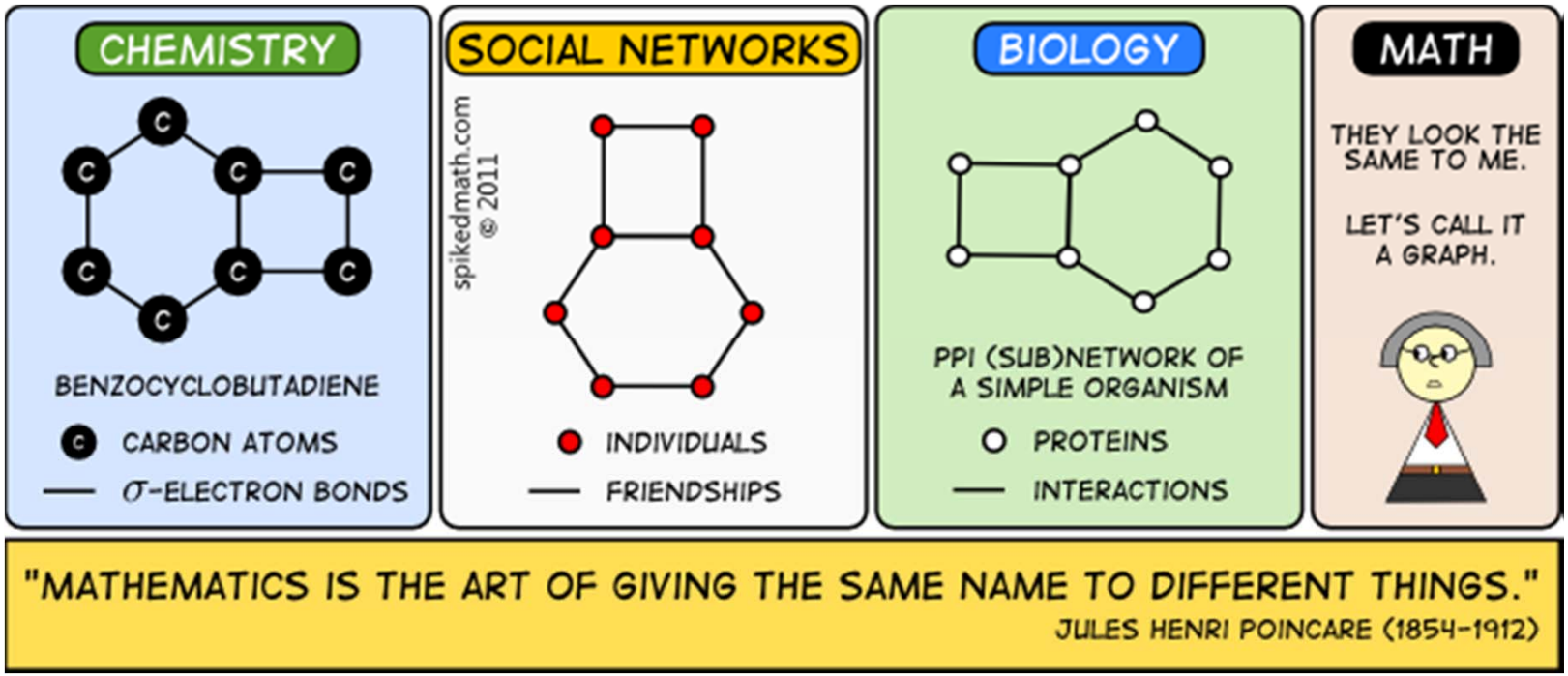
UNIVERSITY of WASHINGTON

Data Science UW
Methods for Data Analysis

CLT and Intro to Regression
Lecture 5
Steve Elston


CHEMISTRY
C
BENZOCYCLOBUTADIENE
CARBON ATOMS
σ-ELECTRON BONDS
SOCIAL NETWORKS
spikedmath.com
© 2011
INDIVIDUALS
FRIENDSHIPS
BIOLOGY
PPI (SUB)NETWORK OF
A SIMPLE ORGANISM
PROTEINS
INTERACTIONS
MATH
THEY LOOK THE SAME TO ME.
LET'S CALL IT A GRAPH.
"MATHEMATICS IS THE ART OF GIVING THE SAME NAME TO DIFFERENT THINGS."
JULES HENRI POINCARE (1854-1912)

Topics

- Review
- Central Limit Theorem
- Linear Regression
- More on presenting data science results

Review

- Outliers
- Group Testing - ANOVA
- Resampling
  - Bootstrap
  - Jackknife
  - Cross Validation
- Presenting Data Science Results
  - State clear conclusion
  - Support with evidence
  - Simplify!

Central Limit Theorem

- Sample a population many times, the distribution of means of all samples are normally distributed, regardless of the population distribution.

$\bar{X}$=sample mean.

$$\bar{X} \sim N(mean, \frac{st.dev}{\sqrt{n}})$$

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

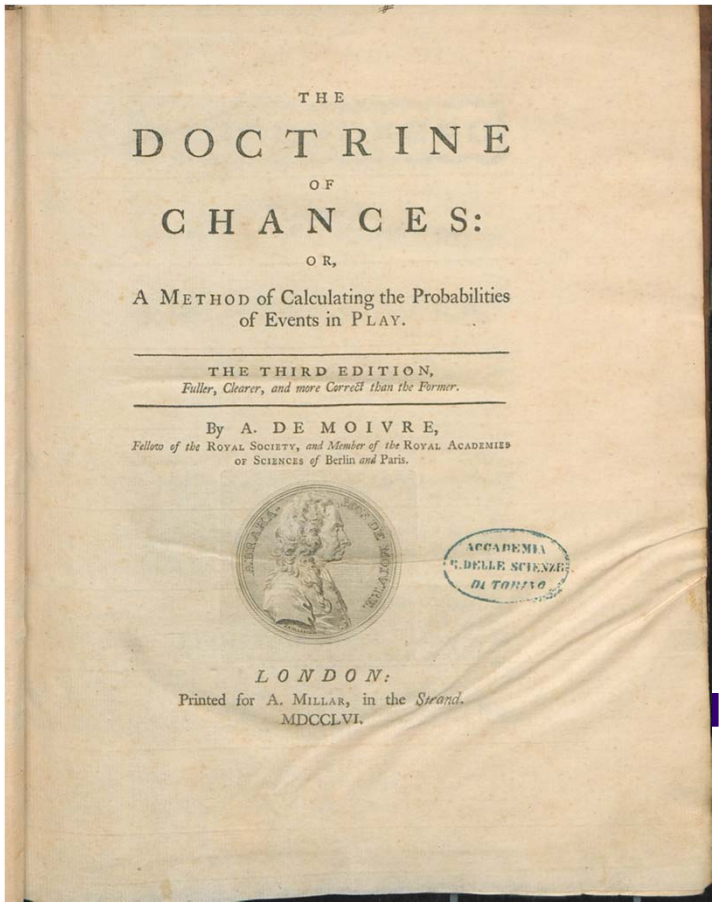
Why is the CLT so important?

CLT is foundational!

- CLT enables sampling methods
- Without a CLT we could not reliably compute confidence intervals
- Most statistical methods and machine learning algorithms rely on CLT
- e.g. Hypothesis tests rest on the CLT

Central Limit Theorem

- de Moivre, 1738 – proof of special case for Bernoulli trials
- Laplace 1776, 1785, 1820
- Chebyshev, 1887 – rigorous proof



THE

DOCTRINE

OF

CHANCES:

OR,

A METHOD of Calculating the Probabilities of Events in PLAY.

THE THIRD EDITION,
Fuller, Clearer, and more Correct than the Former.

By A. DE MOIVRE,
Fellow of the ROYAL SOCIETY, *and Member of the* ROYAL ACADEMIES OF SCIENCES *of* Berlin *and* Paris.

LONDON:
Printed for A. MILLAR, in the *Strand*.
MDCCLVI.

Central Limit Theorem

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

- We can use this central limit theorem to generate confidence intervals on expressing the population mean.
- If we know the sample mean, sample variance, and number of samples:
  - Then we know how our estimate of the population mean is distributed (from above formula).
  - We can then generate 90%, 95%, … confidence intervals around our sample mean.

Confidence Intervals

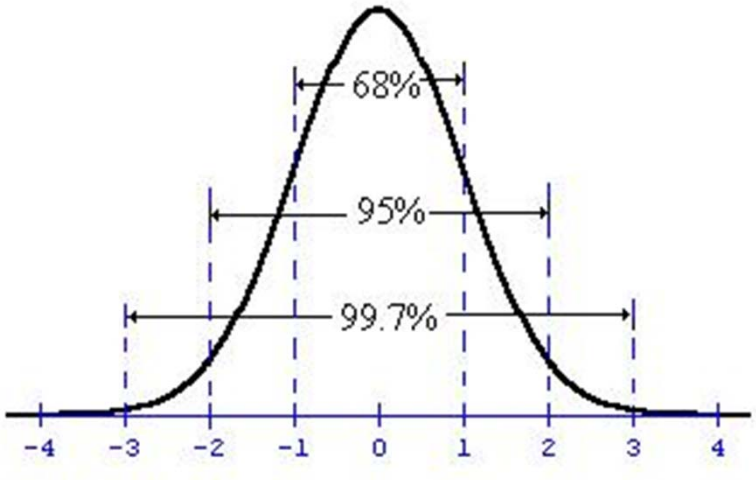
- Confidence intervals are a way to express uncertainty in *population* parameters, as estimated by the sample.
- E.g. If we create a 95% confidence interval for the population mean, say $\hat{\mu} = \bar{X} = 10 \pm 5$
  - Then we say that the true population mean, μ, has a 95% chance of being between 5 and 15.

It is **not** correct to say:

- ~~"95% of the sample values are in this range."~~
- ~~"There is a 95% chance that the mean of another sample will be in this range."~~

Confidence Intervals

- To create confidence intervals for population means, we use the central limit theorem and create confidence intervals based on the normal distribution.
  - Repeatedly sample from the population.
  - Calculate the mean for each sample.
  - Use the average of the sample means as the population estimate and create a C.I. based on the s.d. of the sample means.
  - R demo



Regression Models

- The goal of regression is to produce a model that represents the ‘best fit’ to some observed data.
- Typically the model is a function describing some type of curve (lines, parabolas, etc.) that is determined by a set of parameters (e.g., slope and intercept).
- “Best fit” means that there is an optimal set of parameters according to an evaluation criteria we choose.

