Simple Linear Regression Model Assumptions

Module 2

DATA 5600
Introduction to Regression and Machine Learning for Analytics
Marc Dotson

Module Overview

- Assumptions underlying a linear model
- Consequences of unmet assumptions
- Graphical and numerical diagnostics for determining if assumptions are met
- Remedial measures for when assumptions are not met

Model Checking

- Before trusting any conclusions (of significance) from a statistical model, we must first check that the model assumptions are met. Why?

Linear Regression Model Assumptions

1. **L** – X vs Y is linear
2. **I** – The residuals are independent (ex: knowing e_1 is positive should provide no information about e_{i+1})
3. **N** – The residuals are normally distributed and centered at zero
4. **E** – The residuals have equal (constant) variance σ^2 across all values of X (homoscedastic)
5. **A** – The model describes all observations (i.e., there are no influential points)
6. **R** – Additional predictor variables are not required

Linear Regression Model Assumptions

- We check assumptions using “diagnostics”
- We fix violated assumptions using “remedial measures”
- Violations are most apparent in the error terms $(\epsilon_1, \epsilon_2, \ldots \epsilon_n)$, so we focus on residuals $(e_1, e_2, \ldots, e_n)$. WHY?

- Graphical and numerical diagnostics exists – graphical diagnostics are much more informative

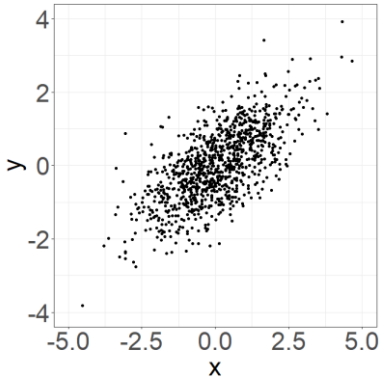
L – X vs Y is linear

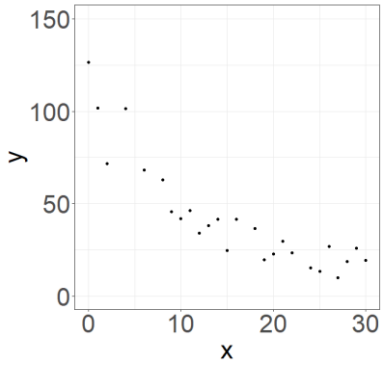
L – X vs Y is Linear Diagnostics

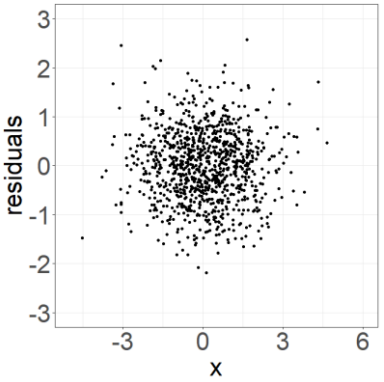
1. Scatterplot of X vs Y
   - Point pattern should be roughly linear or cloud-like (should not be parabolic, sinusoidal, etc.)

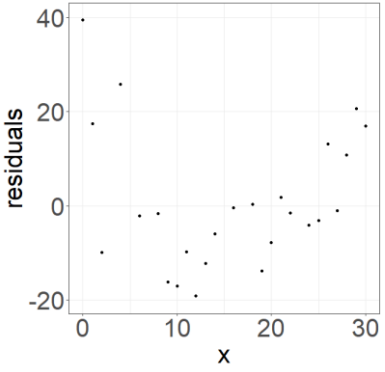
2. Scatterplot of residuals vs X
   - Point pattern should be roughly linear or cloud-like

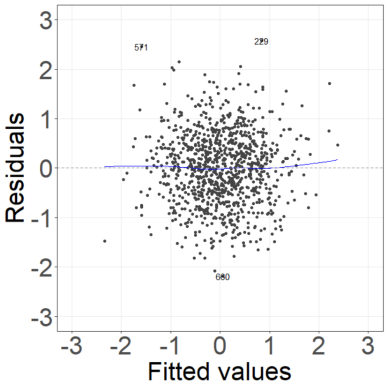
3. Scatterplot of residuals vs fitted values $(\hat{y}_i)$
   - Blue smoother curve should be roughly horizontal

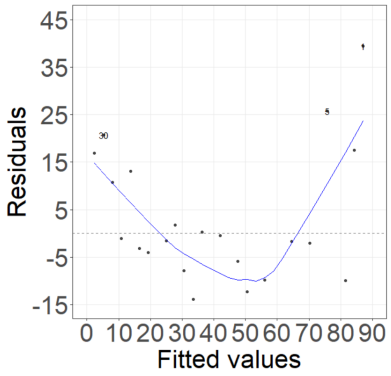












L – X vs Y is Linear Violations

- If X vs Y is **NOT** linear:
  - Everything breaks! Do not fit a line to non-linear data!
  - First, consider transforming X. If X vs Y is still not linear, then consider transforming Y as well (leave the error terms alone, if possible, if linearity is the only assumption violated)
  - You may need a nonlinear model – more about these in DATA 5610

– The residuals are independent

I – The Residuals are Independent

ZA

In linear regression, there is an assumption that the rows are independent. Can you explain what that means exactly, when that assumption would be met, and when it would be violated? Explain in layman's terms.

In linear regression, the assumption that rows are independent means that each data point or observation in your dataset is not influenced by or related to the others in a way that could affect the accuracy of your analysis.

When this assumption is met:

- It means that your data points are collected or sampled in a way where they truly represent different, unrelated cases or events. For example, if you're studying the relationship between a person's age and their income, each person in your dataset should be unique and not connected in a way that one person's income depends on or is influenced by another person's.

When this assumption is violated:

- It means that there is some form of relationship or dependency between the data points that you haven't taken into account. For example, if you're analyzing test scores of students in a classroom, and you have data from multiple tests for each student, the test scores for a particular student are likely not independent because they are influenced by the student's ability, which is constant across tests.

In practical terms, violating the assumption of independence can lead to incorrect or biased results in your linear regression analysis. To address this violation, you might need to consider more advanced statistical techniques or account for the dependencies in your data to obtain accurate insights.

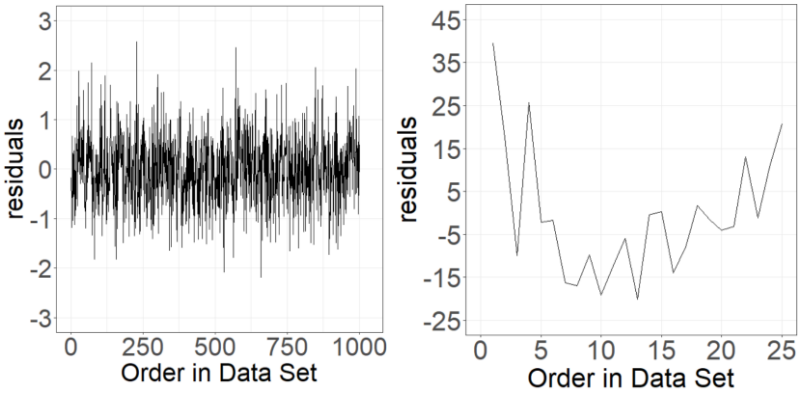
I – The Residuals are Independent

Common ways this assumption is violated:

- Repeated measures
  - Observations are collected from the same individual/unit so that each individual/unit is represented multiple times in the data set
  - Ex: Data of pre- and post-test scores for a class of students
- Clustered data
  - Observations are grouped into clusters such that data within clusters are more similar than data across clusters
  - Ex: Data of elementary students from three different schools
- Temporal correlation
  - Observations are collected in regular time intervals such that sequential data points vary in similar ways
  - Ex: daily wastewater measurements
- Spatial correlation
  - Observations are collected across space/geography such that data points that are next to each other are similar
  - Ex: Data from oil drilling sites in Texas

I – The Residuals are Independent Diagnostics

1. Sequence Plot *only if appropriate*
   - Plot residuals $(e_1, \ldots, e_n)$ vs row order $(1, \ldots, n)$
   - *Meaningful only if the observations are in some natural order* (e.g. measured each day for a year)
   - Should be no trends in the mean or variance



2. Think about how the data was collected (*see next slide)
   - Does the data description specifically mention the data were randomly sampled?
   - Is there any reason to believe one row is correlated with another row?

