Week 7 Inference for regression

2. Inference for prediction

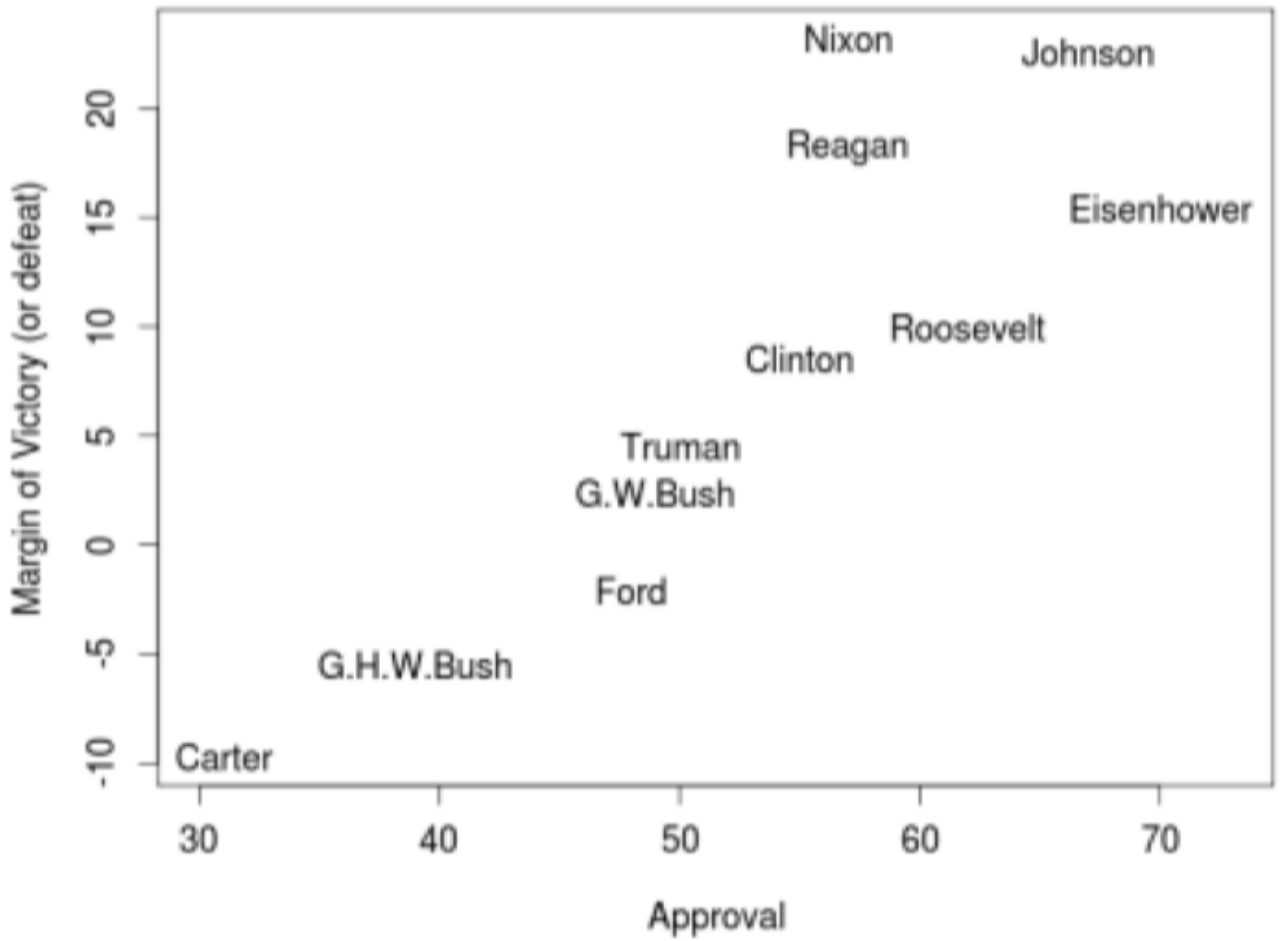
Stat 140 - 04

Mount Holyoke College

Dr. Shan Shan

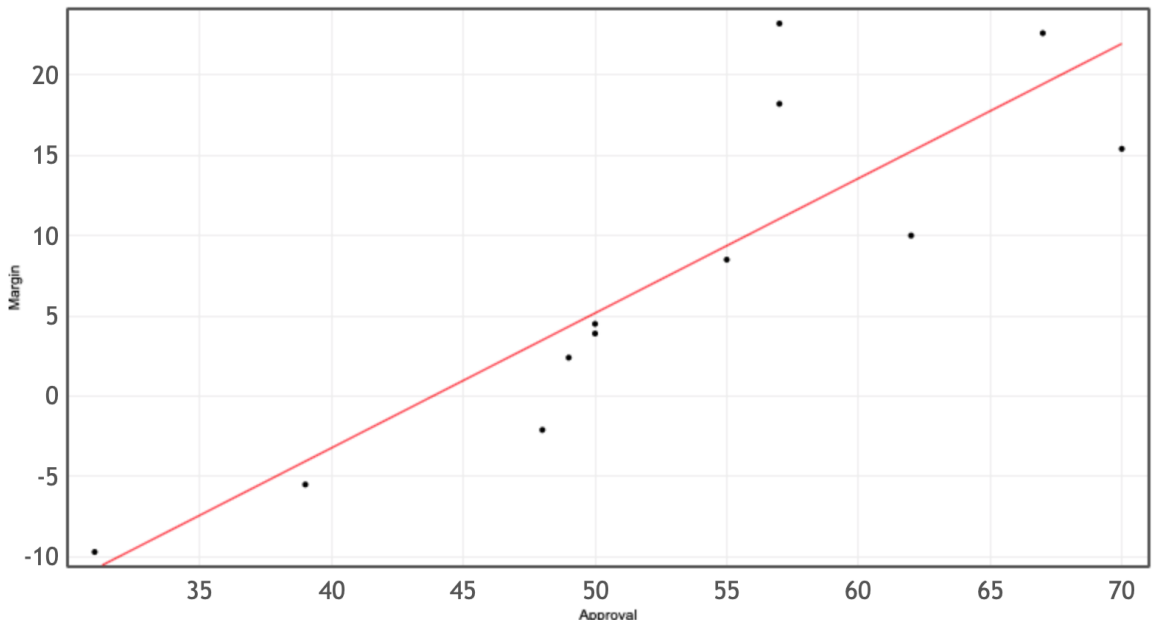
Slides posted at *http://sshanshans.github.io/stat140*

We can build a model using data from past elections to predict an incumbent's margin of victory based on approval rating



Example: Presidential Elections

Election Margin-2e (Margin vs Approval) | Show Data Table | Edit Data | Upload File | Change Column(s)



Summary Statistics Switch Variables

Statistic	Approval	Margin
Mean	52.917	7.617
Standard Deviation	11.041	10.717
Sample Size	12	
Correlation	0.864	
Slope	0.839	
Intercept	-36.76	

Scatterplot Controls

☑ Show Regression Line

What was Obama's predicted margin of victory, based on his approval rating on the day of the election (50%)?

