STAT 101
Dr. Kari Lock Morgan

Simple Linear Regression

SECTIONS 9.3

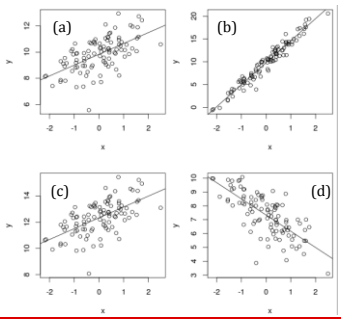
- Confidence and prediction intervals (9.3)
- Conditions for inference (9.1)

Want More Stats???

- If you have enjoyed learning how to analyze data, and want to learn more:
  - take **STAT 210** (Regression Analysis)
  - Applied, focused on data analysis
  - Recommended for any major involving data analysis
  - Only prerequisite is STAT 101
- If you like math and want to learn more of the mathematical theory behind what we've learned:
  - take **STAT 230** (Probability)
  - and then **STAT 250** (Mathematical Statistics)
  - Prerequisite: multivariable calculus

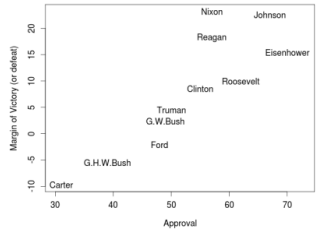
Review

Which plot goes with the line $\hat{y} = x + 10$?

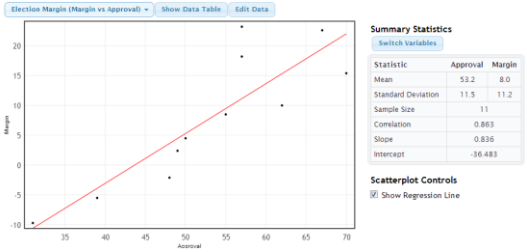


Presidential Elections

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



Presidential Elections



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

Prediction

- 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***
- We will predict the value of y at x = x*

Point Estimate

• The point estimate for the average y value at x=x* is simply the predicted value:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

• Alternatively, you can think of it as the value on the line above the x value

• The uncertainty in this point estimate comes from the uncertainty in the coefficients

Confidence Intervals

• We can calculate a ***confidence interval*** for the ***average*** *y* value for a certain *x* value

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

• Equivalently, the confidence interval is for the point estimate, or the predicted value

• This is the amount the line is free to "wiggle," and the width of the interval decreases as the sample size increases

Bootstrapping

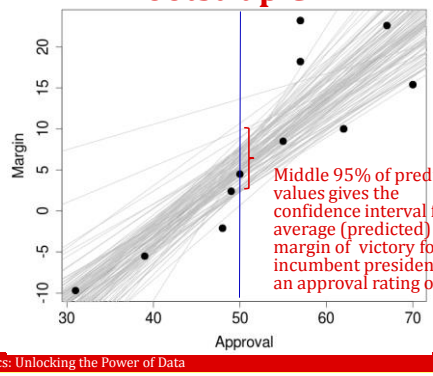
• We need a way to assess the uncertainty in predicted y values for a certain x value… any ideas?

• Take repeated samples, with replacement, from the original sample data (bootstrap)

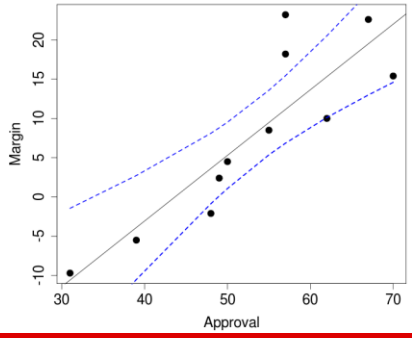
• Each sample gives a slightly different fitted line

• 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*

Bootstrap CI



Confidence Interval



Confidence Interval

• For x* = 50%: (1.07, 9.52)

• 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%!