Assessing Independence Examples

Is the independence assumption met?

1. Several average socioeconomic variables were collected from all 50 U.S. states. Researchers are interested in if they can predict life expectancy based on these variables.

2. Daily RedBox movie rental sales were collected over the year of 2018. Researchers are interested in if the number of new releases is related to the number of rentals.

3. Students registered in Fall 2023 were randomly sampled, and their academic performances were recorded. Researchers are interested in if average GPA depends on the number of registered classes.

4. We have several years' worth of college basketball teams' data. We want to determine if the teams' adjusted defensive efficiency can predict the number of wins.

5. Elementary students' vegetable intake is recorded in a school district. Each school within the district performed different interventions to try to increase vegetable intake. The goal is to see which intervention was most successful.

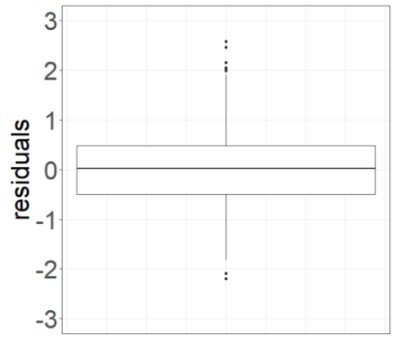
I – The Residuals are Independent Violations

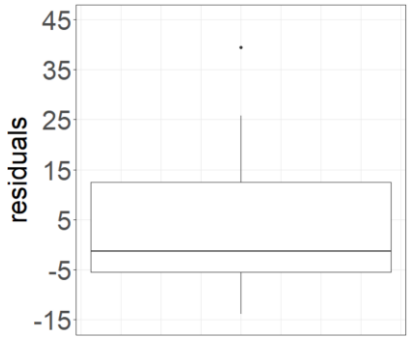
- If the residuals are **NOT** independent:
  - Standard errors of the estimates are typically too small, resulting in artificially narrow confidence intervals and "significant" results
  - Consider adding omitted variables that may explain the bias
  - More likely, you'll need a different model
    - Time series model
    - Spatial model
    - Hierarchical model
    - Repeated measures/longitudinal data model

N – The residuals are normally distributed and centered at zero

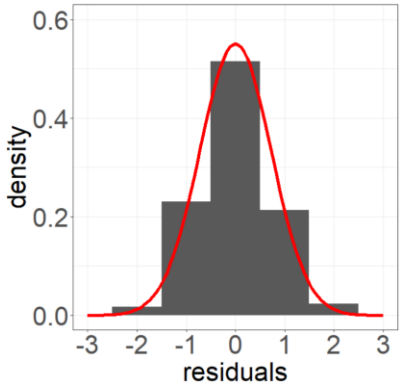
N – Residuals are Normally Distributed Diagnostics

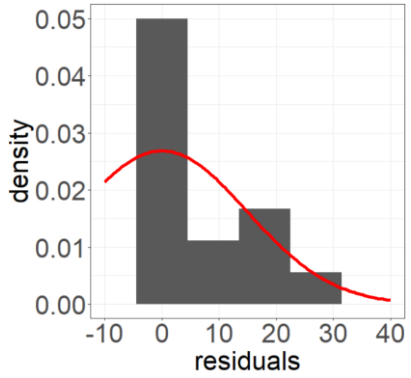
1. Boxplot
   - Plot ordered residuals $(e_{(1)}, \ldots, e_{(n)})$
   - Recall:
     - "Box" is the first and third quartile
     - "Whiskers" usually go up to 1.5×"IQR"
     - Points beyond whiskers are potential outliers





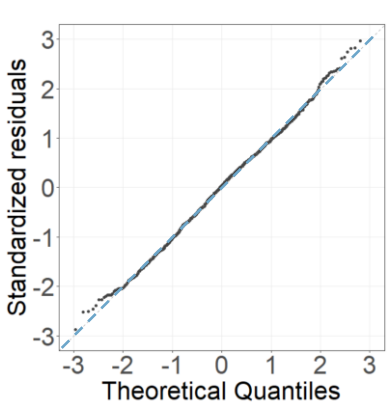
2. Histogram
   - Plot ordered (standardized) residuals $(e_{(1)}, \ldots, e_{(n)})$
   - Superimpose a normal curve

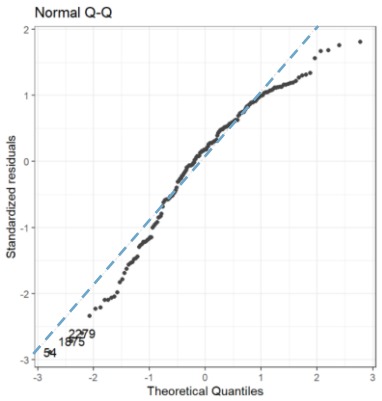




N – Residuals are Normally Distributed Diagnostics

3. Q-Q Plot
   - Plot ordered residuals $(e_{(1)}, \ldots, e_{(n)})$ against the expected values from the normal distribution
   - Key idea: If we have n observations from a particular distribution, then the observed and the expected percentiles should be fairly close
   - Points should roughly follow a diagonal line, especially within ± 2 on the X axis





4. Shapiro-Wilk Test
   - Mathematical details are beyond the pre-requisites for this class
     - H_0: Data come from a normal distribution
     - H_A: Data do not come from a normal distribution

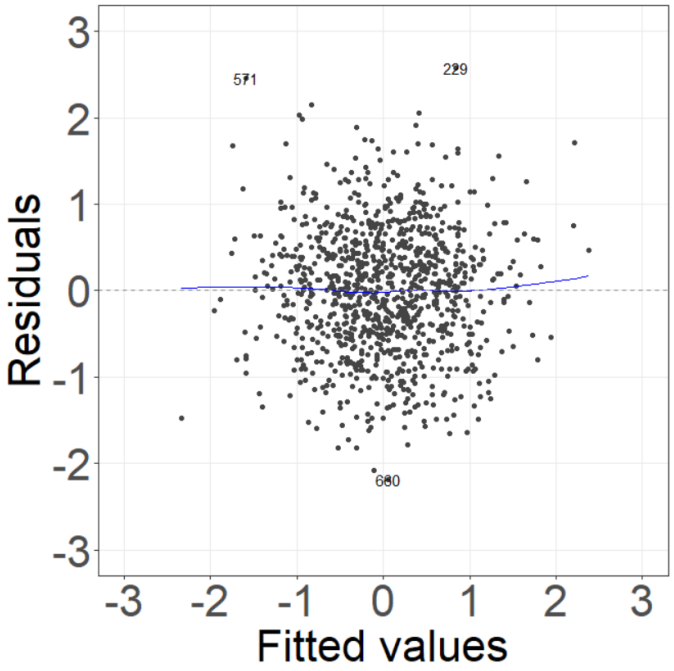
N – Residuals are Normally Distributed Violations