- We would like to use the regression equation to predict y for a certain value of x
- For useful predictions, we also want **interval estimates**
  - Confidence interval
  - Prediction interval

Outline

1. Confidence interval for prediction

2. Prediction interval

- How many US presidents have approval rating equal $= 50\%$ on the election day in our sample?
- When you compute the predicted margin of victory based on approval rating 50%, how many predicted values did you get?
- What does this prediction value really mean??

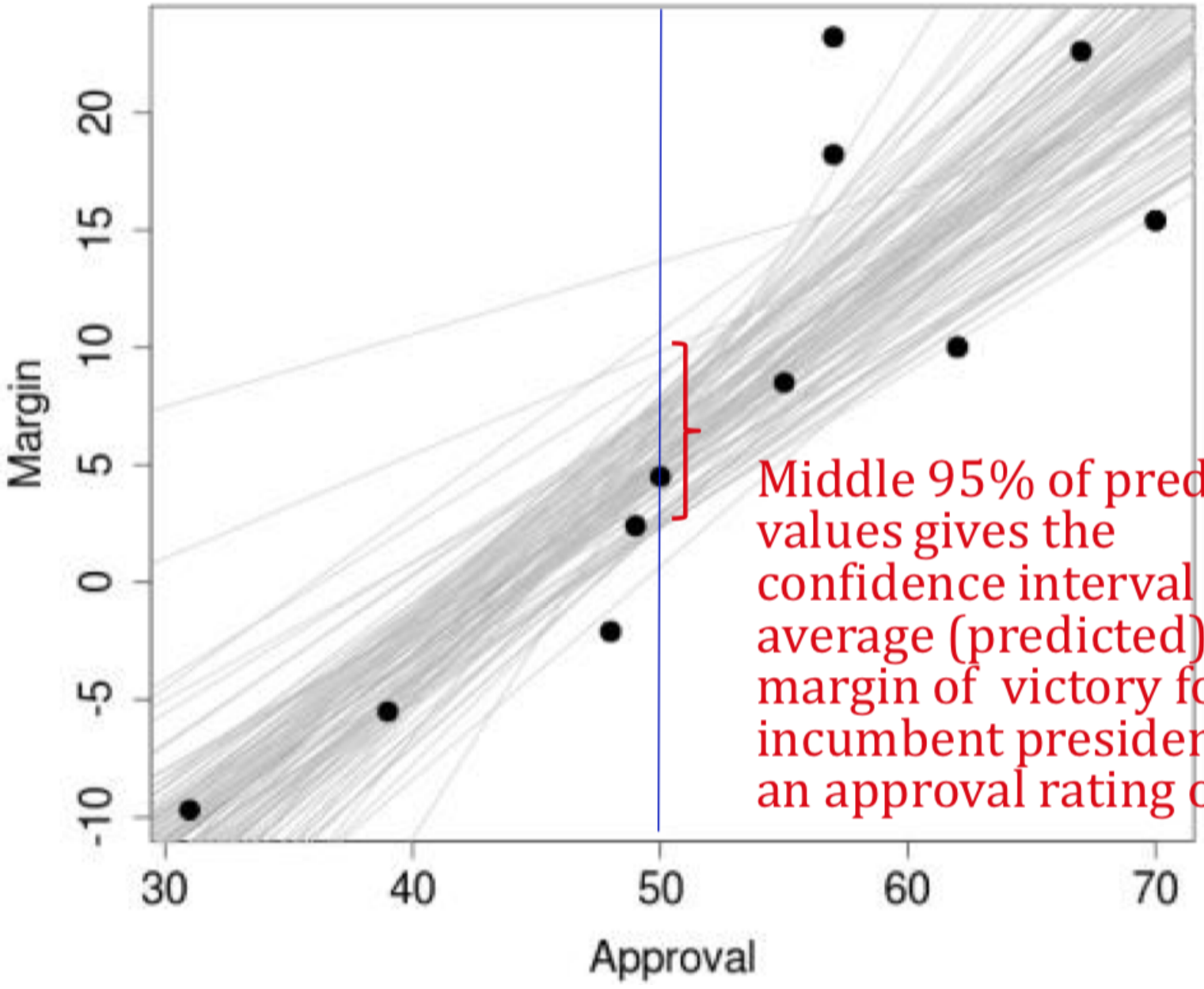
- How many US presidents have approval rating equal $= 50\%$ on the election day in our sample?
- When you compute the predicted margin of victory based on approval rating 50%, how many predicted values did you get?
- What does this prediction value really mean??