Why Focus on Regression Models

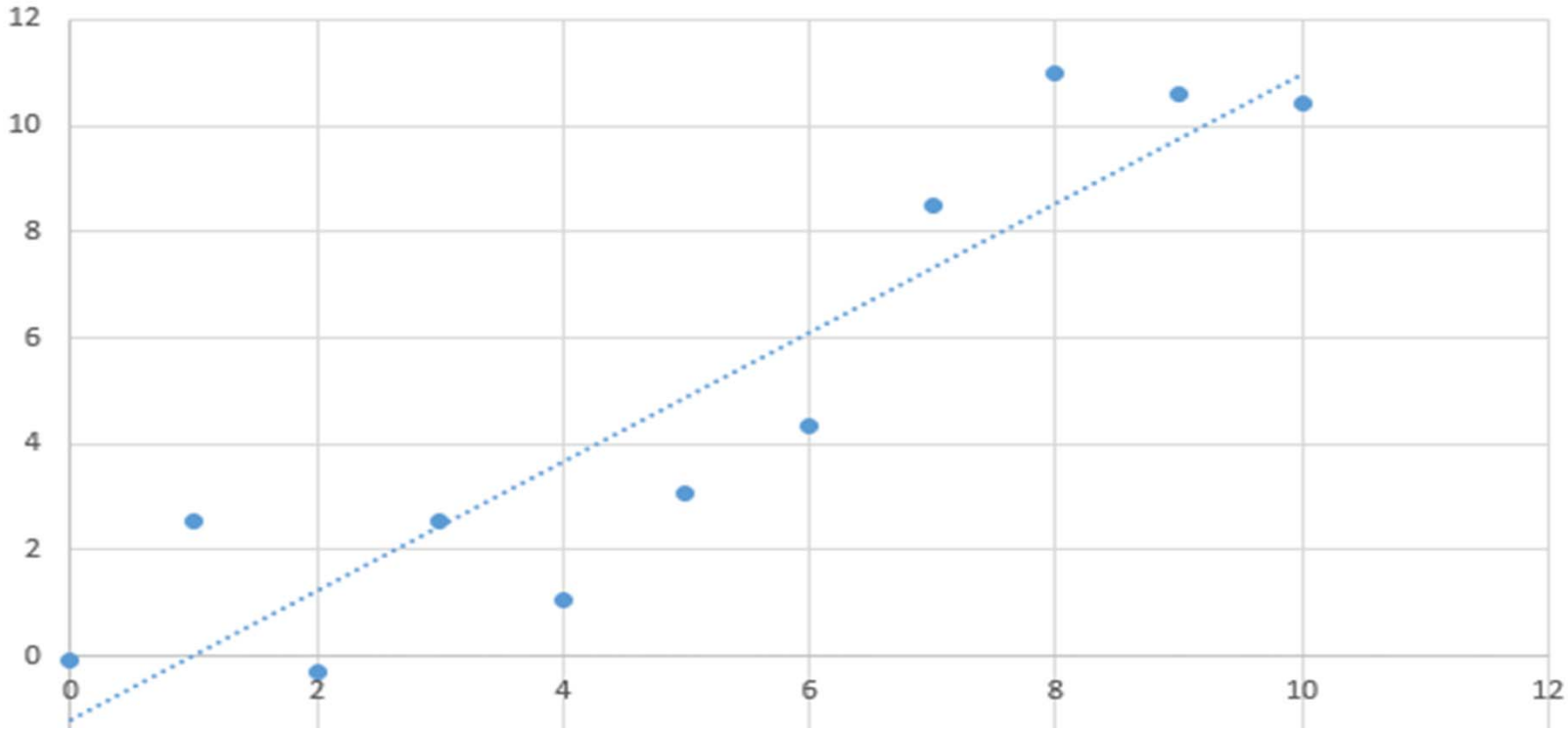
Linear models are foundational

- Regression is a linear model
- Linear models derived with linear algebra
- Basis of many machine learning models
- Understanding linear models is basis for understanding behavior of stats and ML models
- Basis of time series models

History

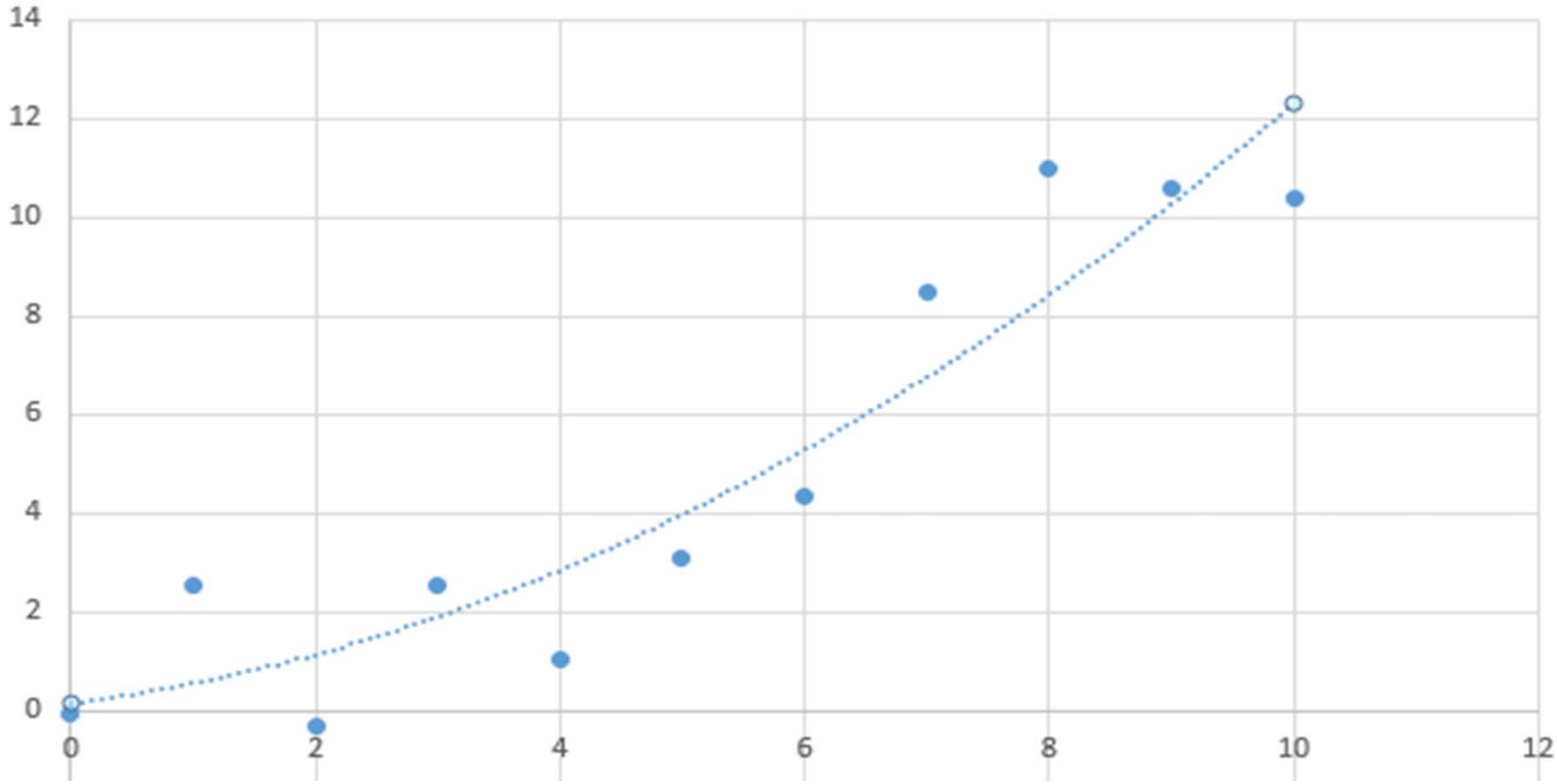
- Least squares is an old idea
  - Tobias Mayer, 1750, Laplace 1788: method of averages
  - Carl Fredrich Gauss, 1809: least squares
- Long history of regression in statistical theory
  - Galton 1894
  - Pearson 1898
  - Joseph Berkson, 1944: logistic regression
  - Many others over the decades