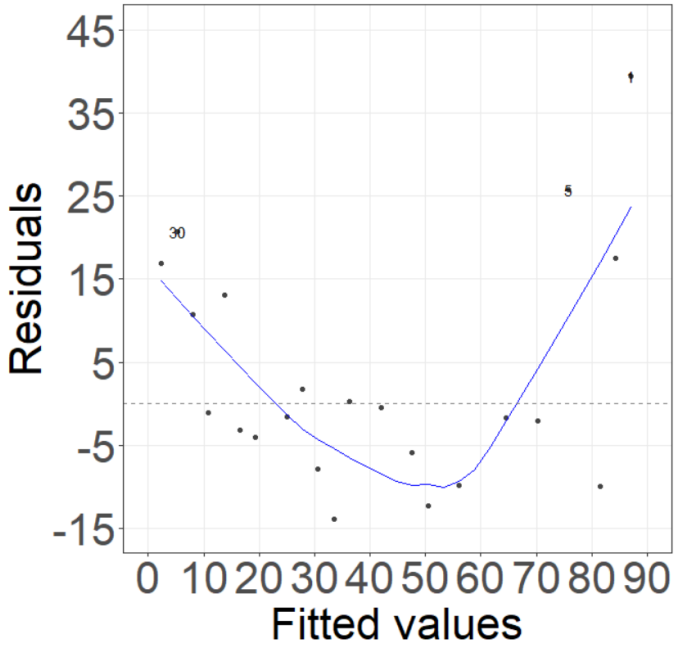
- If the residuals are **NOT** normally distributed and centered at zero:
  - Confidence intervals for the estimates are wrong (we cannot use the t-distribution)
  - First, consider transforming Y. If residuals are still not normally distributed, then consider transforming X as well

E – The residuals have equal (constant) variance across all values of X (homoscedastic)

E – Residuals have Equal Variance Diagnostics

1. Scatterplot of residuals vs fitted values $(\hat{y}_i)$
   - Vertical spread from right-to-left should be constant in height with no distinct patterns (should not see funnel-shaped point patterns)





E – Residuals have Equal Variance Violations

- If the residuals do **NOT** have equal variance σ^2 across all values of X:
  - Standard errors of the estimates are wrong (and we don't know how wrong or in what direction)
  - First, consider transforming Y. If residuals still do not have equal variance, then consider transforming X as well

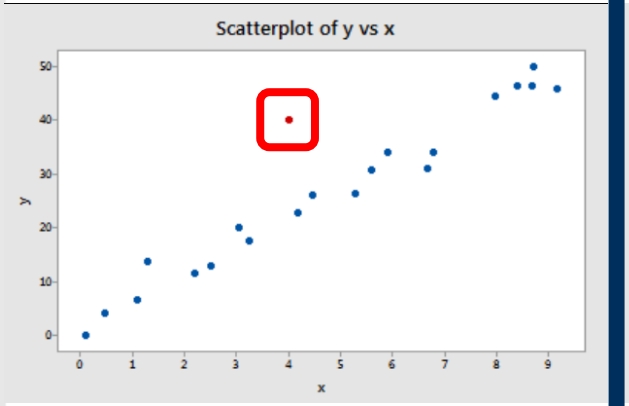
A – The model describes all observations (i.e., there are no influential points)

Influential Points vs Outliers

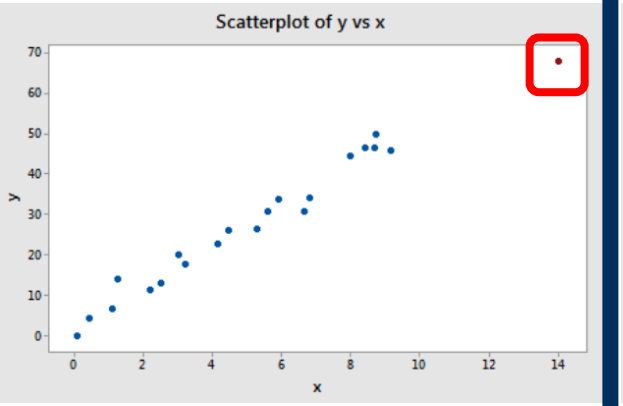
- An influential point is one that unduly influences any part of the regression analysis
- "Outliers" and "leverage points" are both **potential** influential points
  - An outlier is a point whose response does not follow the general trend of the rest of the data (extreme Y values, given the trend)
  - A leverage point is a point with an "extreme" predictor value (extreme X values)

Influential Points vs Outliers

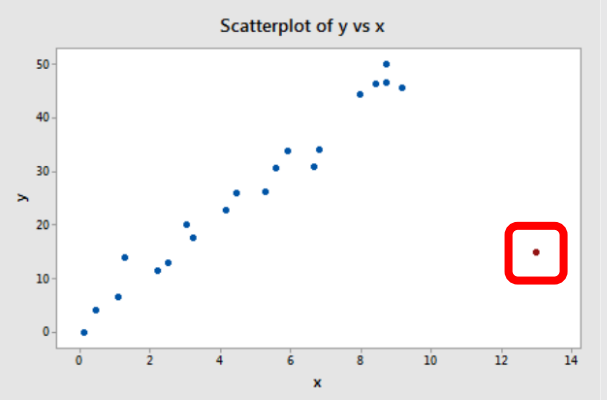
A



B



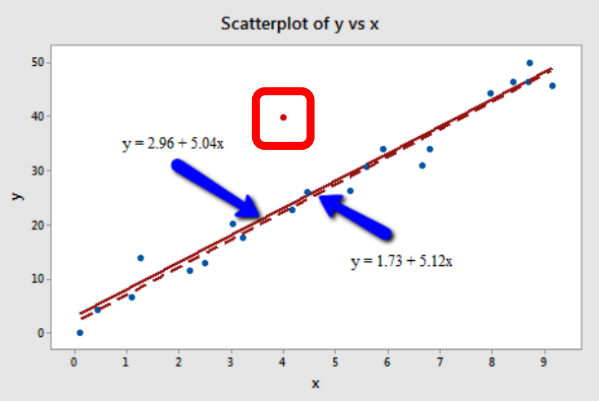
C



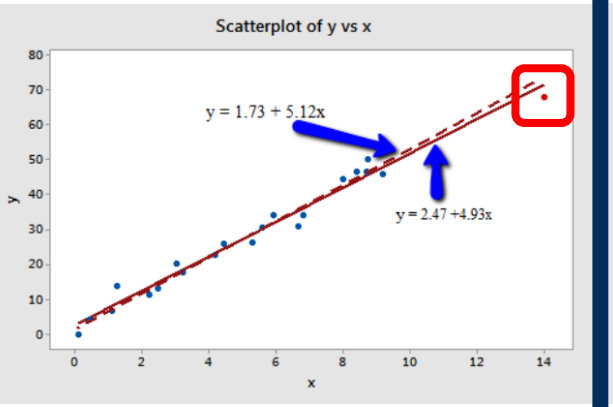
Examples taken from https://newonlinecourses.science.psu.edu/stat462/node/170/#targetText=In%20short%3A,is%20particularly%20high%20or%20low.

Influential Points vs Outliers

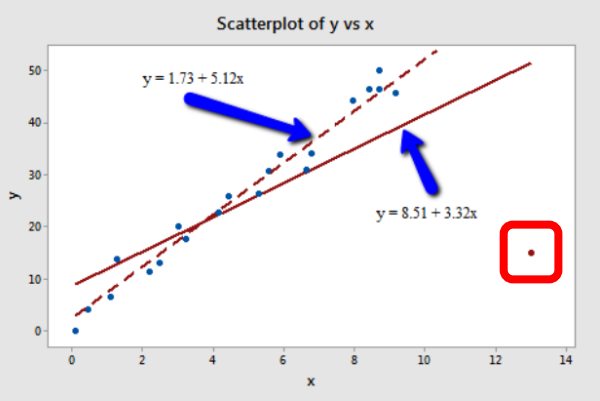
A



B



C



Examples taken from https://newonlinecourses.science.psu.edu/stat462/node/170/#targetText=In%20short%3A,is%20particularly%20high%20or%20low.

