Julie Yoon
Samuel Garcia 11:00 a.m.-12:30 p.m. February 5, 2014

STA 371H Scribes Notes

<u>Prediction Intervals</u>

- Initially, we tried to figure out how much a value deviated from the least squares line.
- With an interval we determined accuracy by seeing how many 'cases' fall within it.

  Odometer vs. Mileage graph:
  - Standard deviation: how much a data point deviates from the line.
  - How spread out is the data?
    - We look at the shaded area (1 standard deviation)
      - How accurate is this shaded area?
        - Figure out by counting how many cases are in the area
        - 1 standard deviation= 65%-75%
        - 2 standard deviation = 95%

<u>Brain vs. Body graph:</u>

- Transform the graph to a log scale to get linear data
- Dashed lines: confidence interval borders
  - This is prediction of confidence interval on log scale
  - On the original scale, the confidence intervals (dashed lines) would be curved.

- In the past, we used the equation:

$$y_i = \hat{B}_0 + \hat{B}_1 x_i + e_i$$

- When estimating the value of y* for a new x*:

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

$$y^* = \hat{y}^* \pm ?$$

$$\left.\begin{array}{l} \pm 1\hat{\sigma}_e \\ \pm 2\hat{\sigma}_e \end{array}\right] \text{- likely size of error}$$

  - The predicted y could be +/- 1 standard deviation or 2 standard deviations

- So the equation to find the upper and lower bound 1 standard deviation away would be:
  - **Upper**:

$$y^* = \hat{B}_0 + \hat{B}_1 x_* + 1\hat{\sigma}_e$$

- **Lower**:

$$y^* = \hat{B}_0 + \hat{B}_1 x_* - 1\hat{\sigma}_e$$

- After converting to the log scale:

$$\log y_i = \hat{B}_0 + \hat{B}_1 \log x_i + \boxed{\varepsilon_i} \rightarrow \hat{\sigma}_\varepsilon$$

- **Upper:**

$$\log y^* = \hat{B}_0 + \hat{B}_1 \log x_* + \hat{\sigma}_\varepsilon$$

$$e^{\log y^*} = e^{\{\hat{B}_0 + \hat{B}_1 \log x_* + \hat{\sigma}_\varepsilon\}}$$

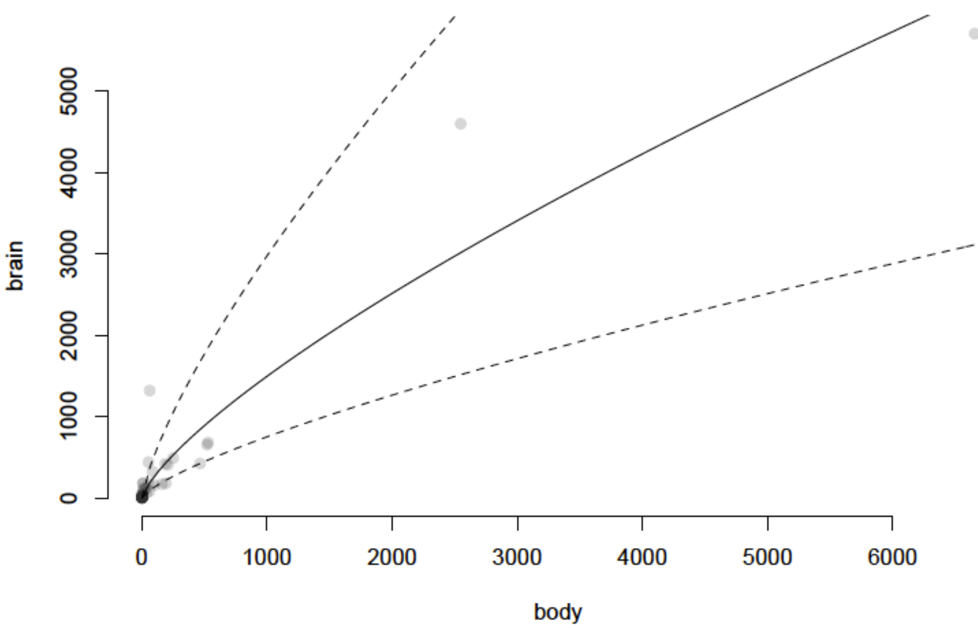
$$y^* = e^{\hat{B}_0} x_*^{\hat{B}_1} e^{\hat{\sigma}_\varepsilon}$$