Regression: Linear



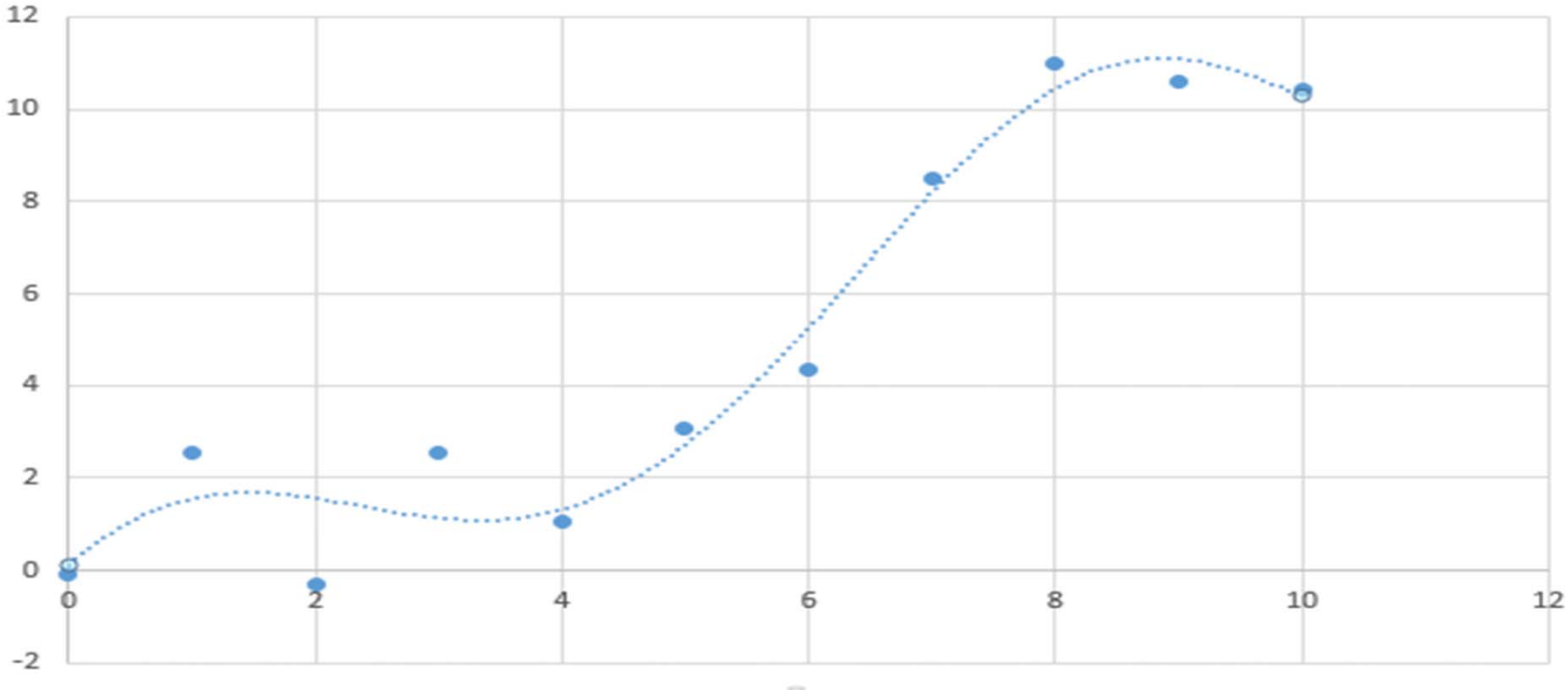
Regression: quadratic

Still linear in coefficients



Regression: High Order

Still linear in coefficients



Regression Models

- Which one of the preceding examples is correct?
- In a sense, all of them are. They all give decent approximations to the data.
- It's hard to tell, just from looking at these plots whether any of them in fact will continue to perform well as more data is received.

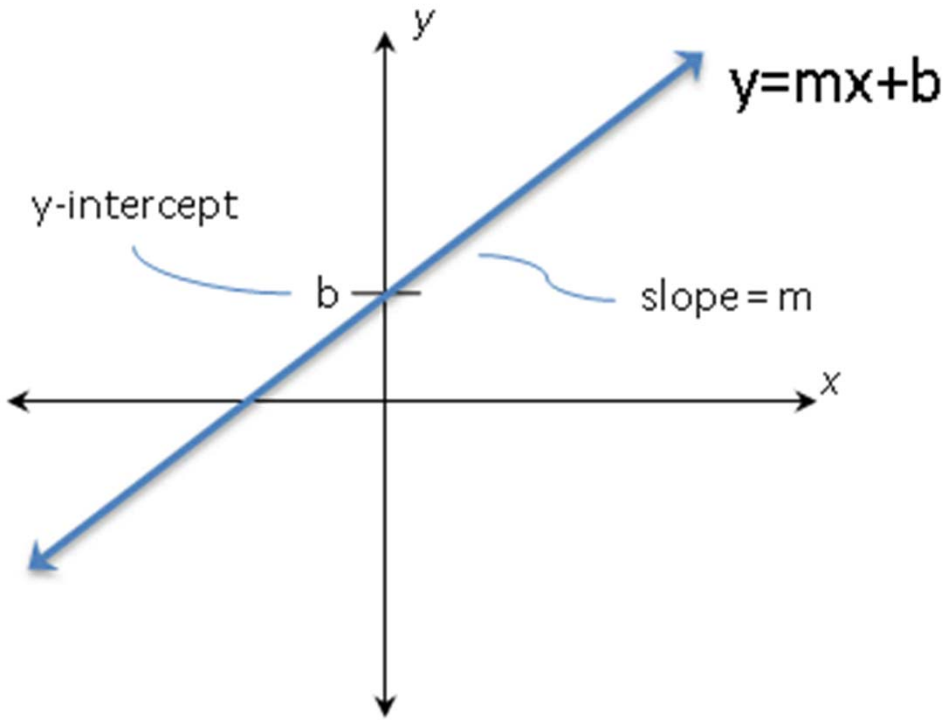
- We don't know if these models will generalize
  - A model which generalizes gives reasonable answers for input values not used for training

Linear Regression

- **Response** (**Dependent**) variable (aka the **Label** in machine learning): the variable of primary interest in a study- the one you are trying to predict or explain.
- **Explanatory** (**Independent**) variables (aka the **Features** in machine learning): the variables that attempt to explain the observed outcomes of the response variable.
- There are two types of parameters in linear models:
  - The intercept (y-intercept).
  - The slope, rise over run, or change in Y divided by change in X

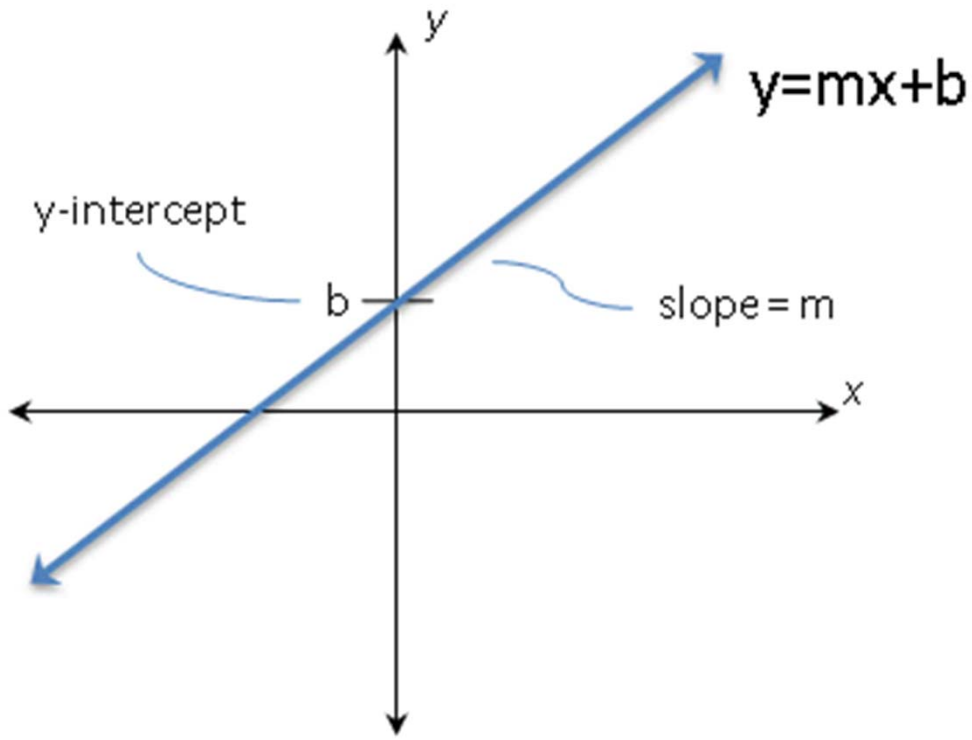
Linear Regression

- When x = 0, then y = b.
- When x = -(b/m) then y = 0.



Linear Regression

- Interpret slope: $m = \frac{rise}{run} = \frac{\Delta y}{\Delta x}$
  - If x changes by Δx, then y must change by Δy in order for the slope to stay the same (and it must).

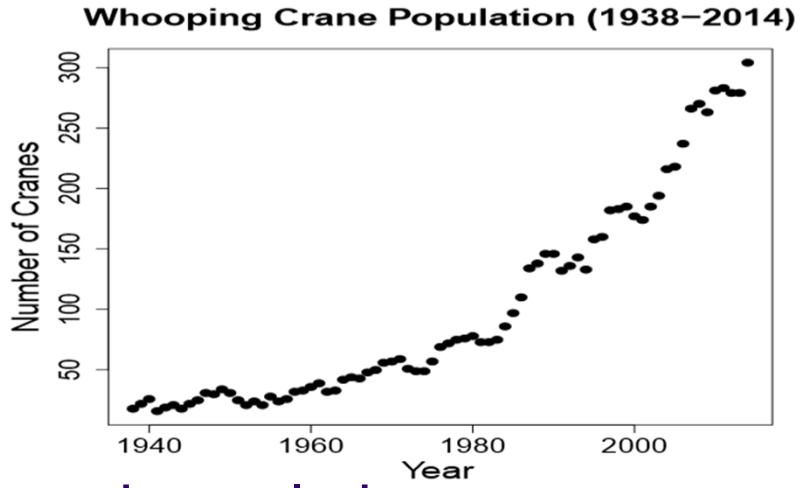


Given two points, (x1,y1), (x2,y2)

$$m = \frac{(y2 - y1)}{(x2 - x1)} = \frac{(y1 - y2)}{(x1 - x2)}$$

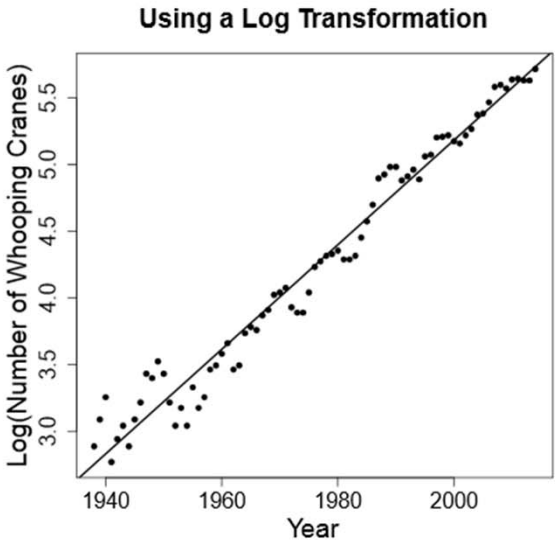
Linear Regression

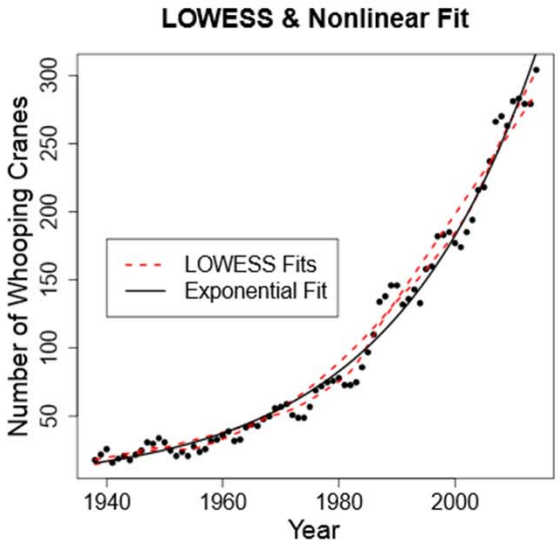
- Consider the relationship between Whooping Crane population and year below.