A – Model Describes All Observations Diagnostics

1. DFBETAS
   - More details in the following slides
   - Look for points that meet **both** of these requirements:
     - Above the red, dashed line
     - Relatively far away from the main body of points

2. DFFITS
   - More details in the following slides
   - Look for points that meet **both** of these requirements:
     - Above the red, dashed line
     - Relatively far away from the main body of points

DFBETAS

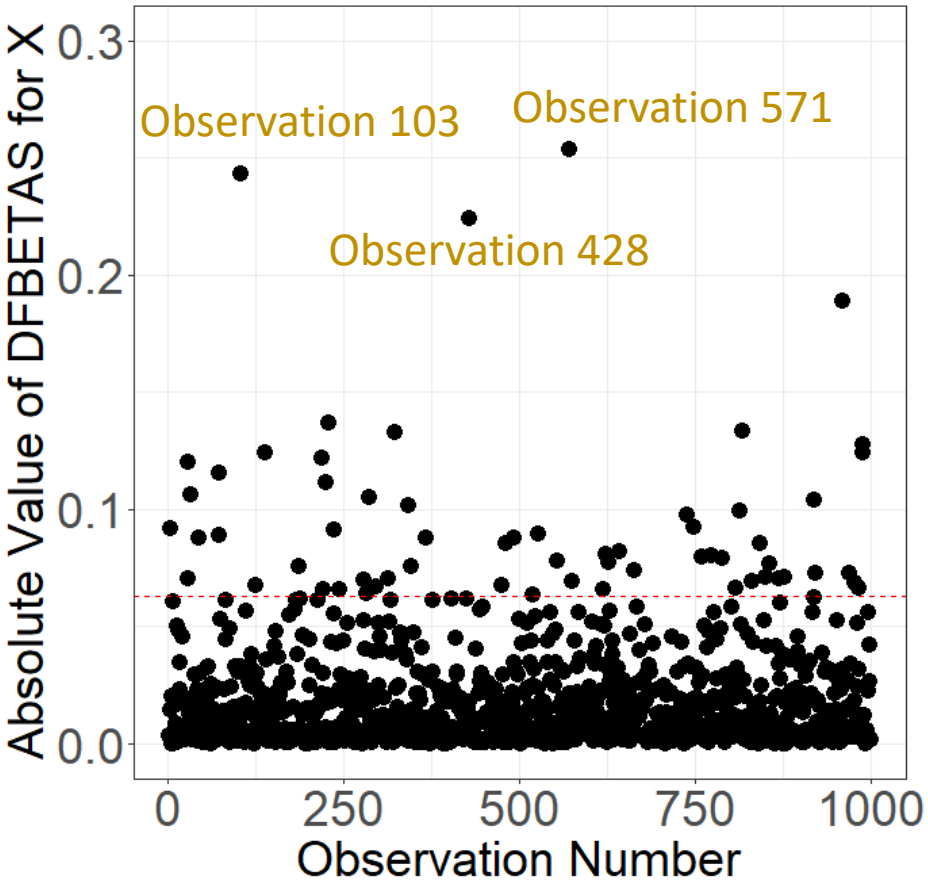
- $DFBETAS$: "DF" means "different", how different would the estimates of β be without an observation in the data set

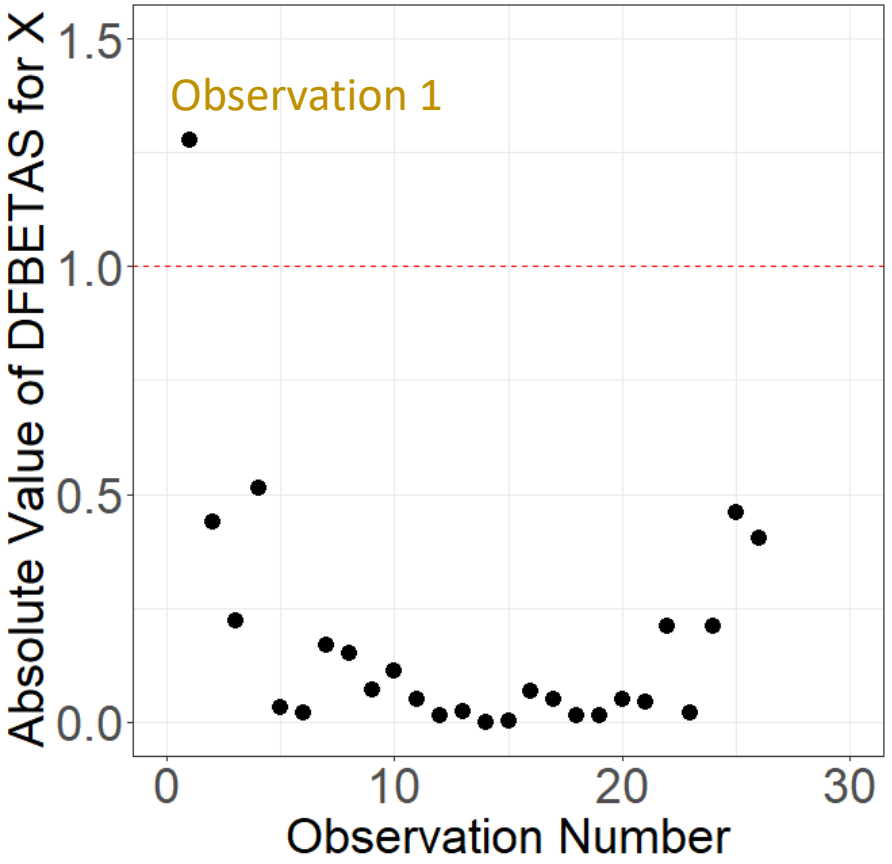
- If $DFBETAS_{(i)} > 0$, observation i pulls $\hat{\beta}$ up
- If $DFBETAS_{(i)} < 0$, observation i pulls $\hat{\beta}$ down

- How "large" to declare observation i influential on $\hat{\beta}$? *Rough* rule of thumb:
  - $\left|DFBETAS_{(i)}\right| > 1$ for $n \leq 30$
  - $\left|DFBETAS_{(i)}\right| > 2/\sqrt{n}$ for $n > 30$

- Common to plot the $DFBETAS$ against the observation number.

DFBETAS


Absolute Value of DFBETAS for X
0.3
0.2
0.1
0.0
Observation 103
Observation 571
Observation 428
0
250
500
750
1000
Observation Number


Absolute Value of DFBETAS for X
1.5
1.0
0.5
0.0
Observation 1
0
10
20
30
Observation Number