- **Lower:**

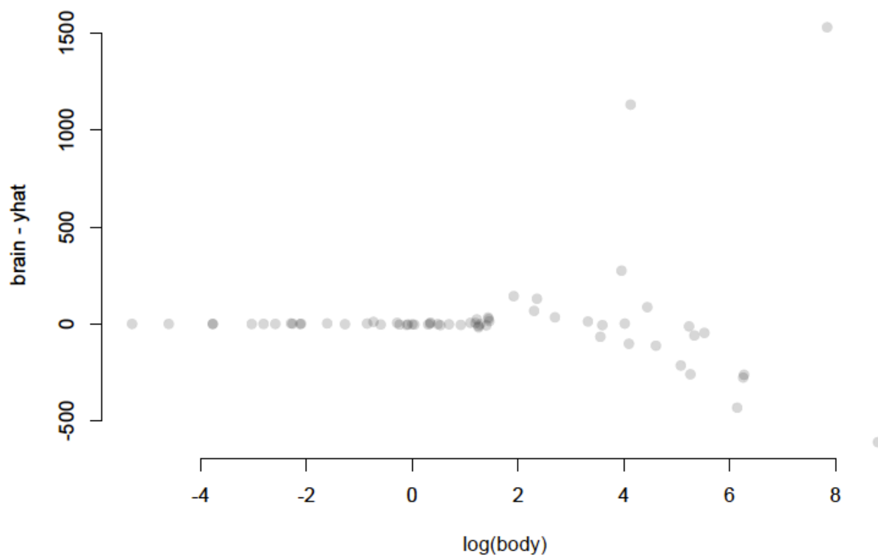
$$\log y^* = \hat{B}_0 + \hat{B}_1 \log x_* - \hat{\sigma}_\varepsilon$$

$$y^* = e^{\hat{B}_0} x_*^{\hat{B}_1} e^{-\hat{\sigma}_\varepsilon}$$

- From the graph, you can see that when the confidence intervals are plotted, they "fan out":



- Bigger errors as you move to the right.
- When you graph the residuals, you can see there are more errors with a larger body size:



Transformation:

1. Estimate everything from model on log scale
2. Transform last

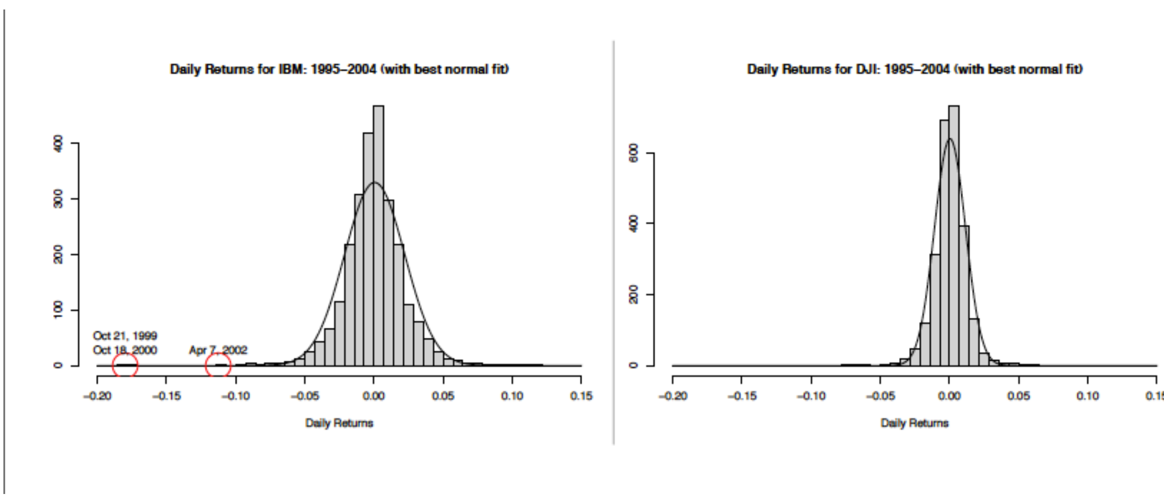
Regression:

- Sampling distribution: distribution of samples fit in linear line
- Confidence in your estimates → stability of those estimates under influence of chance
- Bootstrapping: Think of sample as the population, and then resample from sample with **replacement**.
  - Two criteria:
    - Sample needs to be of the same size as the original sample
    - Sample with replacement or else you will have no variability
  - The goal of bootstrapping is to replicate the variability of the original population
- Residuals also follow normal distribution
  - Put in: (Pg. 111; equations 4.6 & 4.7)

$$\begin{aligned}\widehat{\beta}_0 &\sim \mathrm{N}\left(\beta_0, \sigma_0^2\right)\\ \widehat{\beta}_1 &\sim \mathrm{N}\left(\beta_1, \sigma_1^2\right).\end{aligned}$$

- Get out: errors bars (sampling distribution from B_0 & B_1)

- Residuals are an aggregate of nudges ("stuff" left out of the model) that you cannot forecast using variable x.
  - Mad libs example
    - Aggregate of positive and negative words to move the data point away from the line
    - Coin flip to decide whether to use a positive or negative word = binomial distribution
      - Good, we moved up one, bad we moved down one
    - We discovered the actual finished point did not deviate too much from the line.
- **Conclusion:** Cumulative effect of residuals is an aggregate of nudges that are described using a **binomial distribution (**if there are enough "nudges," a normal distribution will form)
  - Ex: NASDAQ vs. IBM
    - NASDAQ is a better representation of the trend of the economy because it is an aggregate of different company's stocks put together.
    - IBM is just one "nudge," so it will not fit the normal distribution
    - As we add more nudges the distribution of residuals becomes more normal and less variable



  - Ex: School of fish vs. shark
    - School of fish is an aggregate of individual fish acting together, which will create a normal distribution
    - Shark is just the movement of one fish

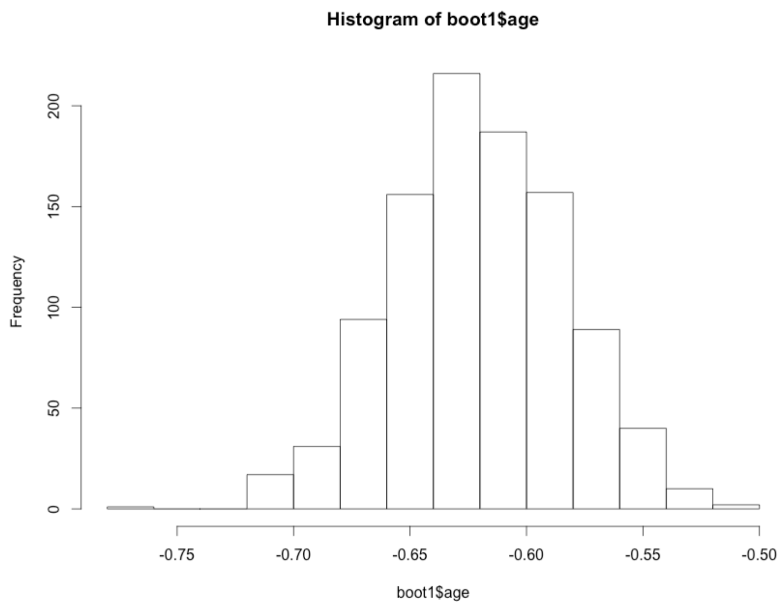
<u>Creatine R. & Creatine csv.</u>

- This data set shows the effectiveness of the kidney by measuring creatine difference between the kidney and urine.
- "summary(lm1)"

- Makes assumption of normal distribution of residuals and gives you standard errors:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 103.5040    63.9207   1.619    0.144
volume        3.9652     0.5028   7.886 4.84e-05
```

- "boot1 = do(1000)*lm(creatclear~age, data=resample(creatinine))"
  - Perform bootstrapping 1000 times
  - Standard error of histogram:



hist(boot1$age)
sd(boot1$age) → [1] 0.03702145

  - Check that standard error and standard deviation are similar

Bootlegging vs. Naïve Prediction Intervals

- Naïve prediction intervals:
  - Ignore uncertainty about parameters
  - Should be wider by the value of standard error to take into account the systematic components that have error

Takeaway Lesson

- Using bootlegging incorporates residual uncertainty as well as uncertainty about predictors (b_0 and b_1)
- The goal of bootstrapping is to replicate the variability of the original population
- As you use bootlegging, you should assume that your residuals follow a normal distribution