The first population parameter

mean response for all elements of the population where the explanatory/predictor value is x^*.

- We need a way to assess the uncertainty in predicted y values for a certain x value... any ideas?

- We need a way to assess the uncertainty in predicted y values for a certain x value... any ideas?
- Take repeated samples, with replacement, from the original sample data (bootstrap)
- Each sample gives a slightly different fitted line
- If we do this repeatedly, take the middle P% of predicted y values at x^* for a confidence interval of the predicted y value at x^*


Margin
20
15
10
5
0
-5
-10
30
40
50
60
70
Approval
Middle 95% of predicted values gives the confidence interval for average (predicted) margin of victory for an incumbent president with an approval rating of 50%

For $x^* = 50\%$, the confidence interval is $(1.07, 9.52)$

This means,

We are 95% confident that the average margin of victory for incumbent U.S. presidents with approval ratings of 50% is between 1.07 and 9.52 percentage points

But wait, this still doesn't tell us about a particular incumbent! We don't care about the average, we care about an interval for one incumbent president with an approval rating of 50%!

Outline

1. Confidence interval for prediction

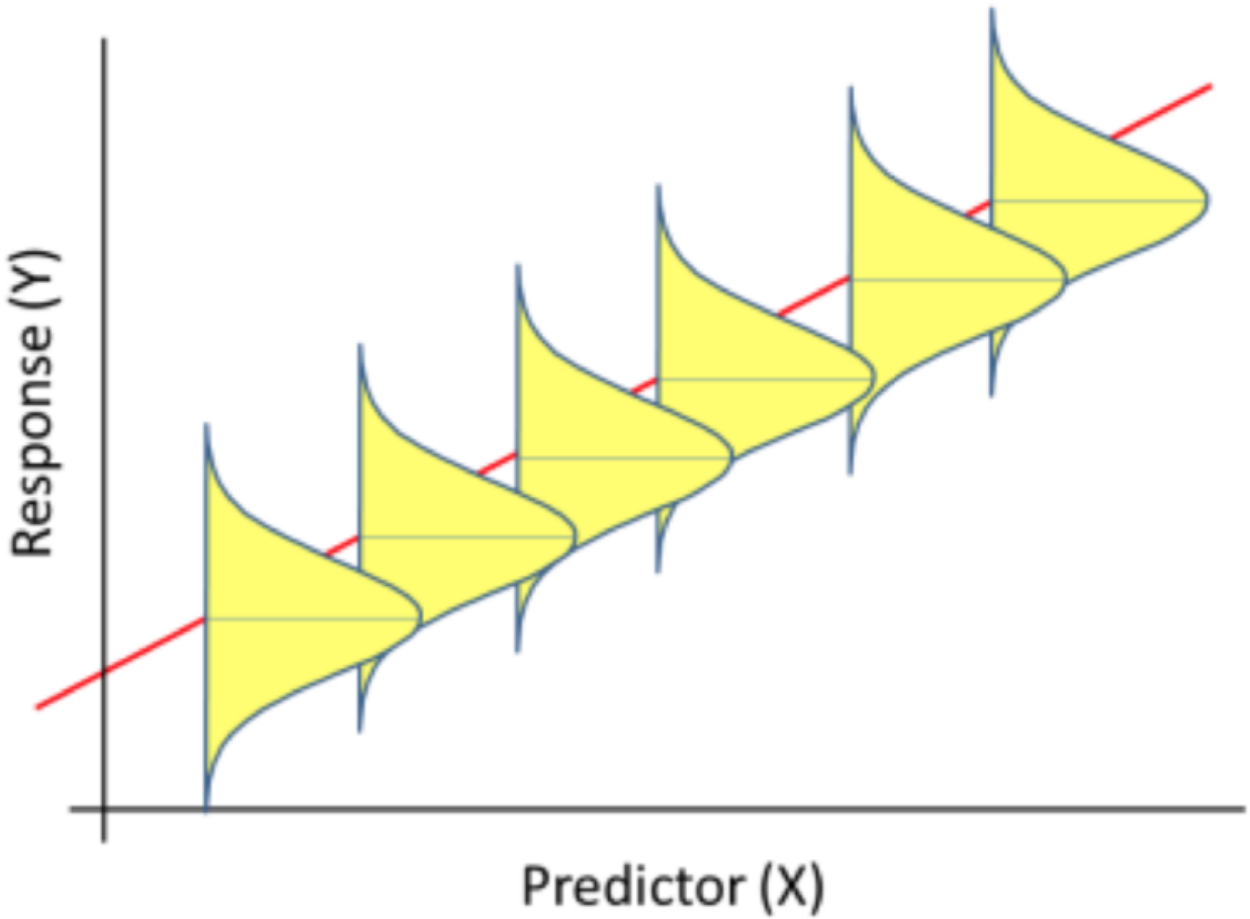
2. Prediction interval

We can also calculate a prediction interval for y values for a certain x value

"We are 95% confident that the y value for $x = x^*$ lies in this interval"