- Possible regression solutions:
  - Transform the response variable, to linearize the relationship.
  - Fit a higher-order model
  - Fit a nonlinear regression model to the data.

Linear Regression





> How would we decide on a ‘best’ model?

Linear Regression

- We use the method of least squares to find the best fit line: $y_i = mx_i + b + \varepsilon_i$

$$\min_{m,b} \sum_{i=1}^{n} (\varepsilon_i)^2 = \min_{a,b} \sum_{i=1}^{n} (y_i - (mx_i + b))^2$$

- Explicit solutions exist (using calculus).
- Computers are really good at finding minimums of equations. We let them do this for us.

Linear Regression

- The method of least squares finds the best fit line.
  - The mean of the errors from the best fit line is zero.
  - This means there is no ‘bias’ in our prediction.
- Three key quantities
  - SST = sum of squares total
  - SSE = sum of squares error
  - SSR = sum of squares residual

  where

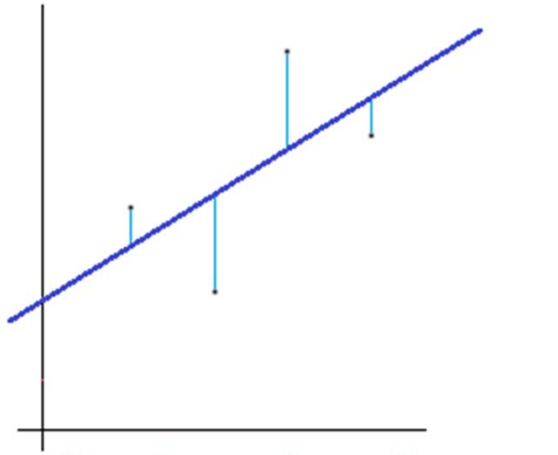
  Residual = observed value – predicted value of dependent variable

  and

  SST = SSE + SSR

Linear Regression

- Linear regression is the most common.
  - Fit a line (2D), plane (3D), or a hyperplane to the observed data.
- We need to define an error metric for a line through points.

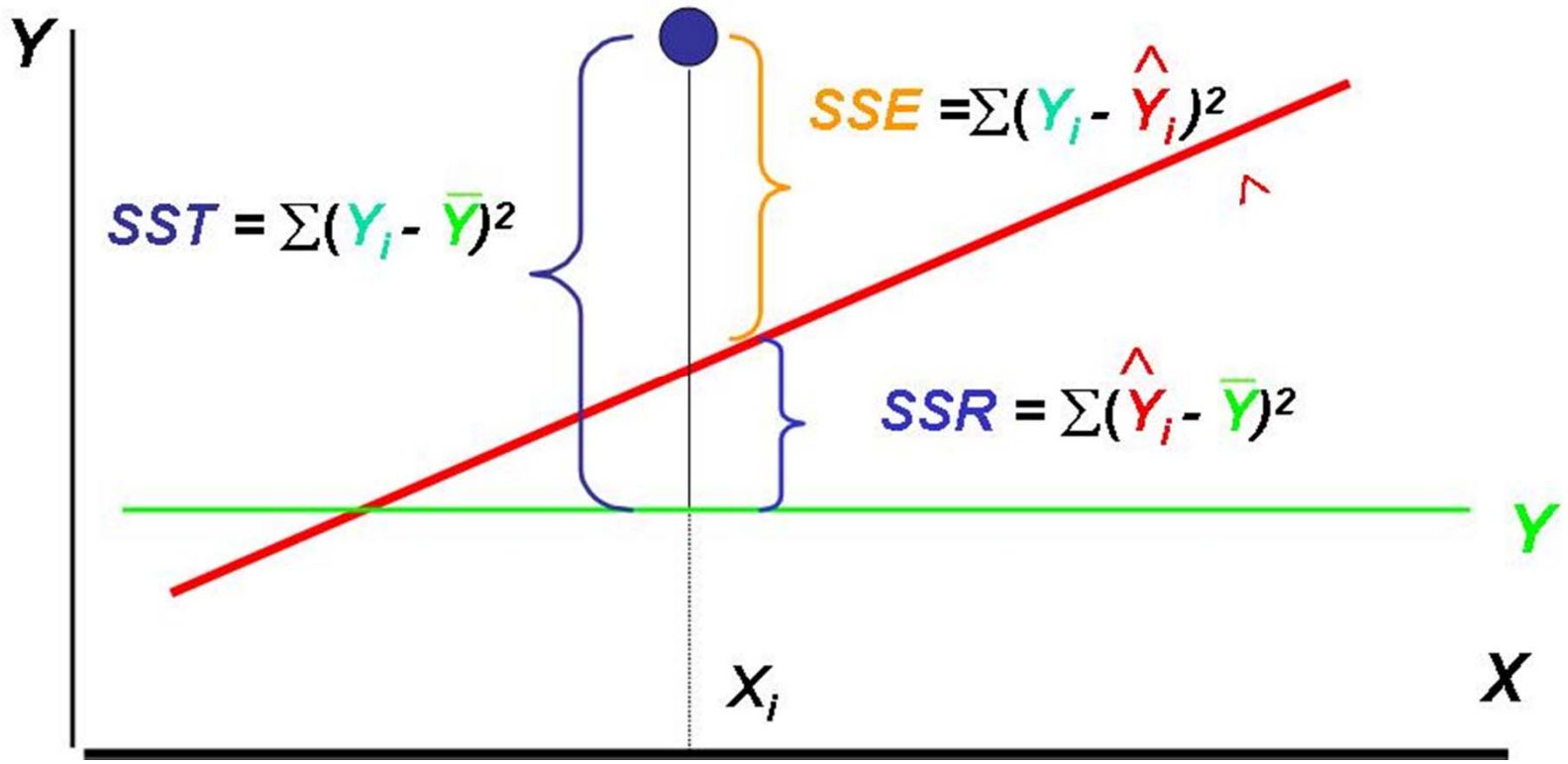


Least Squares Linear Regression

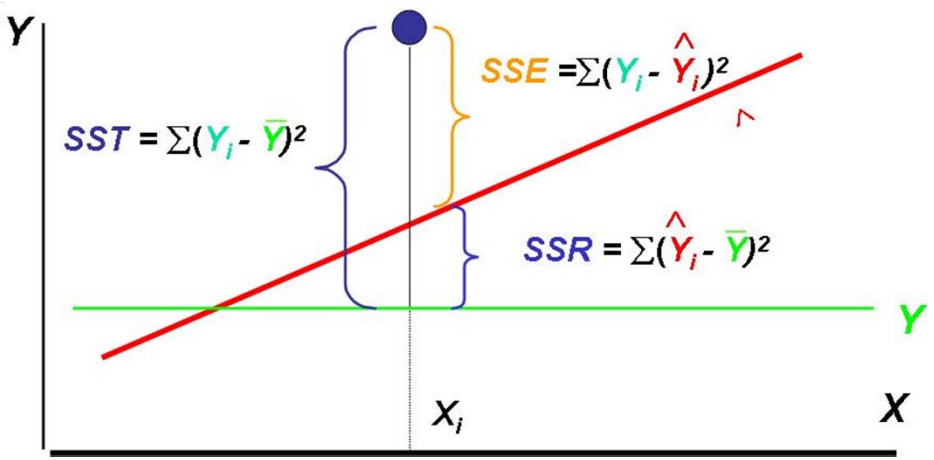
- We use the sum of the squared error between the predicted and actual.

Linear Regression

> With modeling, we are interested in the SSE, SSR, SST.



Linear Regression



- R-squared is called the coefficient of determination.
- R-squared measures how well the data fits a specified model.
- For linear models, we define this as:

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}$$

Linear Regression

- We can also measure accuracy of the line using Root Mean Squared Error (RMSE).
  - Using this as an estimate of the error means we are losing one more degree of freedom than the standard deviation, so we write the RMSE as

