Stacy Liu-sl26667
Anne Theil- at25936
STA 371H 9:30-11

February 3, 2014 Notes

Today we discussed the concept of sampling distribution with an example involving the whole class. After combining our separate data (samples of 10 fish) we were able to draw conclusions on the population of fish. We used the R command `do(10)*lm(weight~volume, data=sample(gonefishing,nsamp))` to calculate 10 different outcomes (or however many you want) using `nsamp` number of samples. We also discussed the idea of bootstrapping and Montecarlo simulations.

1.) Homework 3

1a.) After graphing the preliminary function, convert it to a logarithmic function, plotting a line of best-fit and finding the beta coefficients.From the beta weights from the logarithmic function, we got an exponential equation
1b.) In order to adjust for body size, take the residuals of the original dataset's equation and plot the body size versus the residuals of the brain size.
1c.) In order to find the 95% confidence interval, we used the predict function which gave us the (65.366, 1096.633) interval.

2.) We chose $yi = \beta0 + \beta1xi + \beta2xi2 + \beta3xi3 + ei$ since it looked like it fit better than the $yi = \beta0 + \beta1xi + \beta2xi2 + ei$ and it was simpler than $yi = \beta0 + \beta1xi + \beta2xi2 + \beta3xi3 + \beta4xi4$ + ei. Again, we used the predict function to find the 95% interval

2.) Activity - Fish Bucket

- Form into groups of 2 and get 10 fish
- Enter fish data into excel and import into RStudio
- Find a linear model to predict volume of fish
- Enter all our intercepts and slopes into the class data

```
#FISH
volume = (FISH$Weight)*(FISH$Length)*(FISH$Height)
plot(Weight~volume, data=FISH)
lm1=lm(Weight~volume, data=FISH)
coef(lm1)
abline(lm1)
```

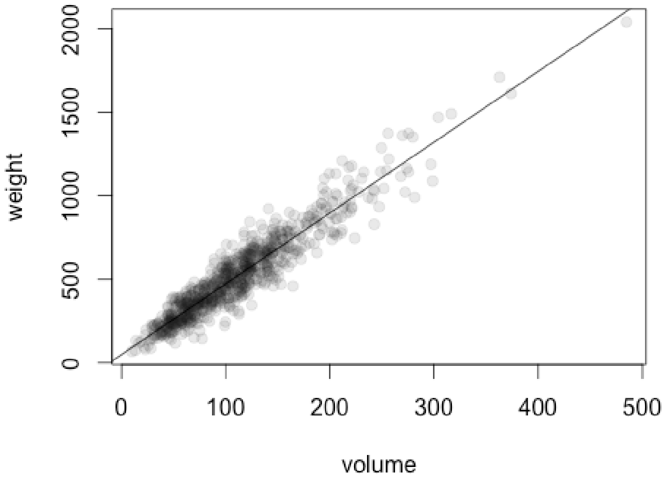
- Theory behind what we are doing: True State of the World

  Yi = b0 + b1Xi +error

  Our sample: (Xi, Y1), (X2, Y2),...., (Xn, Yn) with n=10. We fit a linear model to our sample data and obtained parameter estimates, b(hat)0 and b(hat)1. When we combined all of our parameter estimates and fit a model to it, we got a **sampling distribution** of b(hat)1 and b(hat)0. **We aim to know more about the population from our samples.**

- Sampling Distribution b(hat)1: We aggregated all the b(hat)1s and made a histogram from them. We noticed a higher frequency around the center.

3.) Gone fishing



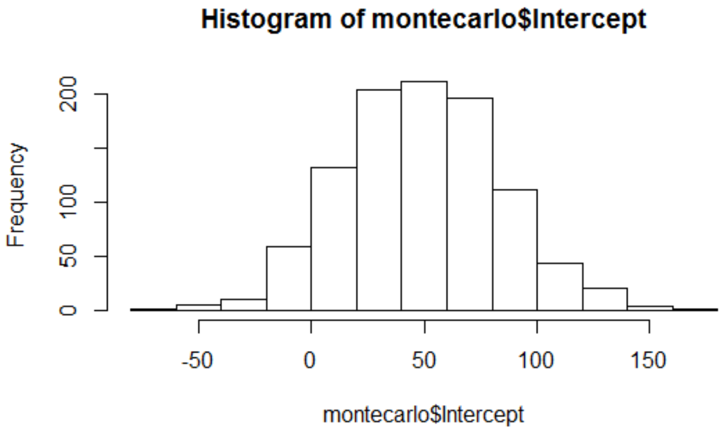
- We can use RStudio to take thousands of samples from the population.

```
# First define the sample size
nsamp = 30 (sample that you are taking from the population)

# Try taking a sample a few different times
lmsamp = lm(weight~volume, data=sample(gonefishing,nsamp))
coef(lmsamp)

# We can automate the process of taking multiple samples
do(1000)*lm(weight~volume, data=sample(gonefishing,30))
```

- Next, we created a histogram of all the intercepts of the 1000 samples.



- This model allowed us to report uncertainty about B1 by being able to create a confidence interval using the sampling distribution.

- We can also compute **standard error** using standard deviation of the estimates of the parameter

```
sd(montecarlo$volume)
```

- **Montecarlo simulation**: gives you a range of possible outcomes and the probabilities they will occur for any choice of action
- We can check if the estimator is **biased** by taking the mean of each column

```
colMeans(montecarlo)
```

- To extract a 95% confidence interval (also know as the prediction interval):

```
confint(montecarlo, level=0.95)
```

4.) What if we only have one sample?

- **Bootstrapping**: In this case, the **sample** is the **population**
  - Take samples from our sample
  - Sample with replacement of samples

```
# Try a single bootstrapped sample from your sample
lmboot = lm(weight~volume, data=resample(myfishingtrip))
coef(lmboot)
#This coef varies from the coef from just our sample
# Now 1000
myboot = do(1000)*lm(weight~volume, data=resample(myfishingtrip))
```

We can also find a confidence interval from our bootstrapped samples. The true population intercept is always within that confidence interval

```
# Coverage interval from the bootstrapped samples
confint(myboot, level=0.95)
```