DFFITS

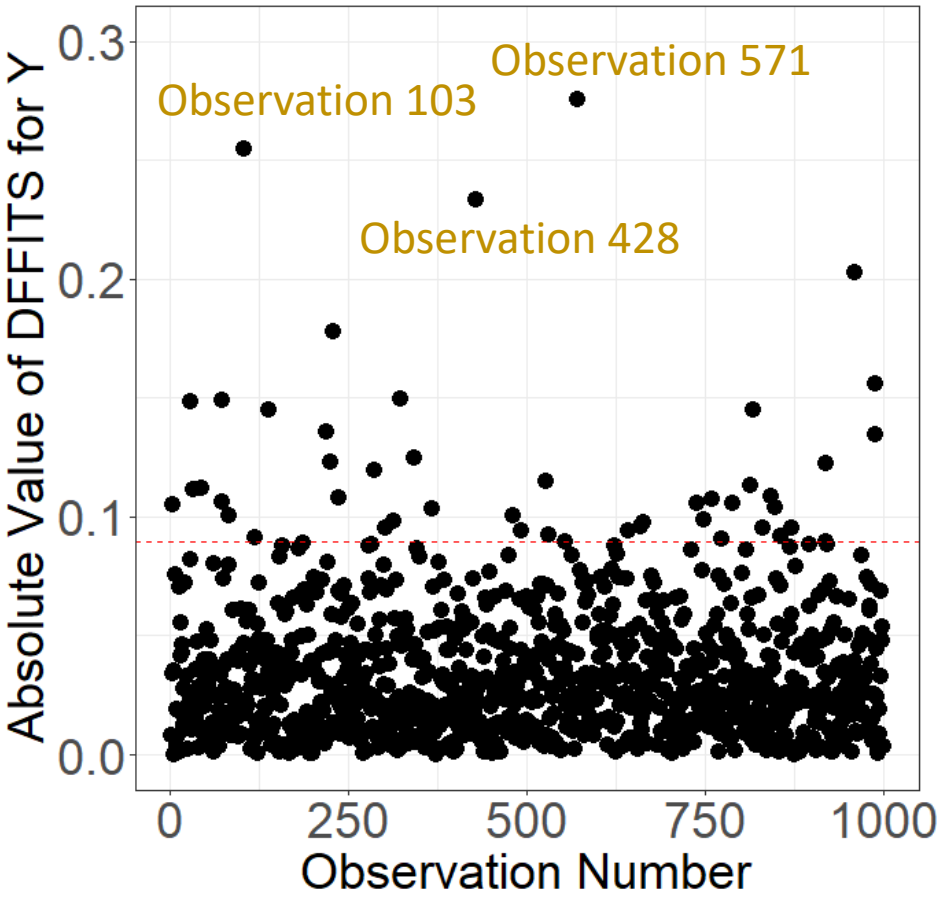
- $DFFITS$: How different would $\hat{y}_i$ be if observation i were not used to fit the model

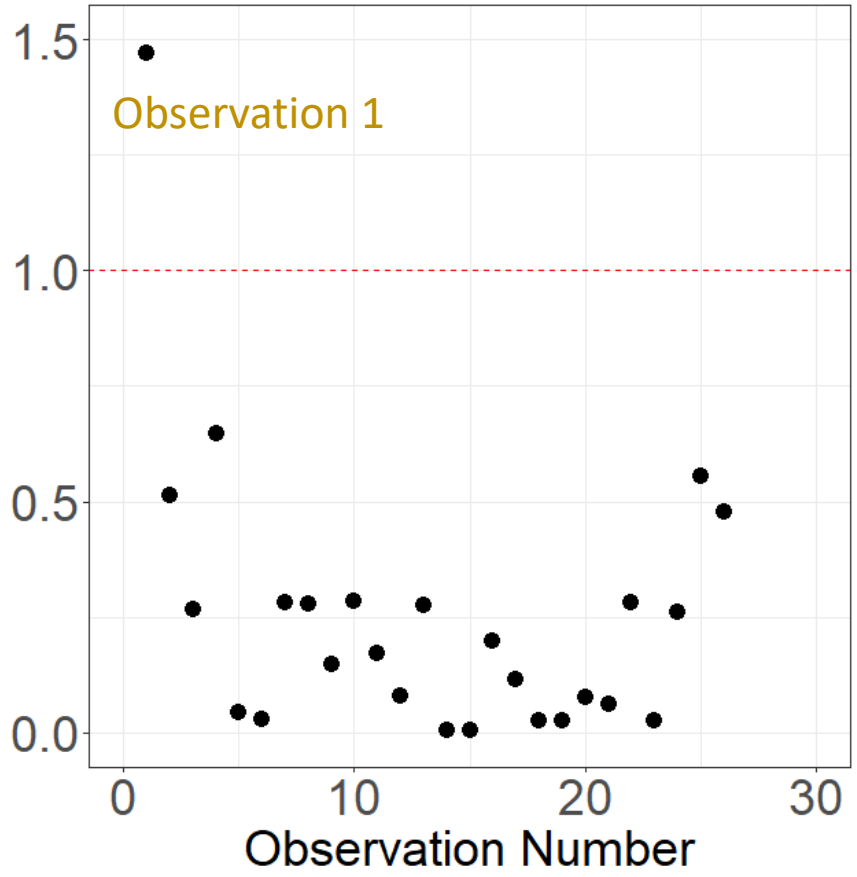
- How "large" to declare observation i influential on $\hat{y}_i$? *Rough* rule of thumb:
  - $|DFFITS_i| > 1$ for $n \leq 30$
  - $|DFFITS_i| > 2\sqrt{p/n}$ for $n > 30$

  where p represents the number of βs in the model (including the intercept).

- Common to plot the $DFFITS$ against the observation number.

DFFITS


Observation 571
Observation 103
Observation 428
Absolute Value of DFFITS for Y
0.3
0.2
0.1
0.0
0
250
500
750
1000
Observation Number


Observation 1
1.5
1.0
0.5
0.0
0
10
20
30
Observation Number

A – Model Describes All Observations Violations

- If the model does **NOT** describe *all* observations (i.e., there **ARE** influential points):
  - We will likely be misled when using model inference since estimates and standard errors could be changed greatly just by one data point
  - DO NOT SIMPLY THROW AWAY INFLUENTIAL POINT(S)!
  - Look into the influential point(s) to learn more about the data
  - Run the analysis with and without the influential point(s) to see how much that point(s) affected the analysis
  - Only remove the observation completely if it makes sense. You can then report the results of both models

R – Additional predictor variables are not required

R – Additional Predictors are Not Required Diagnostics

1. Think about if there are other variables that could explain the response well

- There are almost always other variables that could be useful…so, this assumption is almost always violated!

R – Additional Predictors are Not Required Violations

- If additional predictor variables **ARE** required:
  - We will likely be misled when using model inference since trends could be very different within subgroups
  - Be extremely careful when interpreting models based on observational data!
  - Trends observed are observational and NOT causal!

Diagnostics Summary

Relationships

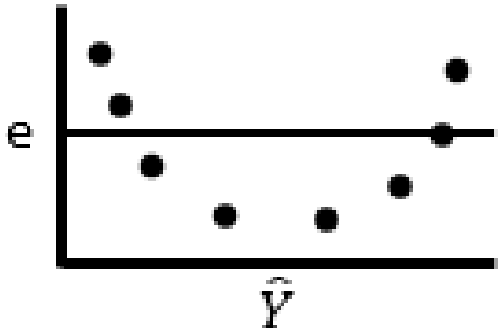
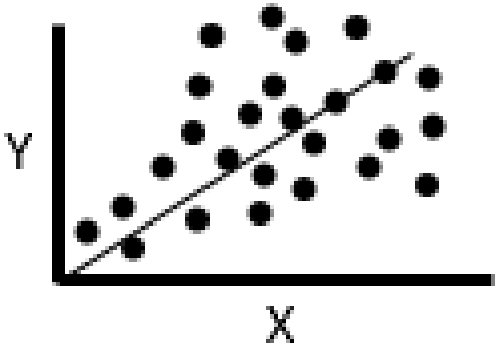
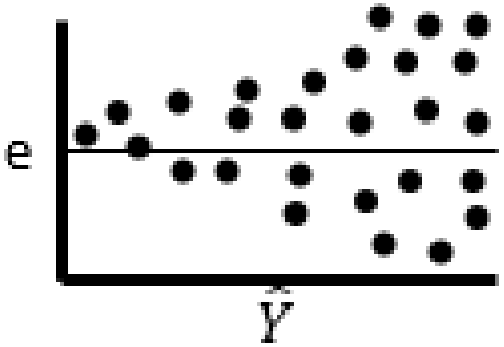
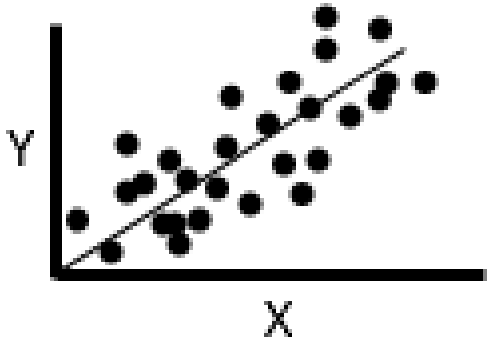
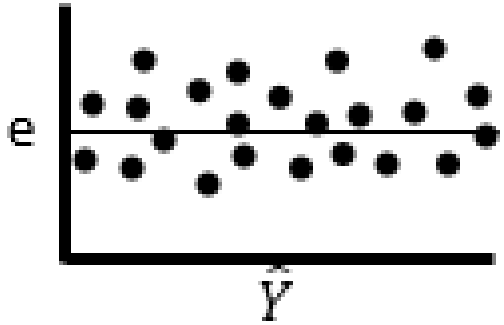
DISTRIBUTION (OF RESIDUALS)