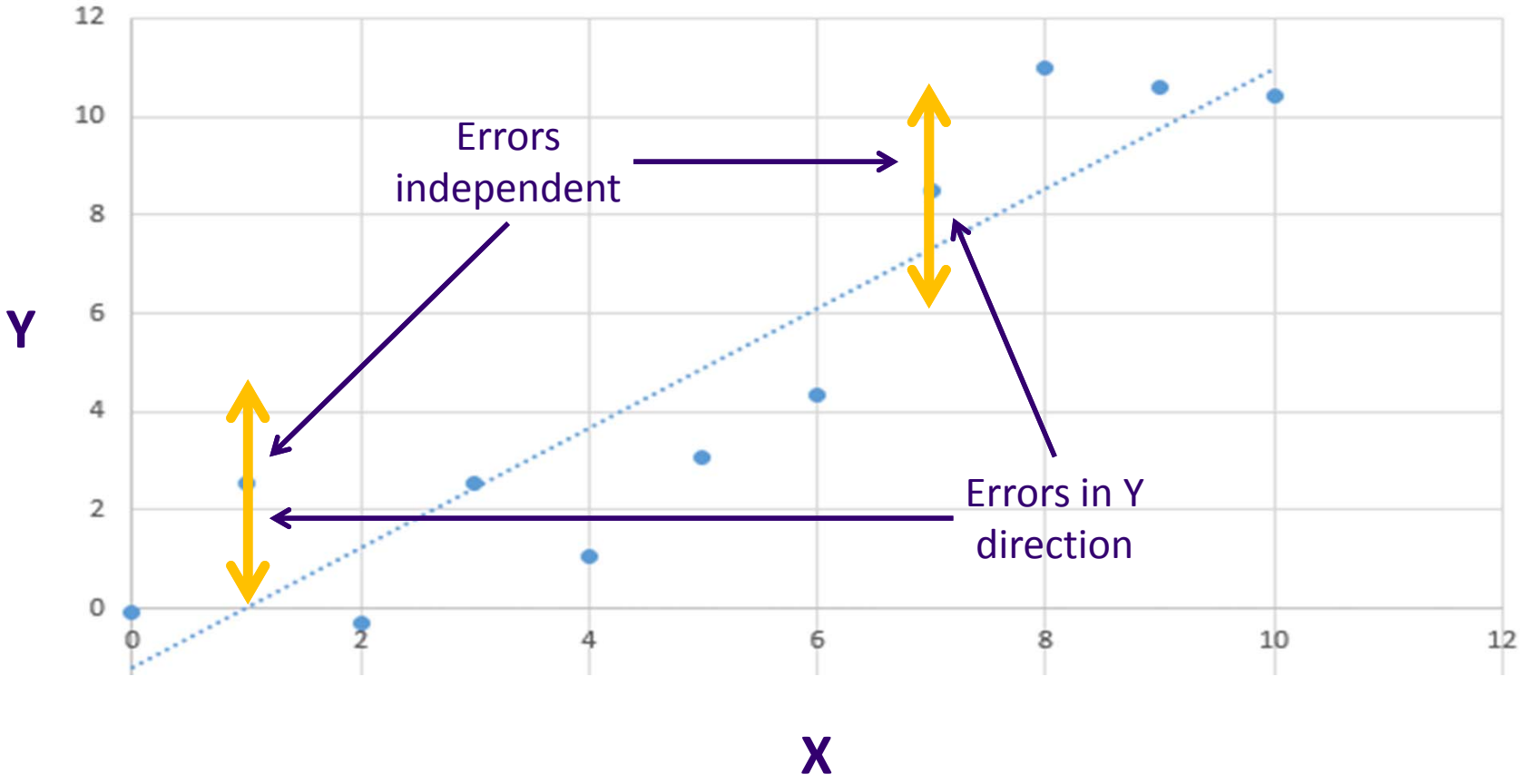
$$RMSE = \frac{SSE}{n-2}$$

Linear Regression Assumptions

- Linear relationship between dependent variable and independent variables.
- Measurement error is independent and random.
- Errors arise from the dependent variable
- No multicolinearity, e.g. no significant correlation between independent variables
- Residuals are homoscedastic (constant variance). I.e, the errors are the same across all groups of independent variables.

iid = independent identical distribution

Linear Regression Errors



Homoscedasticity

- Our assumed model:

$$y_i = mx_i + b + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma)$$

  - Know that m,b, σ are all fixed parameters in that model.
  - ε_i is iid

- Compare with:

$$y_i = mx_i + b + \epsilon_i$$

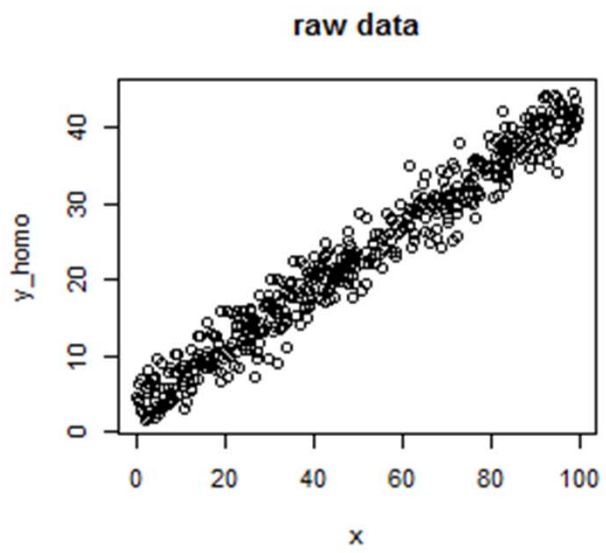
$$\epsilon_i \sim N(0, f(x_i))$$

e.g.

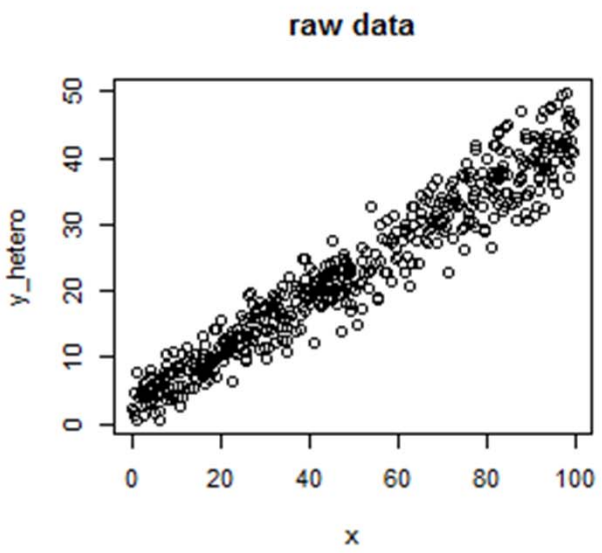
$$\epsilon_i \sim N(0, e^{x_i})$$

Homoscedasticity

homoscedastic



heteroscedastic



Interpreting R's Output

```
x = 1:20
y = x + rnorm(20)
best_fit = lm(y~x)
summary(best_fit)
```

```
Call: lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-2.04153 -0.43442 -0.01455 0.73806 1.93583

Coefficients:
            Estimate      Std.    Error t value Pr(>|t|)
(Intercept) -0.23738   0.53628   -0.443   0.663
x            1.02521   0.04477   22.901   9.19e-15 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.154 on 18 degrees of freedom
Multiple R-squared: 0.9668,      Adjusted R-squared: 0.965
F-statistic: 524.4 on 1 and 18 DF,   p-value: 9.186e-15
```

Distribution of residuals:
Residuals = (actual – pred)

Interpreting R's Output

```
x = 1:20
y = x + rnorm(20)
best_fit = lm(y~x)
summary(best_fit)
```

```
Call: lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-2.04153 -0.43442 -0.01455 0.73806 1.93583

Coefficients:
            Estimate   Std. Error  t value   Pr(>|t|)
(Intercept) -0.23738      0.53628   -0.443      0.663
x            1.02521      0.04477   22.901   9.19e-15 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.154 on 18 degrees of freedom
Multiple R-squared: 0.9668,      Adjusted R-squared: 0.965
F-statistic: 524.4 on 1 and 18 DF,   p-value: 9.186e-15
```

Least Squares estimates of coefficients

Standard Error of coefficients

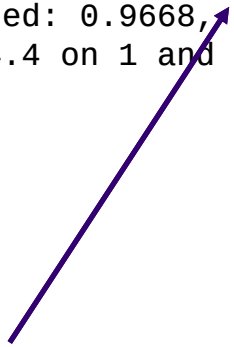
Hypothesis statistic for test that coefficient is NOT equal to zero. (Two tailed). The Null is equal to zero.

P-value for coefficient hypothesis test.

Interpreting R's Output

```
x = 1:20
y = x + rnorm(20)
best_fit = lm(y~x)
summary(best_fit)
```

```
Call: lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-2.04153 -0.43442 -0.01455 0.73806 1.93583

Coefficients:
            Estimate      Std.    Error t value Pr(>|t|)
(Intercept) -0.23738   0.53628   -0.443   0.663
x            1.02521   0.04477   22.901   9.19e-15 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.154 on 18 degrees of freedom
Multiple R-squared: 0.9668,     Adjusted R-squared: 0.965
F-statistic: 524.4 on 1 and 18 DF,   p-value: 9.186e-15
```



Least squares estimate of our standard deviation of error (σ):

$$y_i = mx_i + b + N(0, \sigma)$$

Interpreting R's Output

```
x = 1:20
y = x + rnorm(20)
best_fit = lm(y~x)
summary(best_fit)
```

```
Call: lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-2.04153 -0.43442 -0.01455 0.73806 1.93583

Coefficients:
            Estimate      Std.    Error t value Pr(>|t|)
(Intercept) -0.23738   0.53628   -0.443   0.663
x            1.02521   0.04477   22.901   9.19e-15 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.154 on 18 degrees of freedom
Multiple R-squared: 0.9668,      Adjusted R-squared: 0.965
F-statistic: 524.4 on 1 and 18 DF,   p-value: 9.186e-15
```

R-squared of the model.