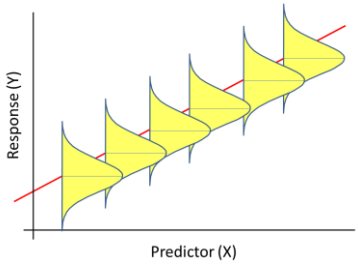
Prediction Intervals

- 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)

Intervals



Intervals

- 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
- For a given x value, which will be wider?

  a) Confidence interval
  b) Prediction interval

Intervals

- As the sample size increases:
  - the standard errors of the coefficients decrease
  - we are more sure of the equation of the line
  - the widths of the confidence intervals decrease
  - for a huge *n*, the width of the CI will be almost 0
- The prediction interval may be wide, even for large *n*, and depends more on the correlation between *x* and *y* (how well *y* can be linearly predicted by *x*)

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

Formulas for Intervals

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

Confidence Interval:

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

s_e : estimate for the standard deviation of the residuals

Prediction Interval:

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

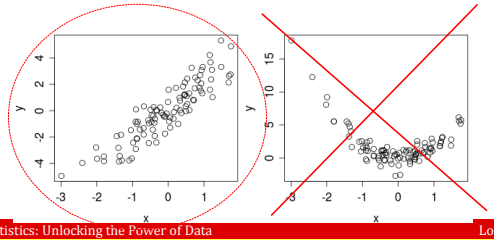
Conditions

Inference based on the simple linear model is only valid if the following conditions hold:

1) Linearity
2) Constant Variability of Residuals
3) Normality of Residuals

Linearity

- The relationship between x and y is linear (it makes sense to draw a line through the scatterplot)

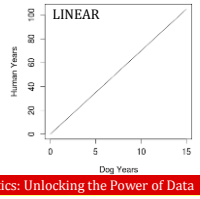


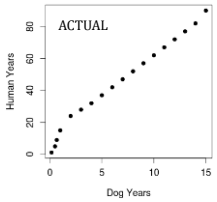
Dog Years

- From www.dogyears.com:

"The old rule-of-thumb that one dog year equals seven years of a human life is not accurate. The ratio is higher with youth and decreases a bit as the dog ages."

- 1 dog year = 7 human years
- Linear: human age = 7×dog age



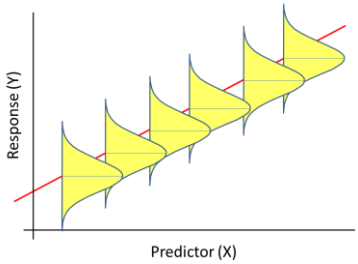


"All models are wrong, but some are useful"

-George Box

Simple Linear Model

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



Residuals (errors)

Conditions for residuals:

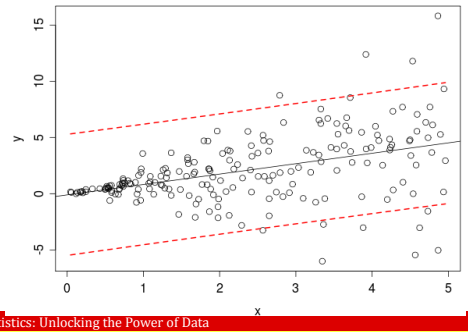
$$\varepsilon_i \sim N\left(0, \sigma_\varepsilon\right)$$

The errors are normally distributed	The average of the errors is 0	The standard deviation of the errors is constant for all cases
Check with a histogram	(Always true for least squares regression)	Constant vertical spread in the residual plot

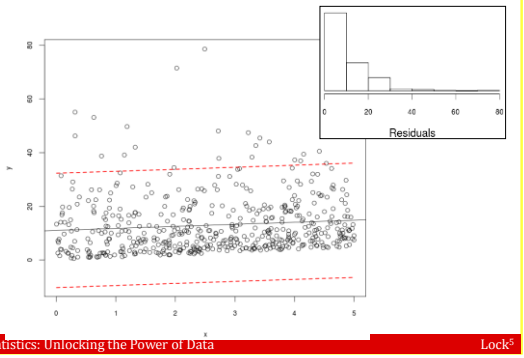
Conditions not Met?