Short-Tailed Left-Skewed Normal Right-Skewed Long-Tailed

HISTOGRAM (OF RESIDUALS)

NORMAL PROBABILITY PLOT (OF RESIDUALS)

Relationships

	Scatterplot	Residuals vs Fitted Values
Non-Linear		
Non-Constant Error Variance (Heteroskedasticity)		
Linear and Constant Variance		

Diagnostics for Each Assumption Summary

1. **L** – X vs Y is linear
   - [best] Scatterplot
   - Residuals vs. Fitted Values Plot
   - Residuals vs. Predictor Plot

1. **I** – The residuals are independent
   - [best] Think about how the data was collected
   - Sequence Plot *only if appropriate*

2. **N** – The residuals are normally distributed and centered at zero
   - Boxplot
   - Histogram
   - [best] Q-Q (Normal Probability) Plot
   - Shapiro-Wilk Test

4. **E** – The residuals have equal (constant) variance σ^2 across all values of X (homoscedastic)
   - Residuals vs. Fitted Values Plot

5. **A** – The model describes all observations (i.e., there are no influential points)
   - DFBETAS
   - DFFITS

6. **R** – Additional predictor variables are not required
   - Think about it

Transformation as a Remedial Measure

Transformations

- When the following assumptions are broken, they can often be remedied by transforming the data:
- If the (L) X vs Y is linear assumption is broken:
  - First, consider transforming X.
  - If X vs Y is still not linear, then consider transforming Y as well (leave the error terms alone, if possible, if linearity is the only assumption violated)
- If the (N) The residuals are normally distributed and centered at zero assumption is broken:
  - First, consider transforming Y.
  - If residuals are still not normally distributed, then consider transforming X as well
- If the (E) The residuals have equal (constant) variance σ^2 across all values of X (homoscedastic) assumption is broken:
  - First, consider transforming Y.
  - If residuals still do not have equal variance, then consider transforming X as well

Transformations (for Y)

- Right-skewed data is often more common than left-skewed data
- While our model requires the residuals to be normally distributed (NOT the response), it can be helpful to plot a histogram of the response with various transformations applied and choose the one that makes the response look the most normal. This often results in the residuals looking more normal, too.
- Box-Cox and Yeo-Johnson transformations can be a guide, but I haven't found reliable results in Python yet.

DISTRIBUTION (OF RESIDUALS)

Left-Skewed | Right-Skewed

HISTOGRAM

NORMAL PROBABILITY PLOT

TRANSFORMATIONS

$\sqrt{\text{constant} - Y}$
$\log(\text{constant} - Y)$
Y^2

$\sqrt{Y}$
$\log(Y)$
$1/\sqrt{Y}$
$1/Y$

Transformations

- Usually, it is best to try multiple transformations (perhaps on just X, just Y, and then on both X and Y) to find the transformation that yields the model that best meets the assumptions.
- For example, you may start with this model that does not meet the assumptions

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \text{ where } \epsilon_i \sim N(0, \sigma^2)$$

- Then, you may try the following transformations (along with others) before deciding which (if any) model is the best

$$y_i = \beta_0 + \beta_1 \log(x_i) + \epsilon_i, \text{ where } \epsilon_i \sim N(0, \sigma^2)$$

$$\log(y_i) = \beta_0 + \beta_1 x_i + \epsilon_i \text{ , where } \epsilon_i \sim N(0, \sigma^2)$$

$$\log(y_i) = \beta_0 + \beta_1 \log(x_i) + \epsilon_i \text{ , where } \epsilon_i \sim N(0, \sigma^2)$$

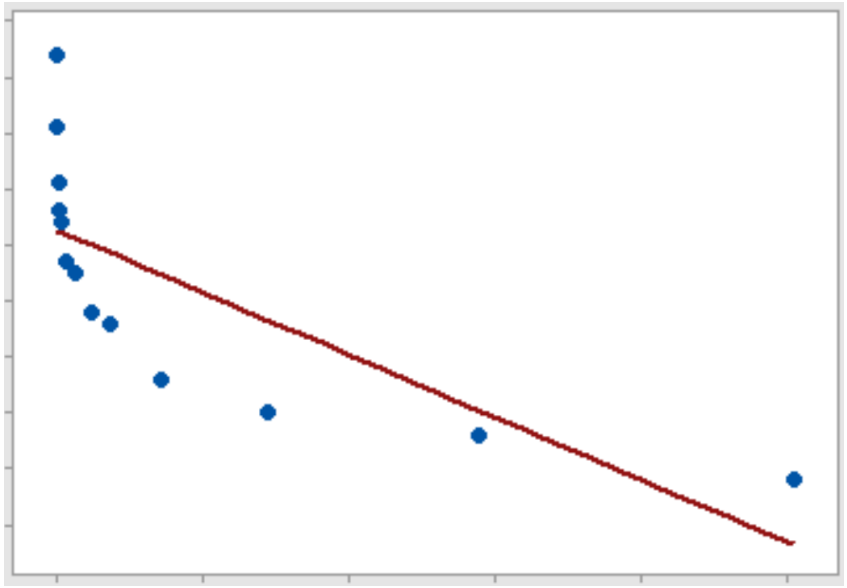
$$y_i = \beta_0 + \beta_1 \sqrt{x_i} + \epsilon_i, \text{ where } \epsilon_i \sim N(0, \sigma^2)$$

$$\sqrt{y_i} = \beta_0 + \beta_1 \sqrt{x_i} + \epsilon_i, \text{ where } \epsilon_i \sim N(0, \sigma^2)$$

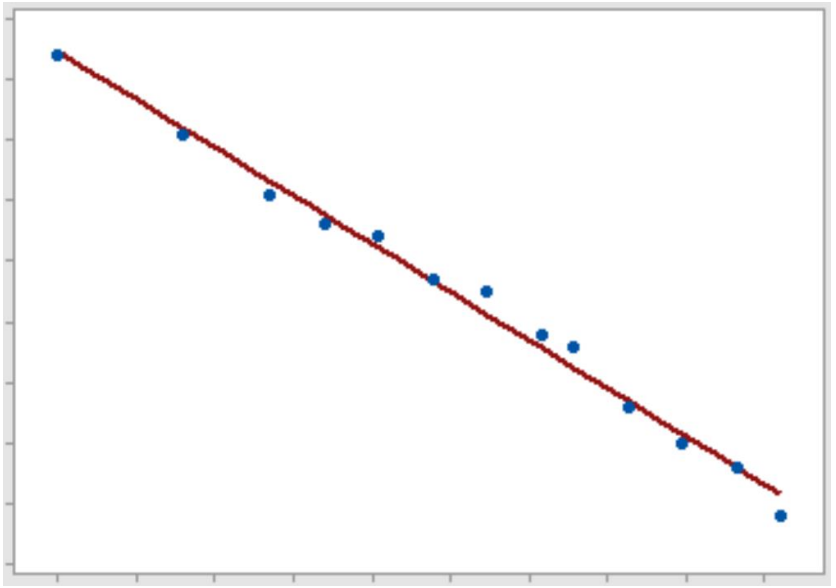
Example

Figure taken from: https://online.stat.psu.edu/stat501/book/export/html/956

Scatterplot (before)