Adjusted R-squared, accounts for complexity of formula:

$$R^2 - adj = 1 - \frac{(1-R^2)(n-1)}{(n-1)-p}$$ for p<n-1

Interpreting R's Output

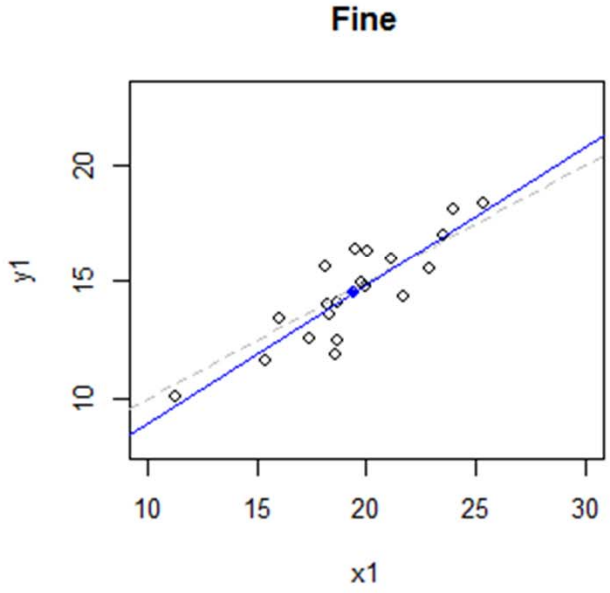
```
x = 1:20
y = x + rnorm(20)
best_fit = lm(y~x)
summary(best_fit)
```

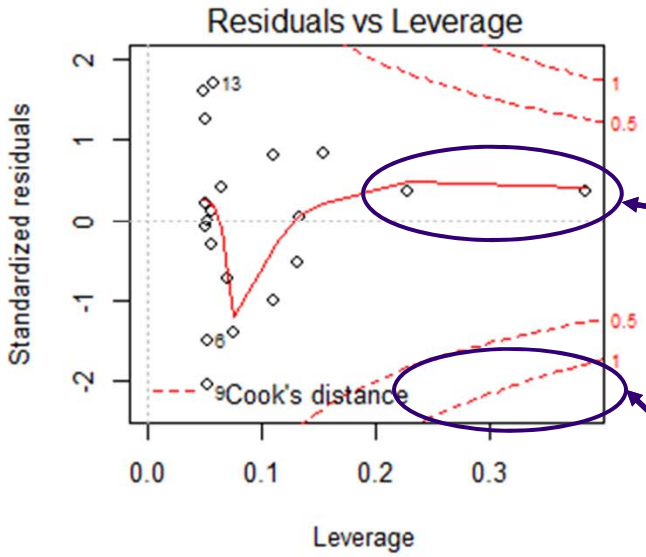
```
Call: lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-2.04153 -0.43442 -0.01455 0.73806 1.93583

Coefficients:
            Estimate      Std.    Error t value Pr(>|t|)
(Intercept) -0.23738   0.53628   -0.443   0.663
x            1.02521   0.04477   22.901   9.19e-15 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.154 on 18 degrees of freedom
Multiple R-squared: 0.9668,      Adjusted R-squared: 0.965
F-statistic: 524.4 on 1 and 18 DF,   p-value: 9.186e-15
```

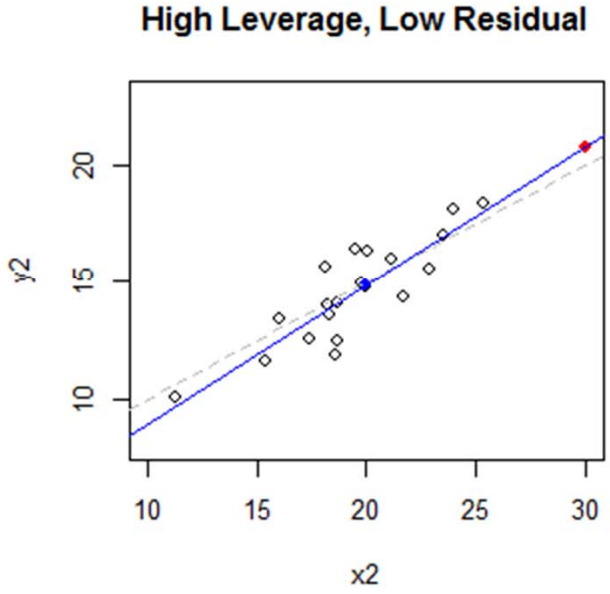
- The F-statistic is a statistic for the hypothesis test that the linear model is a better fit than the mean of y. ($H_0: R^2 = 0$)

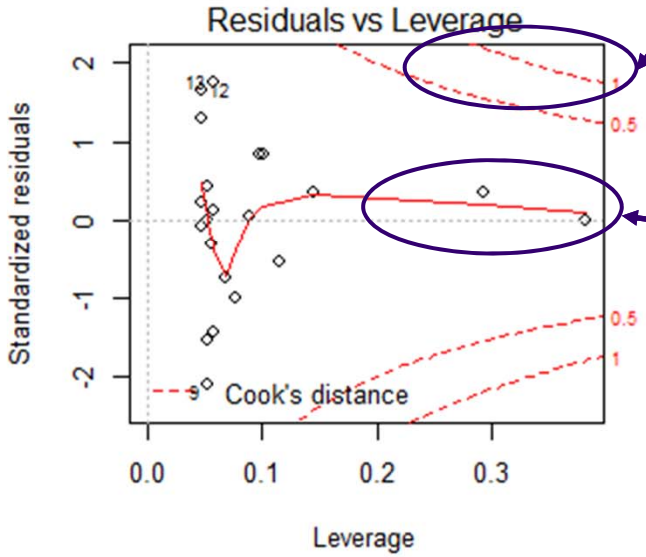
Leverage and Cook's Distance

- Linear regression fits a line based on the means of the y and x values. It fits a line that goes through the means of both values.
- The line pivots around that point relative to the pull of each point. Points that are further away from the mean pull harder on the slope.
- Another way to quantify the 'pull' of each point, is to fit the line to the data without each point and see how the parameters move. This is called Cook's Distance.





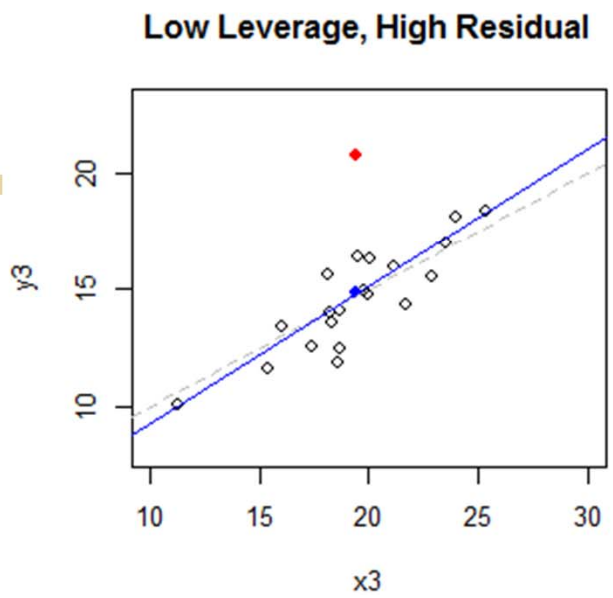


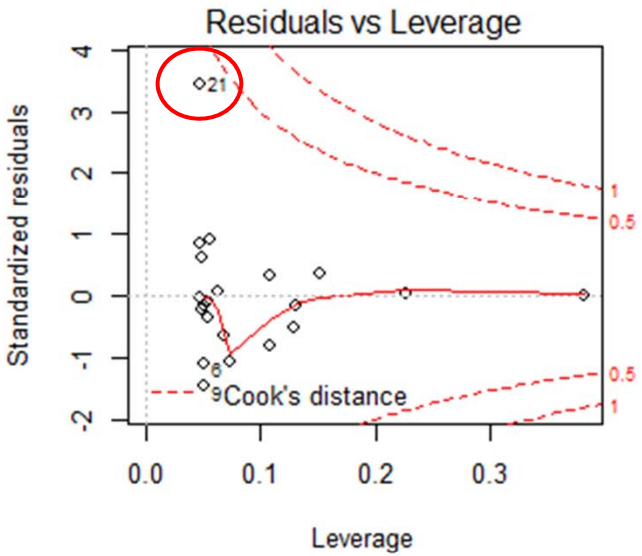


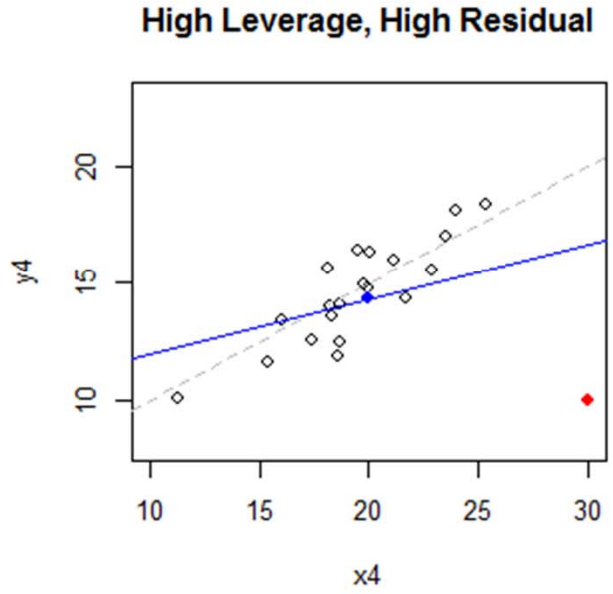
Be aware: High Leverage, Low Residual

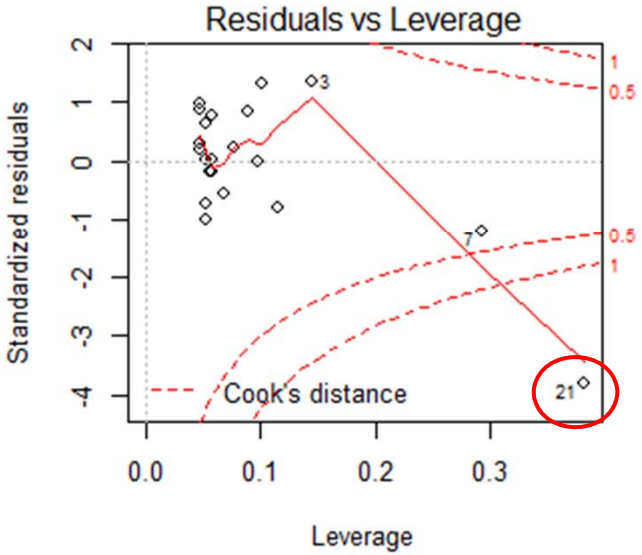
Problem areas: High Leverage, High Residual

Be aware: High Leverage, Low Residual


Low Leverage, High Residual
y3
x3
10
15
20
25
30

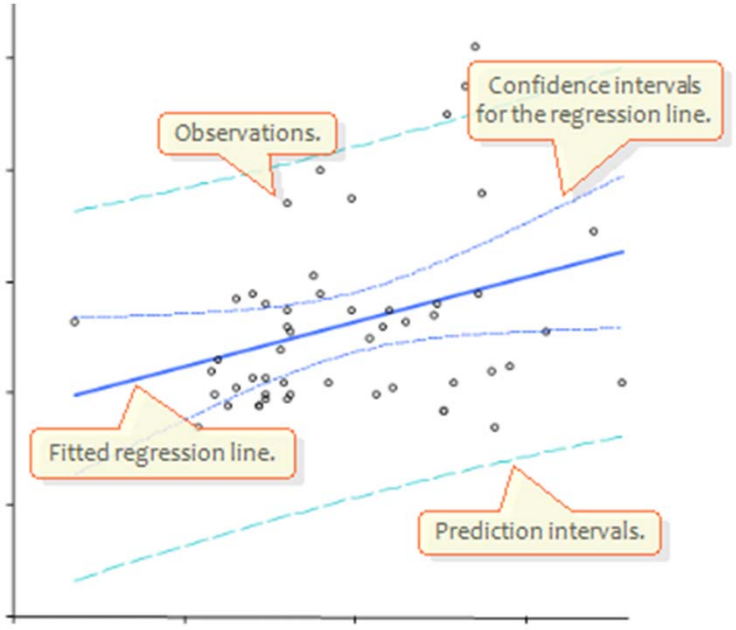

Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
21
6
9
1
0.5


High Leverage, High Residual
y4
x4
10
15
20
25
30


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
21
3
7
1
0.5

Prediction Vs. Confidence in Linear Regression

- Confidence error is the error we assign to the parameters. (m,b,…)
- Prediction error is the error in observing another point.