- If the association isn't linear:
  - ⇒ Try to make it linear (transformation)
  - ⇒ If can't make linear, then simple linear regression isn't a good fit for the data
- If variability is not constant, or residuals are not normal:
  - ⇒ The model itself is still valid, but *inference* may not be accurate

Non-Constant Variability



Non-Normal Residuals

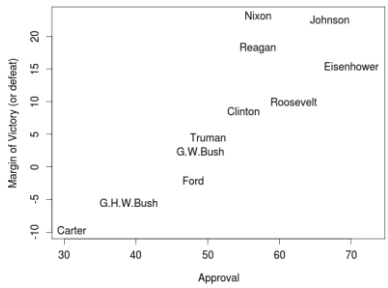


Simple Linear Regression

1) Plot your data!
   - Association approximately linear?
   - Outliers?
   - Constant variability?
2) Fit the model (least squares)
3) Use the model
   - Interpret coefficients
   - Make predictions
4) Look at histogram of residuals (normal?)
5) Inference (extend to population)
   - Inference on slope
   - Confidence and prediction intervals

President Approval and Re-Election

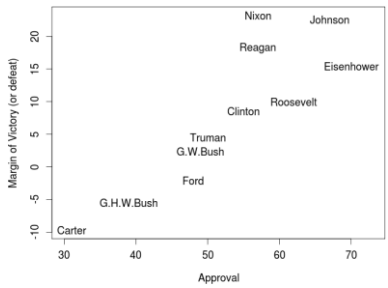
1. Plot the data:



Is the trend approximately linear? (a) Yes (b) No

President Approval and Re-Election

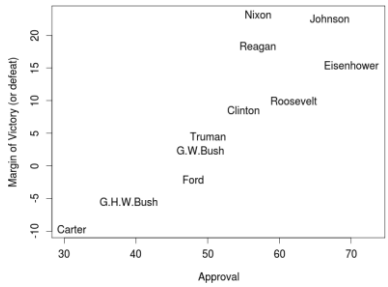
1. Plot the data:



Are there obvious outliers? (a) Yes (b) No

President Approval and Re-Election

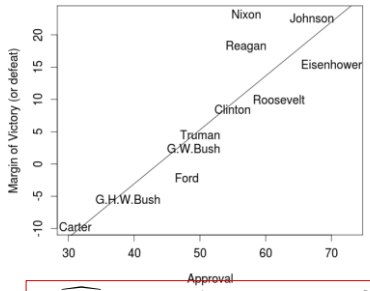
1. Plot the data:



Is there approximately constant variability? (a) Yes (b) No

President Approval and Re-Election

2. Fit the Model:



$\widehat{Margin} = -36.5 + 0.84\ Approval$

President Approval and Re-Election

3. Use the model:

$\widehat{Margin} = -36.5 + 0.84\ Approval$

Which of the following is a correct interpretation?

a) For every percentage point increase in margin of victory, approval increases by 0.84 percentage points
b) For every percentage point increase in approval, predicted margin of victory increases by 0.84 percentage points
c) For every 0.84 increase in approval, predicted margin of victory increases by 1

President Approval and Re-Election

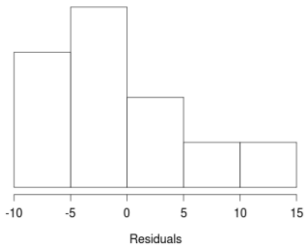
3. Use the model:

$\widehat{Margin} = -36.5 + 0.84\ Approval$

The predicted margin of victory for an incumbent with an approval rating of 50%:

President Approval and Re-Election

4. Look at histogram of residuals:



Are the residuals approximately normally distributed?

(a) Yes (b) No

President Approval and Re-Election

5. Inference

Should we do inference?