This takes into account the variability in the line (in the predicted value) AND the uncertainty around the line (the random errors)

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad \varepsilon \sim N(0, \sigma_\varepsilon)$$



- A confidence interval has a given chance of capturing the mean y value at a specified x value
- A prediction interval has a given chance of capturing the y value for a particular case at a specified x value

Poll question

For a given x value, which will be wider?

a. Confidence interval
b. Prediction interval

Based on the data and the simple linear model:

The predicted margin of victory for an incumbent with an approval rating of 50% is 5.3 percentage points

We are 95% confident that the margin of victory (or defeat) for an incumbent with an approval rating of 50% will be between -8.8 and 19.4 percentage points

NOTE: You will never need to use these formulas in this class – you will just have RStudio do it for you.

Confidence Interval:

$$\hat{y} \pm t^* \times s_e \sqrt{\frac{1}{n} + \frac{\left(x^* - \bar{x}\right)^2}{(n-1)s_x^2}}$$

Prediction Interval:

$$\hat{y} \pm t^* \times s_e \sqrt{1 + \frac{1}{n} + \frac{\left(x^* - \bar{x}\right)^2}{(n-1)s_x^2}}$$

s_e : estimate for the standard deviation of the residuals

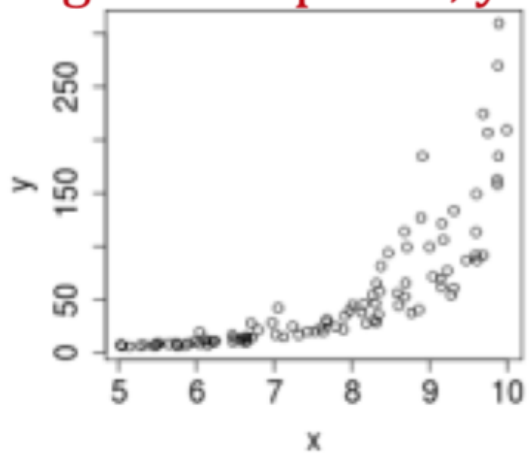
- No outliers (points that don't fit the trend)
- Straight enough?
- Does the plot thicken?
- Sample representative of population
- Independence
- Normally distributed residuals (or large enough sample size)

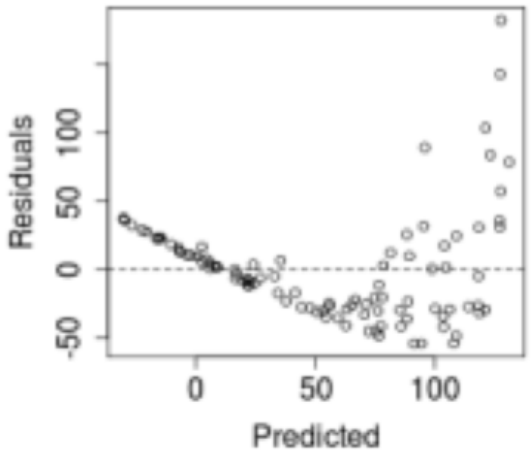
What if the conditions for inference aren't met???