Scatterplot (after)



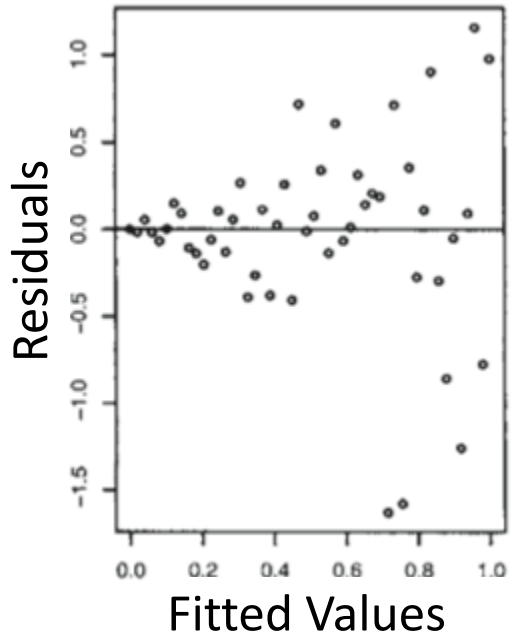
- Which assumption is the data violating?
- What should I do to fix it?

- Label the axes.
- Is this assumption now met?
- Note that you would now need to check the other assumptions again.

Example

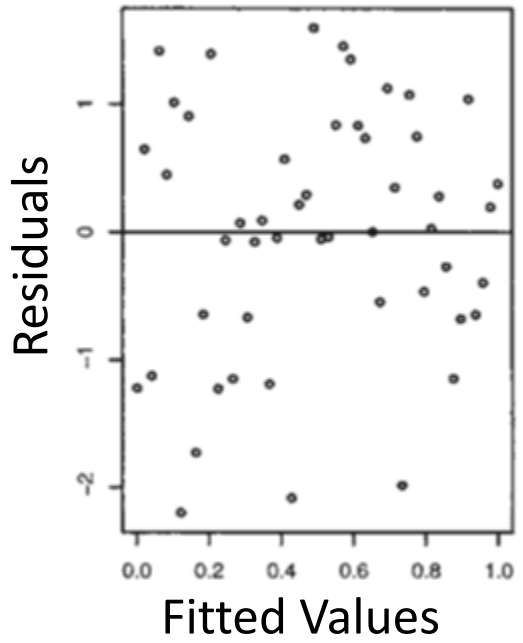
Figure taken from https://stats.stackexchange.com/questions/76226/interpreting-the-residuals-vs-fitted-values-plot-for-verifying-the-assumptions (originally from Faraway's Linear Models with R (2005, p. 59))

Residuals vs. Fitted Values (before)



- Which assumption is the data violating?
- What should I do to fix it?

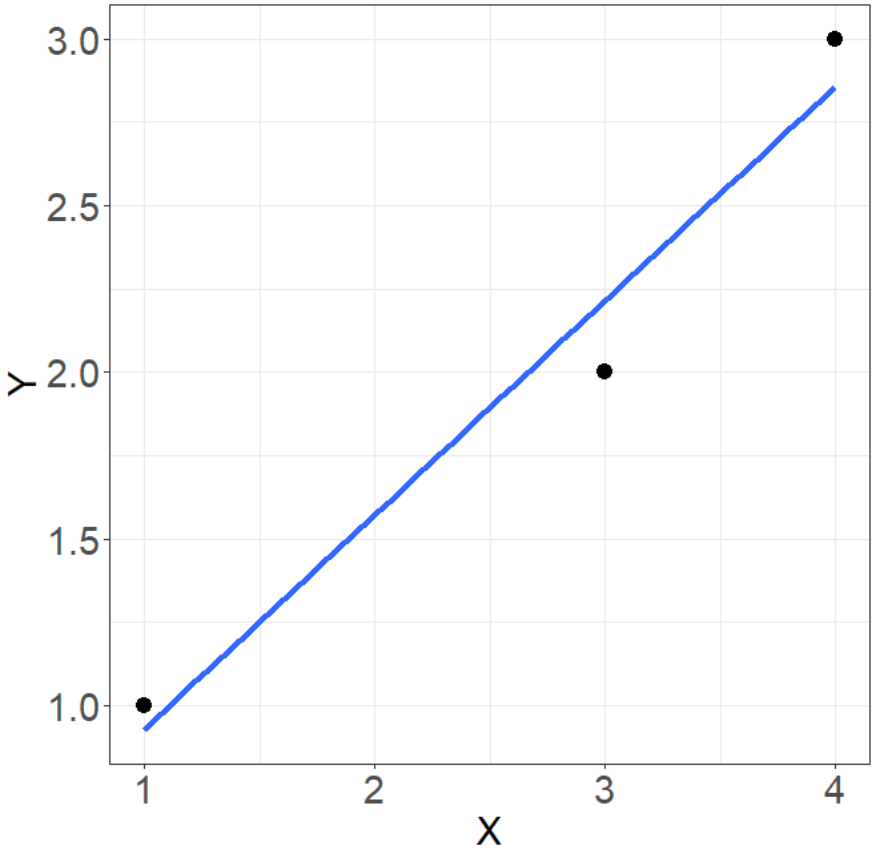
Residuals vs. Fitted Values (after)

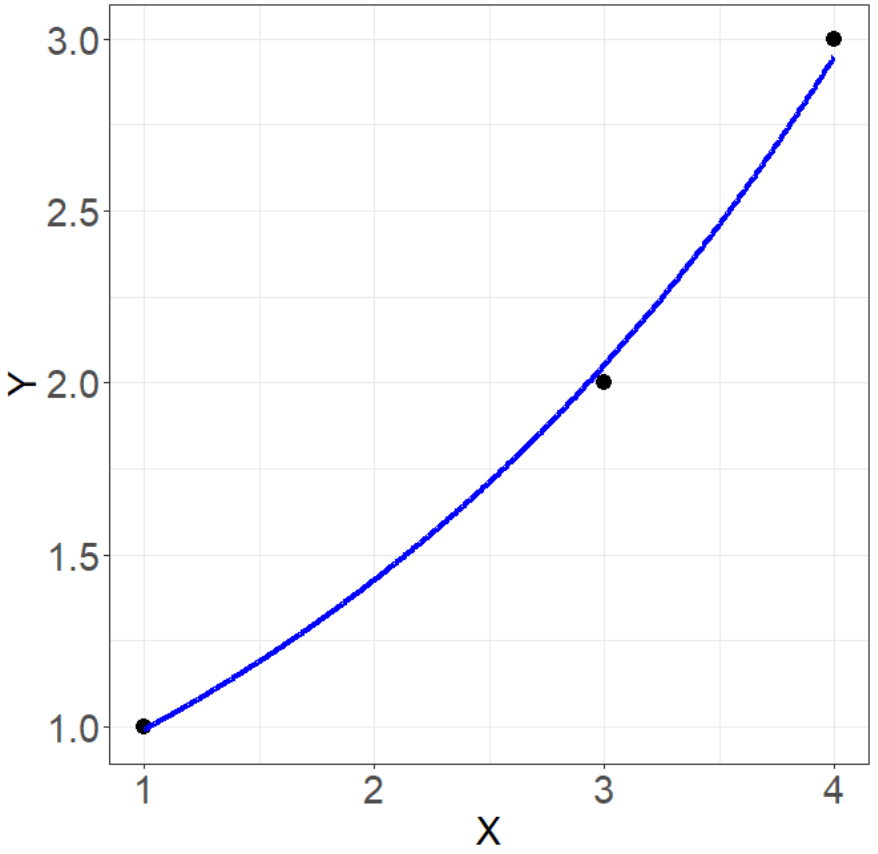


- Is this assumption now met?
- Note that you would now need to check the other assumptions again.