(a) Yes
(b) No

President Approval and Re-Election

5. Inference

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -36.4832     8.8553  -4.120 0.002598 **
Approval      0.8356     0.1631   5.124 0.000624 ***
```

Give a 95% confidence interval for the slope coefficient.

Is it significantly different than 0?

(a) Yes
(b) No

President Approval and Re-Election

5. Inference:

We don't really care about the slope coefficient, we care about the margin of victory for a president with an approval rating of 50%.

A 95% prediction interval for margin of victory for an incumbent with an approval rating of 50% is -8.8 to 19.4.

Obama's margin of victory in 2012: **2.8** (50.6% Obama to 47.8% Romney)

Conditions

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

• Option 1 (best option): Take STAT 210 and learn more about modeling!

• Option 2: Try a transformation…

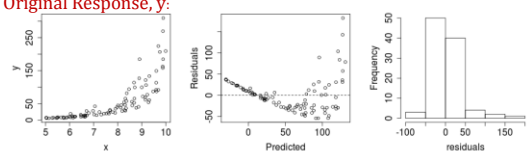
Transformations

• If the conditions are not satisfied, there are some common ***transformations*** you can apply to the response variable

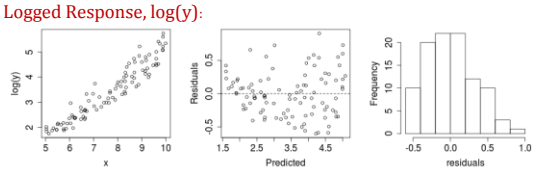
• You can take any function of *y* and use it as the response, but the most common are
  • log(y) (natural logarithm - ln)
  • $\sqrt{y}$ (square root)
  • y^2 (squared)
  • e^y (exponential))

log(y)

Original Response, y:

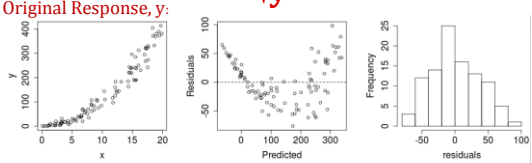


Logged Response, log(y):

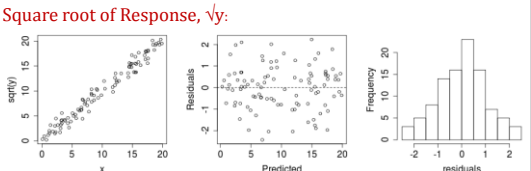


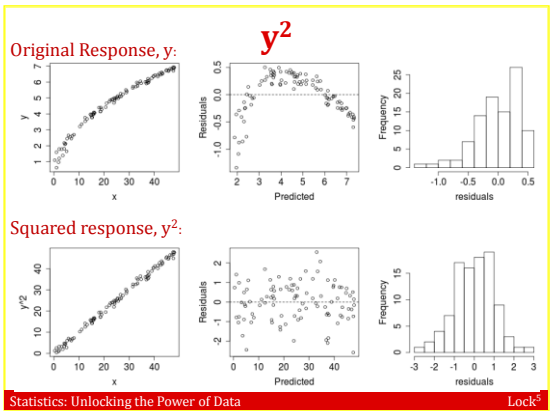
$\sqrt{y}$

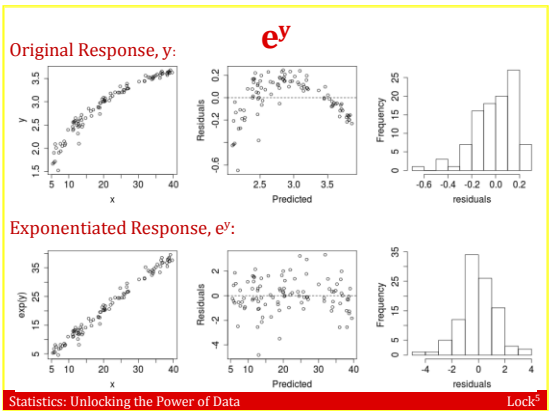
Original Response, y:



Square root of Response, $\sqrt{y}$:







Transformations

- 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
- The slope will now have different meaning, and predictions and confidence/prediction intervals will be for the transformed response

Transformations

- You do NOT need to know which transformation would be appropriate for given data on the exam, but they may help if conditions are not met for Project 2 or for future data you may want to analyze

Exam 2: In-Class

- In class Wednesday 4/2
- Cumulative, but emphasis is on material since Exam 1 (Chapters 5-9, we skipped 8.2 and 9.2)
- Closed book, but allowed 2 double-sided pages of notes prepared by you
- You won't have technology, so won't have to compute p-values, but should be able to tell by looking at a distribution whether something is significant

Key to Success

- WORK PRACTICE PROBLEMS!
- Recommended problems: Units C and D Essential Synthesis and review problems (solutions on course website under documents)
- In Unit D odd essential synthesis and review problems, skip D9, D17, D25, D47, D52-D58 (will cover after exam)
- Want more practice problems??? Full solutions to all odd problems in the book are on reserve in Perkins

To Do

- Read Chapter 9
- Do Homework 7 (due Monday, 3/31)
  - NO LATE HOMEWORK ACCEPTED – SOLUTIONS WILL BE POSTED IMMEDIATELY AFTER CLASS
- Study for Exam 2 (Wednesday 4/2)