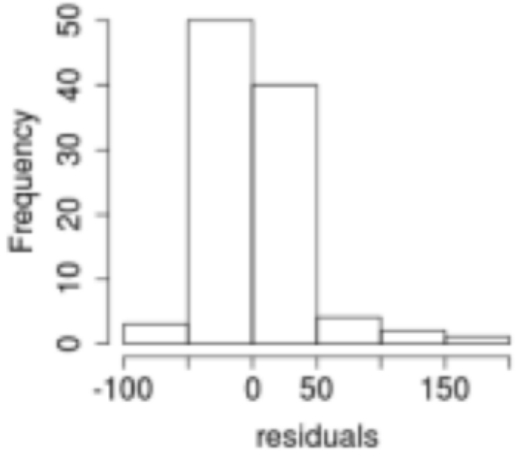
- Option 1: Take 200 level Stat classes and learn more about modeling!
- Option 2: Try a transformation...
  You can take any function of y and use it as the response, but the most common are
  - $\log(y)$ (natural logarithm - ln)
  - $\sqrt{y}$ (square root)
  - y^2 (squared)
  - e^y (exponential))

log(y)

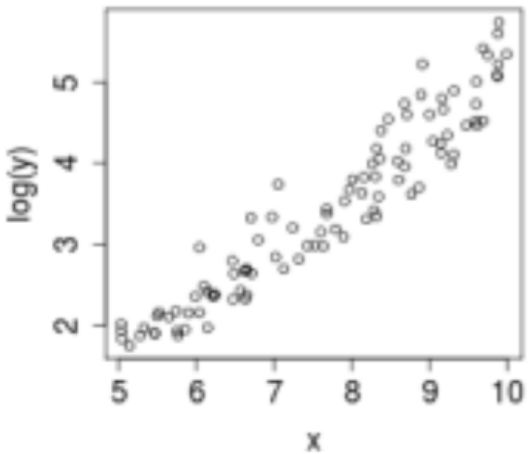
Original Response, y:

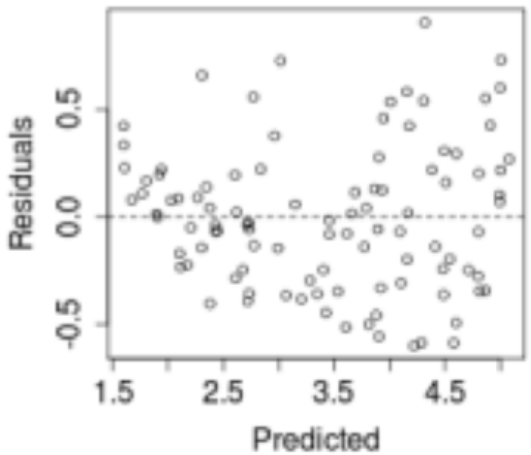


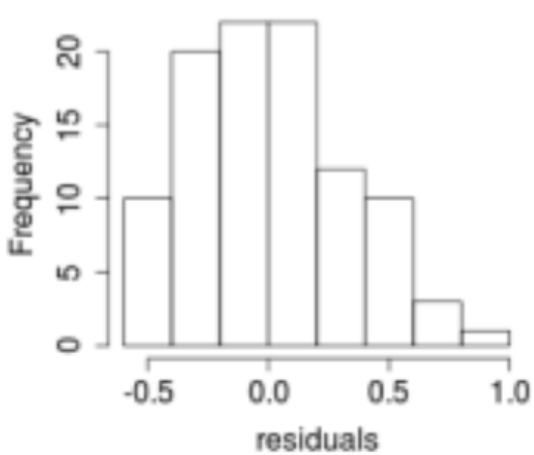




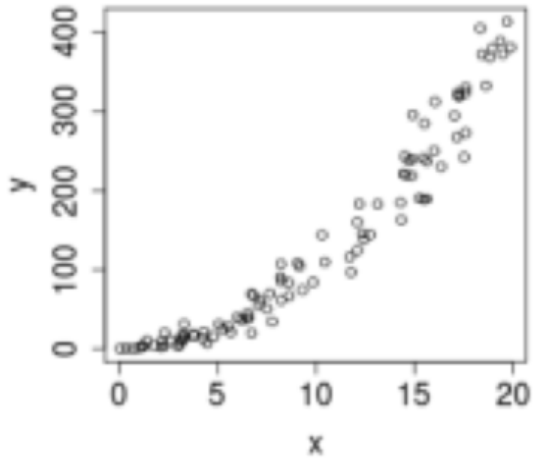
Logged Response, log(y):