Linear Models

- Models

  $y = ax + bz + c + \epsilon$

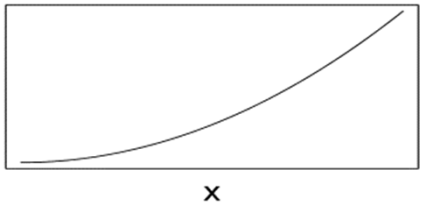
  $y = ax + bz + cx^2 + dxz + \epsilon$

  $y = a\ln(x) + b\sin(z^2) + c + \epsilon$

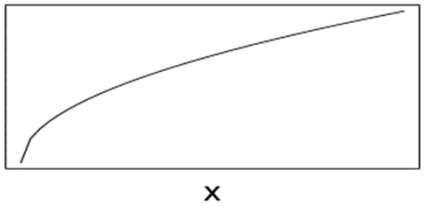
  …
- All of these are linear *in the coefficients*
- Think of these as transformations on the independent variables.
- Methods and interpretation are largely the same as in the simple case

Non-Normal Data

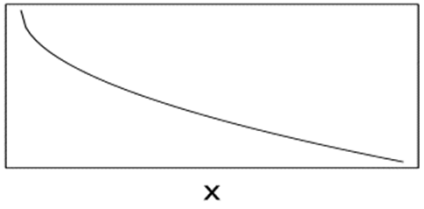
- Transform data to normality:
  - Examples:



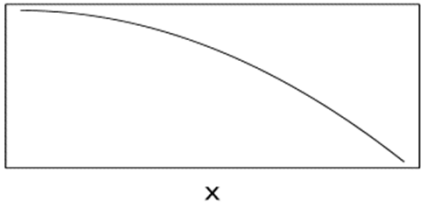
Exp(), power>1



Power>1



Log(), power<1



Power<1, sqrt()

Non-Normal Data

- Transform data to normality:
  - Note you can only use **monotonic** functions for transformations for interpretability.
  - These are functions that preserve:
    - If a < b then f(a) < f(b).
  - Always either non-increasing or non-decreasing.

Non-Linear Relationships & Measurement Error

- Other than transformations, we can't do much here.
- Solutions:
  - Fit a non-linear regression.
  - Use a non-regression machine learning model. (Touch on these at the end of the semester, and it is a large subject of the third class)

- Cannot change measurement error, but you can control for it.

Example: Temperature measurements may have more variance at the high and low ends of an instruments limits. We can add a variable that takes into account how far from the limits the measurement is. More to come on this later.

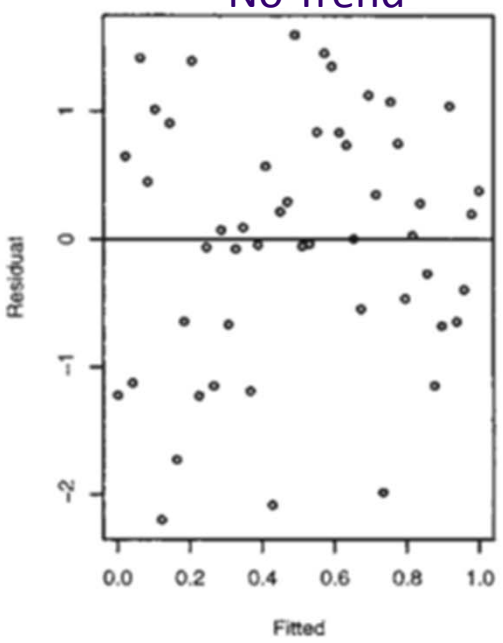
Checking the Residuals

- Always plot the residuals!
- You should always check for trends in the residuals.
- Residuals vs. fitted values:
  - We should never see a trend here. If there is a trend, our linear fitting failed.
- Residuals vs. y:
  - Always positively correlated. The higher the correlation, the worse the fit. This is an effect of the influence of points on the end of the line. A point near the end has much more influence than a point in the middle. Because of this, higher values of y (near the end) tend to have higher residuals (and vice versa for lower values of y).
- Testing the normality of the residuals is also important
  - shapiro.test() in R.

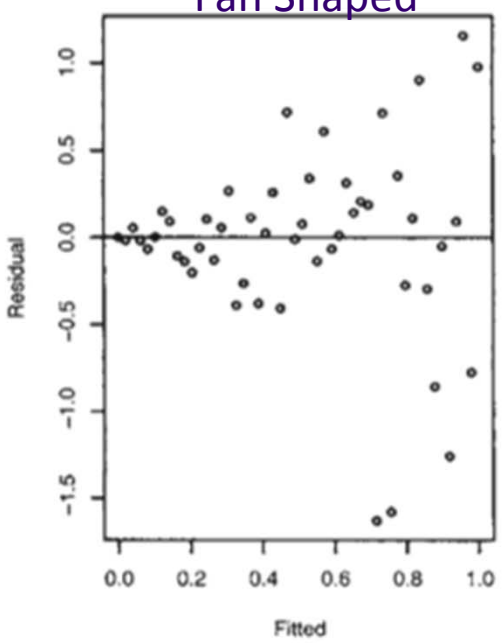
Trends in Residuals vs. Fitted

- Types of outcomes:

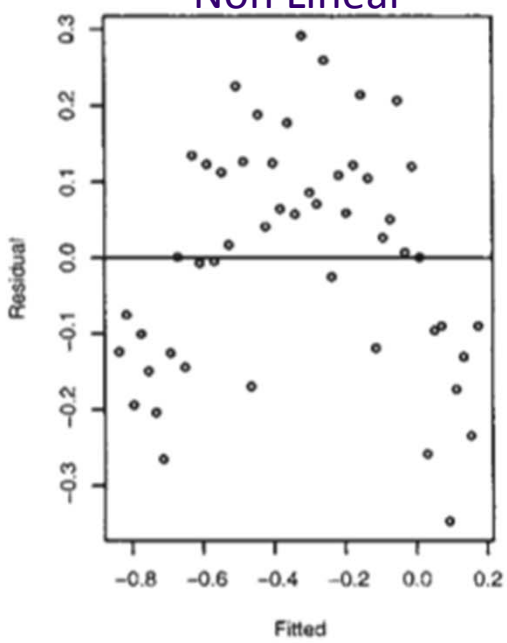
No Trend



Fan Shaped



Non Linear



Trends in Residuals vs. Fitted

- Fan-shaped residuals (Heteroskedastic Residuals):
  - This does not affect our parameter estimates, but only our standard errors.
    - i.e. our population parameter confidence intervals are wider
  - In order to correct for this, we may try transforming our variables
    - Log transforms, sqrt transforms…
- Non-linear residuals
  - Best solution is to use transformation or non-linear regression
- More in-depth statistics information:
  - https://cran.r-project.org/doc/contrib/Faraway-PRA.pdf

Multiple Linear Regression

> Again, by linear, we mean linear in the parameters.

$y = \beta_0 + \beta_1 x_1 + \beta_2 {x_1}^2 + \epsilon$	Univariate Quadratic Model
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$	First Order Model
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$	2nd-order Interaction Model
$y = \beta_0 + \beta_1 e^{x_1} + \beta_2 {x_2}^{0.5} + \epsilon$	First Order Model (with transformations)

β_0 is still the intercept (what y is equal to when all x's = 0).

$\beta_1, \beta_2, \ldots$ Are known as the partial slopes. Each one still represents the change in y per unit increase in x.

Multiple Linear Regression

- Start with a first order model.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

- How do we deal with factor/categorical variables?

Gender
F
M
F
F
M
…

→

Gender
1
0
1
1
0
…

Eye Color
Brown
Brown
Blue
Green
Green
Blue
Brown
…

→

Brown	Blue
1	0
1	0
0	1
0	0
0	0
0	1
1	0
…	…

"One hot encoding"

DayOfWeek
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
…

→

DayOfWeek
1
2
3
4
5
6
7
…

"Factor encoding"

Multiple Linear Regression

- Throwing in all possible variables to help explain our response is sometimes *not* a good thing
  - Variables can be dependent on each other.
  - Variables might not be important to explain the response.
  - Note that the SSE is always larger for reduced models!

How do we choose which combinations of independent variables to use?

