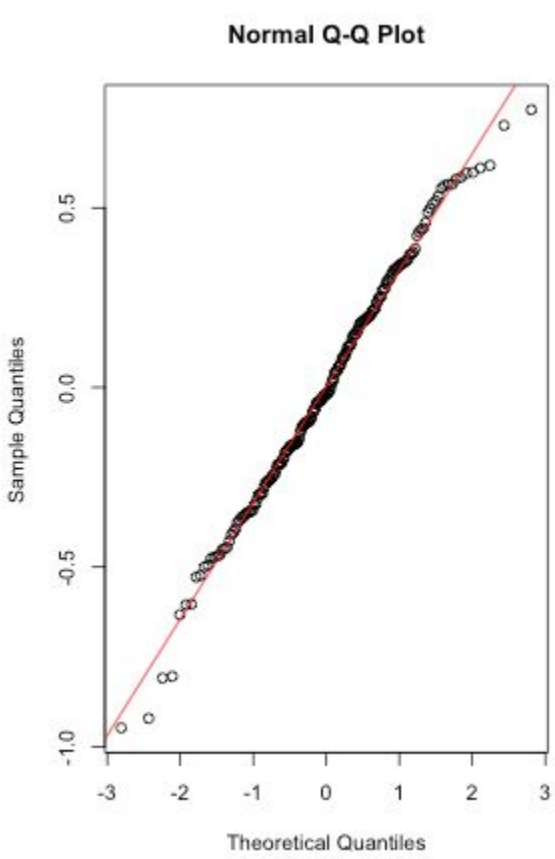


Transforming Your Data

- What if we could limit the scale of Y, and slow its rate of growth? Consider a log transformation on Y:

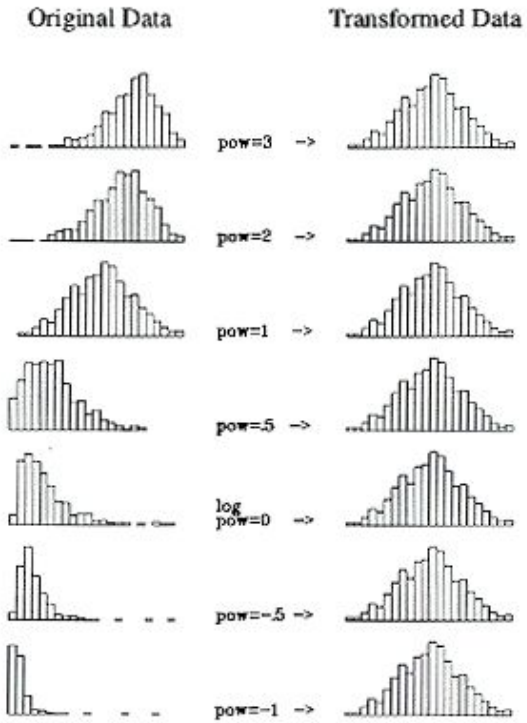






How Can We Choose the Correct Power Transform?

- In general, to help make the distribution of a single variable more symmetric in shape:
  - Powers > 1 are good for transforming data that is skewed left.
  - Powers < 1 are good for transforming data that is skewed right.



The Box-Cox Transformation

- A better method that tends to help correct for multiple violations at once is called the Box-Cox transformation.
- Given a value of λ, the Box-Cox transformation is defined as follows:

$$y' = \begin{cases} \frac{(y^\lambda - 1)}{\lambda} & \lambda \neq 0 \\ log(y) & \lambda = 0 \end{cases}$$

- The Box-Cox procedure iterates along values of λ by maximum likelihood so as to maximize the fit of the transformed variable to a normal distribution.

The Box-Cox Transformation: Caution

- The Box-Cox transformation does not guarantee that the transformed data will be normally distributed. Why?
  - The method is actually attempting to minimize the standard deviation with respect to the choice of λ.
  - We assume that when the standard deviation is smallest, the transformed data has the highest likelihood of being normally distributed.

- The Box-Cox transformation can only be used on positive values. What if we have negative values in our data?
  - Shift all values in your dataset up by a constant that renders all values positive, then apply the Box-Cox transformation procedure.

Transforming Your Data: Pros & Cons

- Some pros of data transformations:
  - Can help remedy some assumption violations and thus increase the validity of our model.
  - Can help strengthen the linear relationship between our variables.

- Some cons of data transformations:
  - Tends to make model interpretability more difficult.
  - Could lead to overfitting.

PART 4

The Coefficient of Determination R^2

How Well Does the Model Fit?

- Once we fit a valid model (i.e., a model that does not violate any assumptions), how can we assess how well the model fits the data?

- We could look at the RSS and RSE as measures of the lack of fit of the model.
  - Recall that these values essentially measure the average deviations of the model estimates to the true values:
    - Smaller RSS & RSE values indicate a better fit.
    - Larger RSS & RSE values indicate a worse fit.

- **NB:** The RSS & RSE are measured in terms of units of Y. What determines "good" or "bad" values? What determines "small" or "large" values?

How Well Does the Model Fit?

- Suppose we have a measure of the total variability of the response variable. If so, we could contrast this with the amount left unexplained by our model (the RSS) in order to gauge overall accuracy.

- To measure the amount of variability inherent in our dependent variable, recall that we define the total sum of squares as follows:

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- Intuitively, this is a measure of the total squared deviation of our response variable from its mean value: our best guess in the worst case scenario (no outside information).

The Coefficient of Determination R^2

- A better measure of model fit is to assess the coefficient of determination R^2. This value is defined as the proportion of total sample variability in the dependent variable that is explained by the regression model:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

- R^2 will always be bound between 0 and 1, regardless of the units of measurement used for our response variable.

- A higher R^2 value indicates a better model fit (i.e., a greater percentage of the variability in the response variable has been explained by the explanatory variable).

PART 5

Review

Review

- Part 1: Simple Linear Regression
  - What is Simple Linear Regression?
  - Simple Linear Regression: Mathematically
  - Accuracy of the Coefficient Estimates
  - Performing Hypothesis Tests
    - The T-Test
    - The F-Test
  - Constructing Confidence Intervals
  - In Tandem: Hypothesis Testing & Confidence Intervals
- Part 2: Assumptions & Diagnostics
  - Assumptions of Simple Linear Regression
    - Linearity
    - Constant Variance
    - Normality
    - Independent Errors
  - Visualizing All the Assumptions
- Part 3: Transformations
  - Transforming Your Data
  - How Can We Choose the Correct Power Transform?
  - The Box-Cox Transformation
    - Caution
  - Pros & Cons
- Part 4: The Coefficient of Determination R^2
  - How Well Does the Model Fit?
  - The Coefficient of Determination R^2