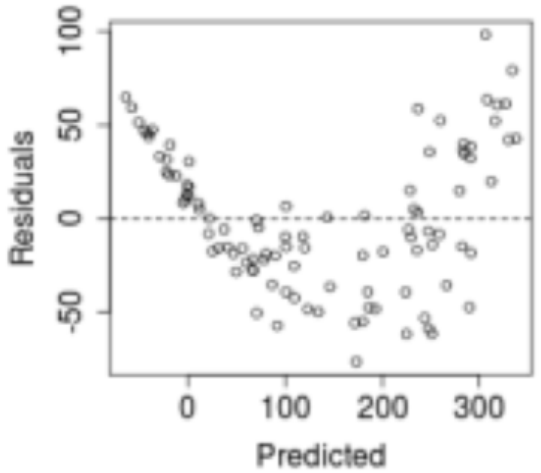


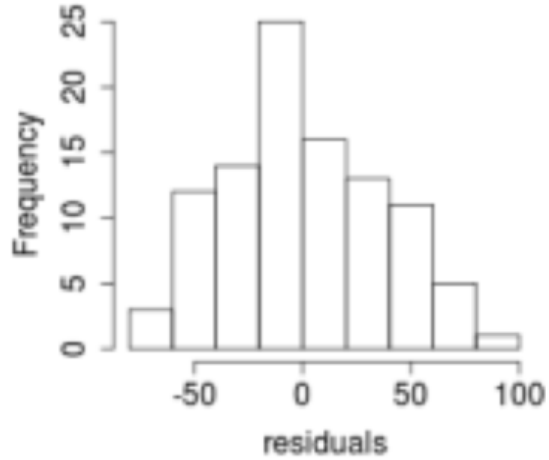




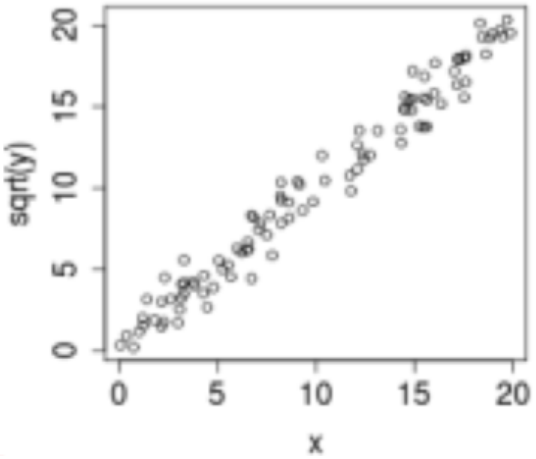
Original Response, y:

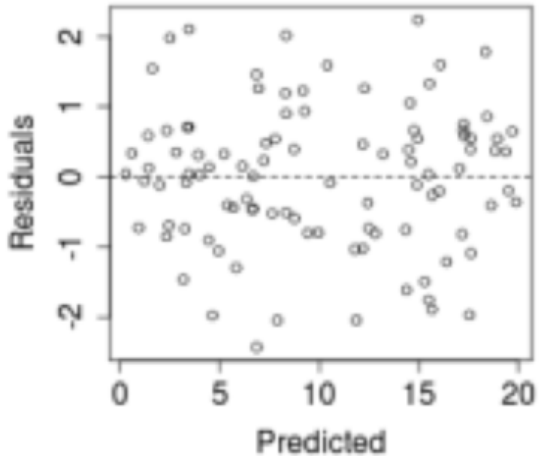


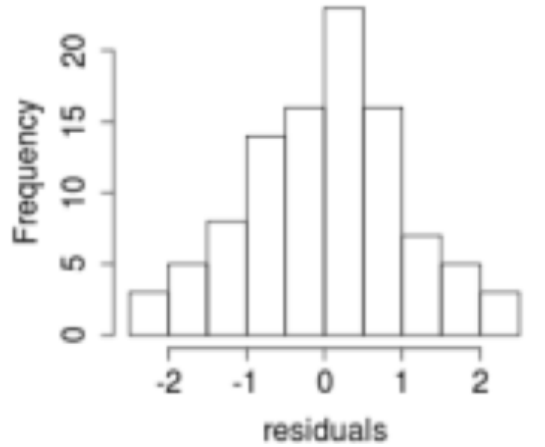




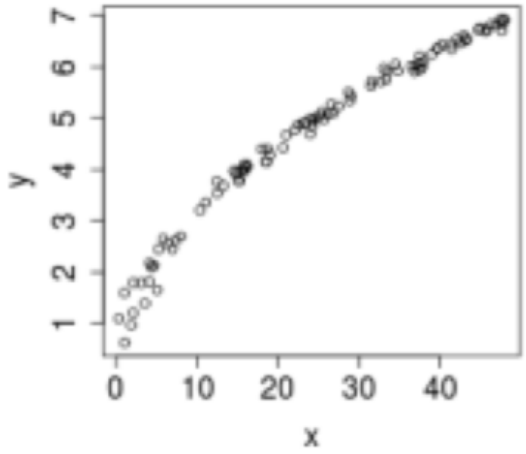
Square root of Response, √y:



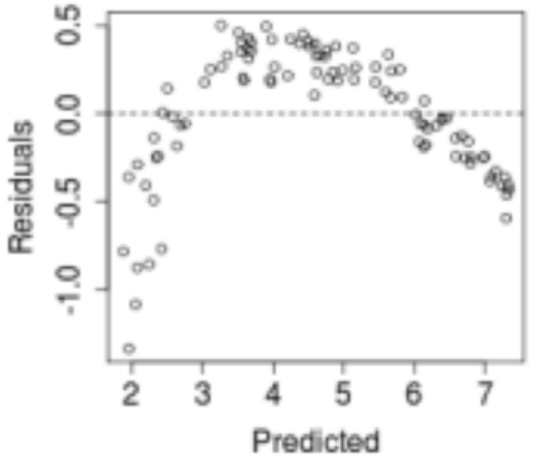


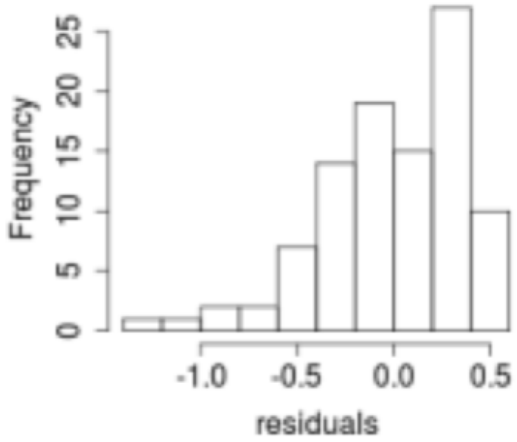


Original Response, y:

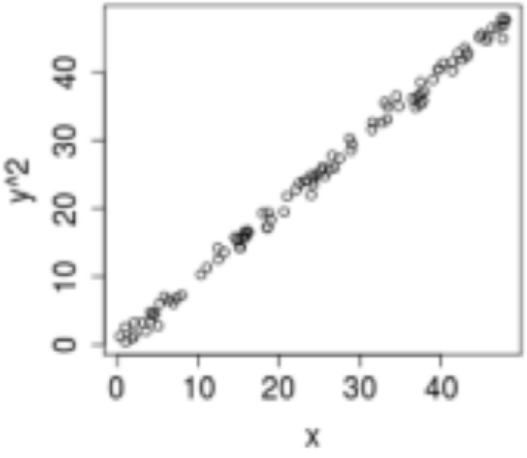


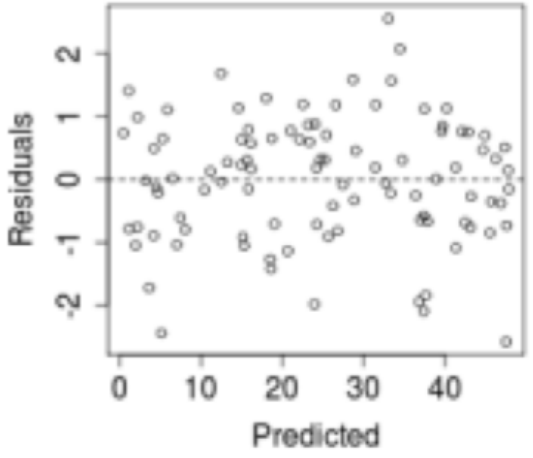
y^2

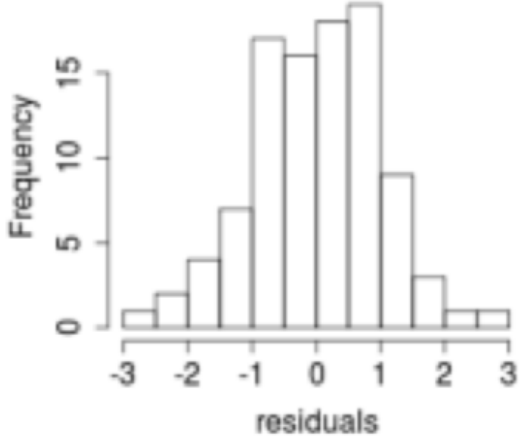




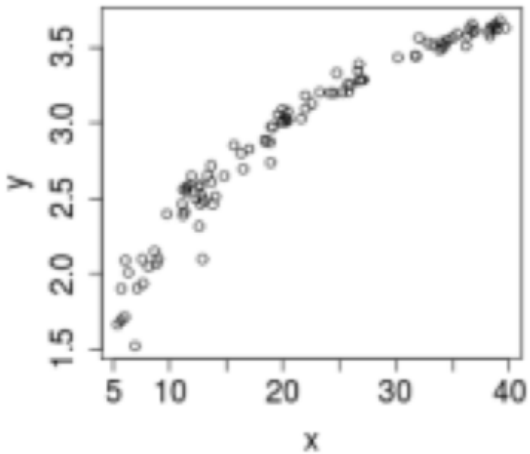
Squared response, y^2:

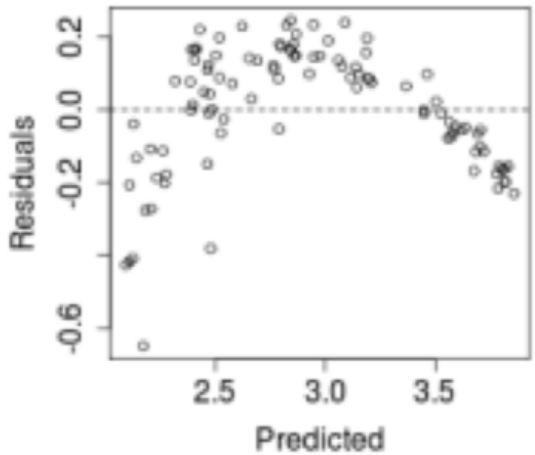


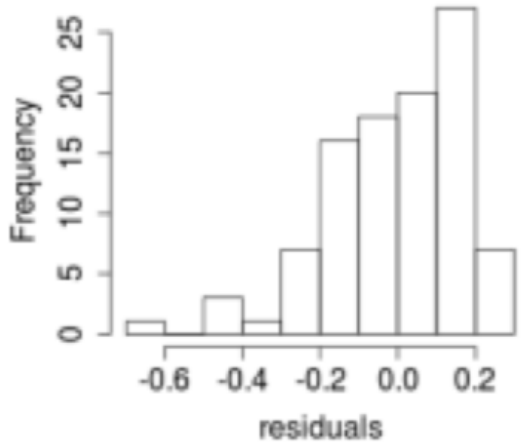




Original Response, y:

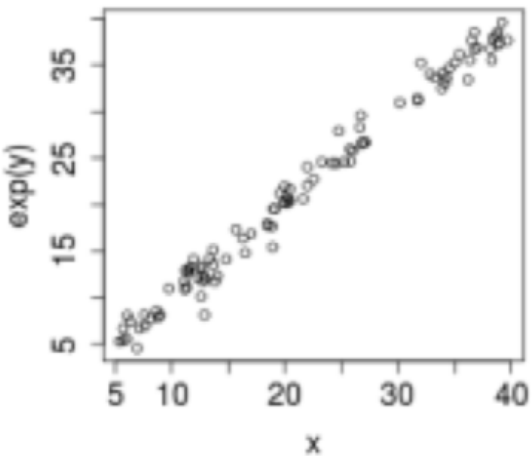


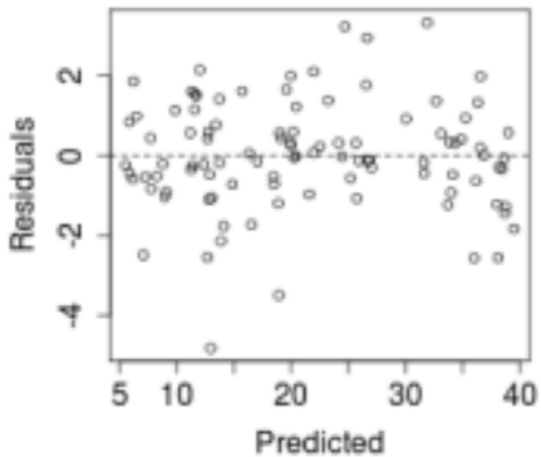


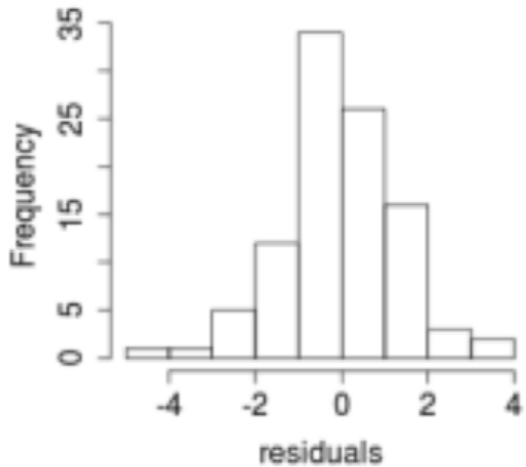


e^y

Exponentiated Response, e^y:







- Interpretation becomes a bit more complicated if you transform the response – it should only be done if it clearly helps the conditions to be met
- If you transform the response, be careful when interpreting coefficients and predictions
- You do NOT need to know which transformation would be appropriate for given data in this class, but they may help if conditions are for future data you may want to analyze