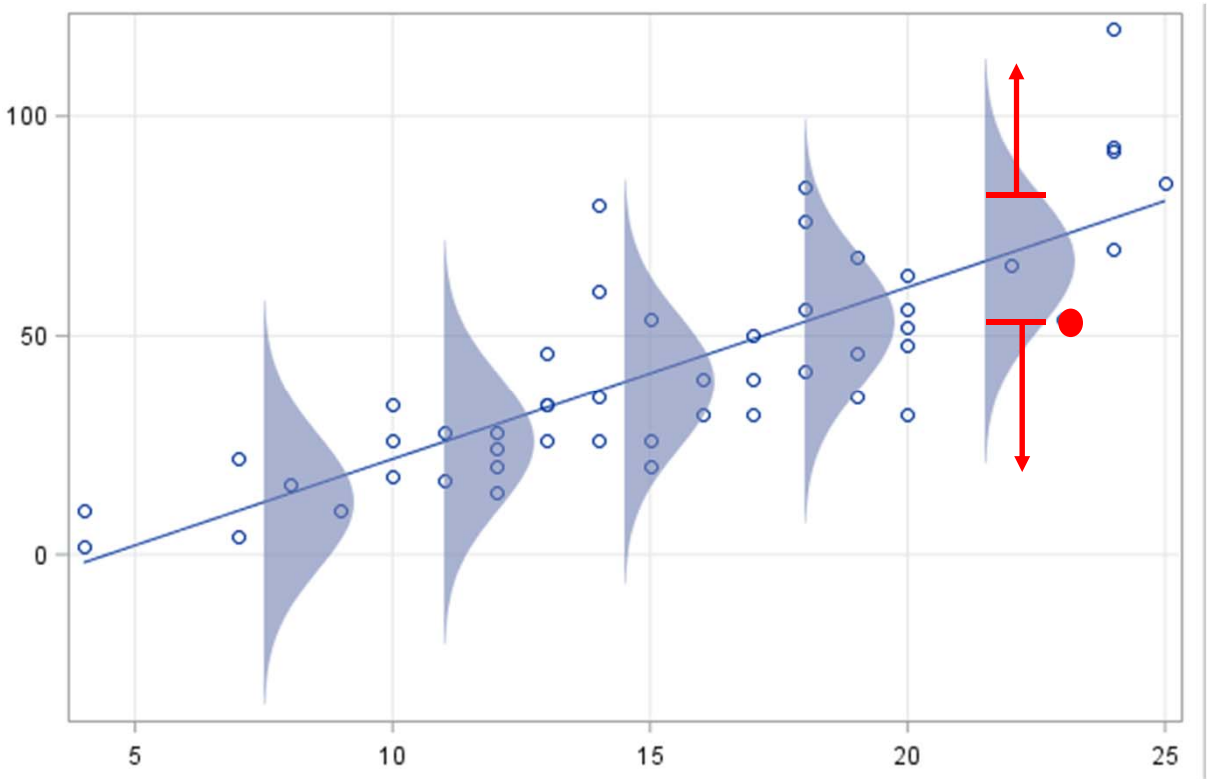
We might consider looking at the difference in SSE between models and the number of explanatory variables.

Multiple Linear Regression

- Given a linear model with known constants:

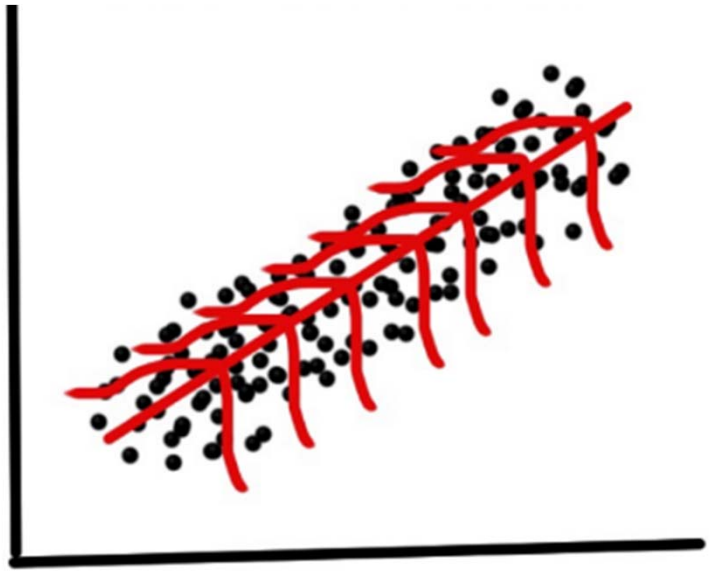
$$y_i = mx_i + b + N(0, \sigma)$$

- Given a point that comes from that line, we can come up with a probability of observing that point.



Multiple Linear Regression

- Linear Model Likelihood
  - We assume errors are normally distributed. From the resulting set of errors, we can come up with a distribution. We then use each residual point and come up with a total error and calculate the probability of that model given our data. This is the likelihood.



To make the calculations easier, we usually take the logarithm of the model (remember we can do this because it is monotonic). This is called the 'log-likelihood'. We will talk more about this when we get to Bayesian Statistics.

Multiple Linear Regression

- Akaike Information Criterion (AIC)
  - Given a model with k-parameters, and a likelihood of L,

$$AIC = 2k - 2\ln(L)$$

  - Note that the more parameters, the higher the AIC.
  - The higher the likelihood, the lower the AIC.
  - Better models have lower AIC values.

Multiple Linear Regression

- How to select the variables in the model?
- Stepwise regression.
  - Forward Selection:
    - Start with no independent variables and add the variables one by one, selecting the variable that improves your criterion the most.
  - Backward Selection
    - Start with all independent variables and remove on at a time. Remove the one that improves the chosen criterion.
- R-demo

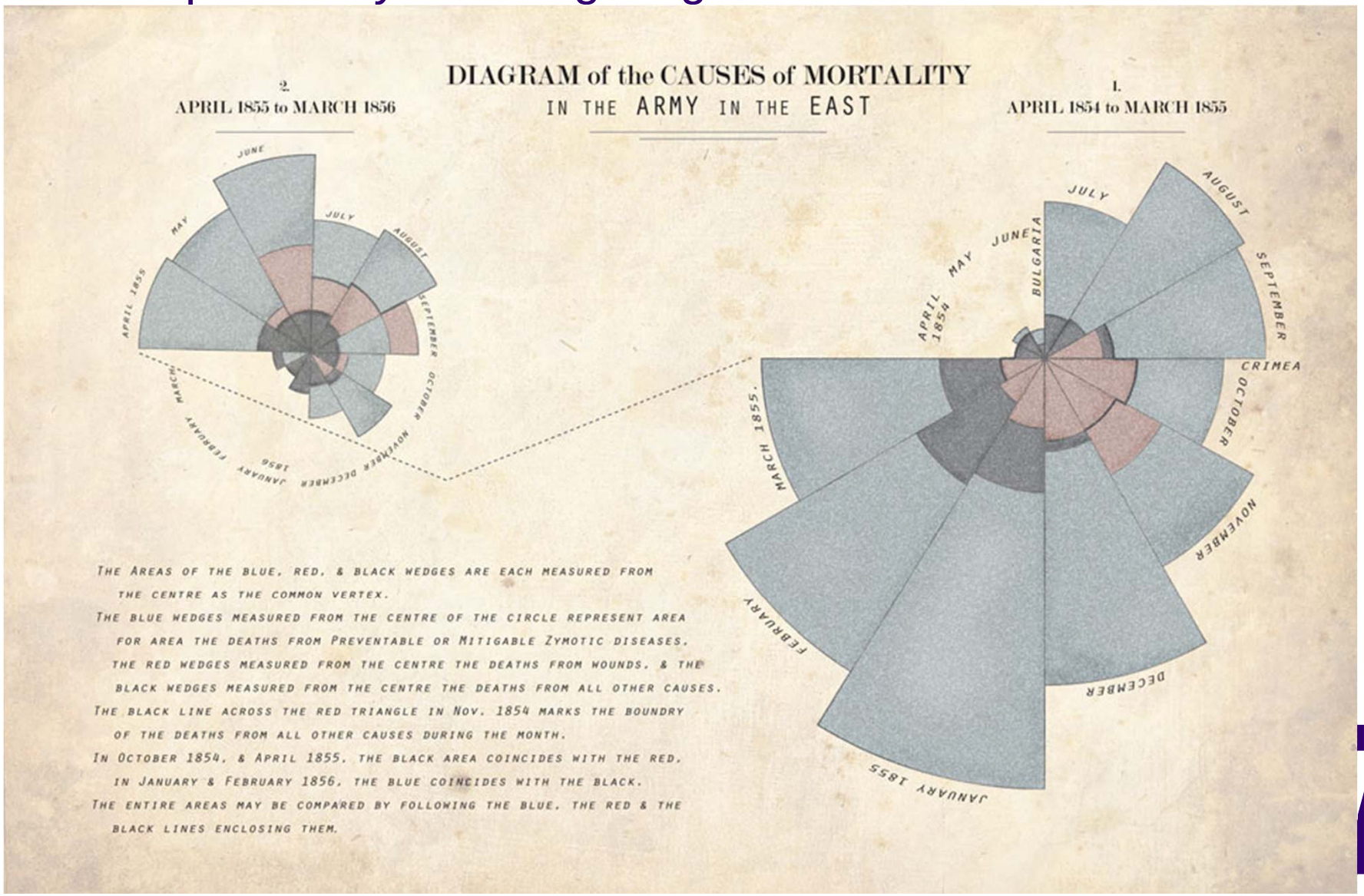
Notes on Final Project

Please email your proposal to me by next week!

- Include the following in your report:
  - Summary and conclusions: with statement of problem
  - Exploration of the dataset: to support conclusions
  - Description of the model
  - Interpretation of the model: to support conclusions
  - End with short statement of conclusions
- R code standards
  - Include sufficient comments to make explain the operation of your code
  - Use functions to simplify and structure you code
  - Label your plots; main label and axis labels + any required legend

Presenting Data Science Results

Example: Analysis of Nightingale data



Assignment

- Complete Homework 5:
  - Construct and evaluate a linear model of automotive price.
    - Model price by engine size, curb weight, and city mpg
      - Hint1: the R model formula is something like:

        ```
        price ~ engine.size + curb.weight + city.mpg
        ```

      - Hint 2: a transformation of either the label (dependent variable) or the features (independent variables) is required.
    - Evaluate the significance of the model coefficients from the model summary
    - Evaluate the performance of the model fit using **both** the diagnostic plots and the model summary.
    - Test normality of residuals (e,g, SW test)
  - You should submit:
    - One R-script.
    - A text document summarizing your findings in the required forma
  - Read Statistical Thinking for Programmers pages 93-97.