Example

X	Y
1	1
3	2
4	3

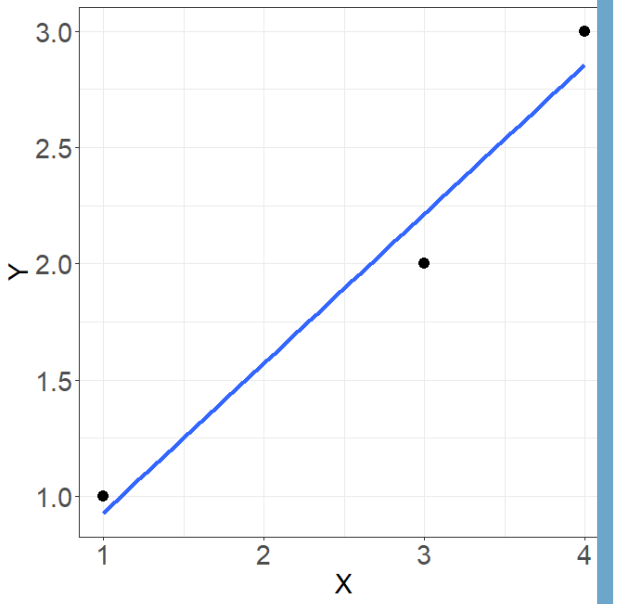




Transformations: The Process

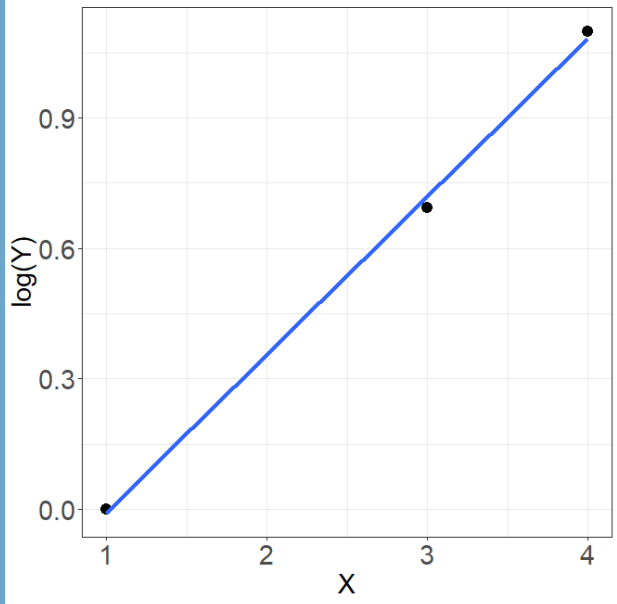
Linear Regression using the model

$\hat{Y}_i = 0.3 + 0.6X_i$



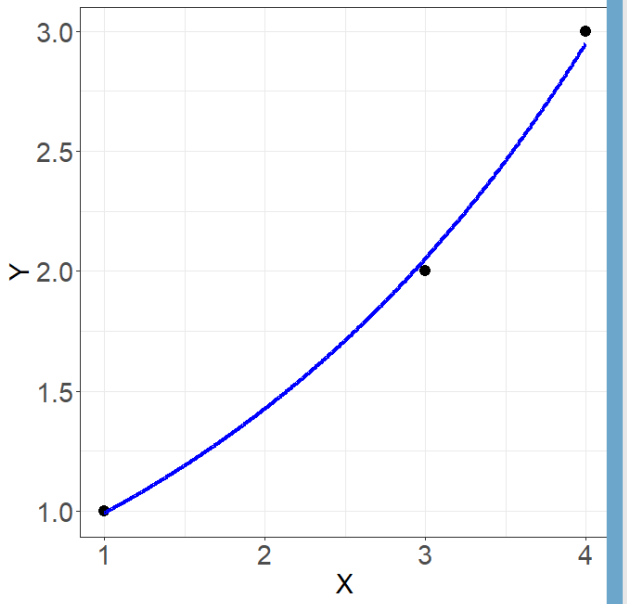
Linear Regression using the model

$\log(\hat{Y}_i) = -0.4 + 0.4X_i$



Linear Regression using the model

$\hat{Y}_i = \exp(-0.4 + 0.4X_i)$



SAME MODEL!

Slope Coefficient Interpretations after Transforming

Transformations & Interpretations Interpret

- For this fitted model

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

- Solve for y

$$\begin{aligned}\hat{y}_i &= \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right) \\ &= \exp\left(\hat{\beta}_0\right)\exp\left(\hat{\beta}_1 x_i\right)\end{aligned}$$

- So, X has a *multiplicative* relationship with Y.
  - Ex: $\hat{\beta}_1 = 0.45$, $\exp\left(\hat{\beta}_1\right) = 1.57$, which means that the average of Y is multiplied by 1.57 for every one unit increase in X.
- To obtain a percent change, calculate

$$\left(\exp\left(\hat{\beta}_1\right) - 1\right) \times 100\%$$

  - Ex: $(1.57 - 1) \times 100\% = 57\%$. A more straightforward interpretation than the previous is: "The average of Y increases by 57% for every one unit increase in X."

Transformations & Interpretations Interpret

- For this fitted model

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

- X still has an $additive$ relationship with Y, but we should interpret that relationship in terms of a percent increase/decrease.
- To do this, calculate $\frac{\hat{\beta}_1}{100}$.
  - Ex: $\hat{\beta}_1 = -0.27$, $\frac{-0.27}{100} = -0.0027$, with the interpretation, "The average of Y decreases by 0.0027 with every 1% increase in X."

Transformations & Interpretations Interpret

- For this fitted model

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

- Start by calculating $\hat{\beta}_1 \times 100\%$.
  - Ex: $\hat{\beta}_1 = -0.3$, $-0.3 \times 100\% = -30\%$ with the interpretation, "The average of Y decreases by 30% with every 1% increase in X."

Transformations & Interpretations Interpret

- For your homework, if you decided on a log transformation, which is quite common, I expect you to be able to interpret the slope like I did in the previous examples.

- If the transformation you choose is *not* the log transformation, it is sufficient to report the coefficient on the transformed scale.
  - Ex: your final fitted model is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1\sqrt{x_i}$, where $\hat{\beta}_1 = 2.1$, so your interpretation would be, "The average of Y increases by 2.1 as $\sqrt{X}$ increases by 1."
  - Ex: your final fitted model is $\frac{1}{\hat{y}_i} = \hat{\beta}_0 + \hat{\beta}_1 x_i$, where $\hat{\beta}_1 = 2.1$, so your interpretation would be, "The average of $\frac{1}{Y}$ increases by 2.1 as X increases by 1."

Parsimony

The key is to get interpretable transformations where the assumptions are met while keeping the model as simple as possible.

Non-Slope Interpretations after Transforming

Transformations & Interpretations Interpret

- Note that to get **fitted values/estimates** for Y or **confidence intervals**, you can simply back-transform.

- For example, for the model:

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

we would get the estimate of Y by exponentiating:

$$\hat{y}_i = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)$$