Jackson Simons Scribe 2/12/14 STA371H
JMS7288 MW 930-11

T- Statistic

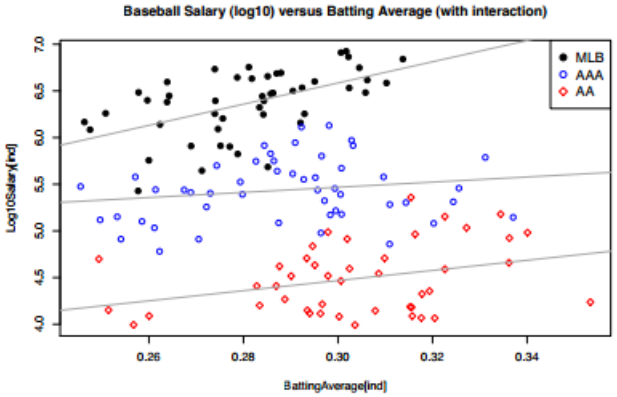
- Often quoted statistic that stands for Test Statistic
- It can be basically thought of as a "signal to noise" ratio
- The basic form of the t statistic is (Sample Statistic – Population parameter (or what you believe it to be)/ standard deviation.
- Since you would do this test if you knew the true population parameter, you are usually using a hypothesis for it
- You can also use find a tstatistic, or t^* ,based off a desired coverage interval and then use it to find a confidence interval for the true parameter given your sample statistic.
- $\hat{\beta_1}$ in the linear regression model is the coefficient for slope
- $\hat{\beta_1}$ is a variable that contains randomness since it is based off a sampling distribution
  - This means that if we took a different sample from the population we would get a different Beta
- We can find the standard error of $\hat{\beta_1}$ by multiplying it with the standard deviation of the sample $\hat{\sigma}$
- We can then use a Confidence Interval, using the t statistic, to get an idea of where the real (population) β_1
- ($\hat{\beta_1}$-$t^*\hat{\sigma}$, $\hat{\beta_1}$+$t^*\hat{\sigma}$) ← Confidence Interval for Beta1
- While this is useful in the linear regression model, T is better used as a quick and dirty way of determining if there is a relationship at all.
- If we assume there is no relationship in the population then we get that t_1=$\hat{\beta_1}/\hat{\sigma}$
- We can then divide through our original confidence interval by $\hat{\sigma}$ to get (t_1-t^* , t_1+t^*)
- Using this method if $|t_1|>t^*$ we can say with the level of confidence of our t statistic that there is a relationship between the X and Y variables used to find the linear regression
  - This idea results from the fact that if that relationship is true then 0 (no relationship) is not contained within upper and lower values of our confidence interval

Confidence Intervals

- Although we can simply define it as how confident we are that a parameter is in a range, there is also a certain technical meaning behind the interval
- Imagine an assembly line being sampled for quality control purposes
  - Of the widgets made 95% are okay 5% are defective
  - Of the ones not sampled how confident are we that the ones in the real world will work: 95%
  - All we have to go on is sampling method
  -
- As we take samples from a population, we receive different statistics and a different variation which leads to varying confidence intervals of the same level
- $\propto (alpha)$- is how much error we will tolerate.

- Taking our assembly line example, if we aggregate all the confidence intervals from our samples and:
  - If (1-∝) of the intervals contain the true value then this assembly line produces (1-∝) confidence intervals
  - Using our example of 95%, alpha would equal 5% and if 95% of the intervals contained the true value then the assembly line is producing 95% confidence intervals
- ^This property described above is called the **Frequentist Coverage Property**
- If sample is biased and not normal (residuals not normally distributed) then because of the Frequentist Coverage Property confidence intervals probably would not be true confidence intervals

Aggregation Paradox



- The aggregation paradox deals with how data looks different when distinct groups are aggregated together and presented
- In the example above we see that when aggregated together, there actually is a seeming negative relationship between batting average and salary for pro baseball players which makes no sense
- However, once we point out the groupings (MLB, AAA, AA three different levels of baseball with corresponding pay grades) we see that there is a positive relationship within each group
- This is idea is exactly what the aggregation paradox is pointing out
- We similarly saw this idea in school admittance with women and men
  - Just looking at the rate of admittance and gender we see that men are being admitted at a much higher rate
  - This leads us to believe there is sexual discrimination
  - However when we disaggregate the genders and look at it by department applied, we see that men are in reality just apply in higher numbers to departments with higher acceptance rates
  - This aggregation then causes us to think that men are being favored because of their higher rate when that is not the reality

Dummy Variables and Interactions

- Looking back at the baseball salary data, we see that the grouping can affect the linear regression model for each group in two ways: Slope and intercept
- This gives us two ways to look at it: Dummy Variable Model and Interaction Model
- In the Dummy Variable model we just look at the shift of intercept between groups (the slope remains the same
  - $Y_i = response, X_{i1} = grouping\ variable, X_{i2} = continuous\ variable$
  - Looking at each group individually we get two linear models with two different intecepts
  - $Y_i = \beta_0^{(Group\ 0)} + \beta_2 X_{i2} + E_i$
  - $Y_i = \beta_1^{(Group\ 1)} + \beta_2 X_{i2} + E_i$
  - As we can see these models have the same slope but different intercepts
  - Combining these modles:
  - $Y_i = \beta_0 + \beta_1 \ \Uparrow \{X_{i1} = 1\} + \beta_2 X_{i2} + E_i$
    - where $\Uparrow \{X_{i1} = 1\}$ means 1 if Xi1=1 and 0 if not
    - This Beta 1 + Beta 2 results in the shift of intercept depending on the grouping
- The second model we see is the Interaction model where slope and intercept both change depending on the grouping
  - $Y_i = \beta_0^{(Group\ 0)} + \beta_2^{Group\ 0} X_{i2} + E_i$
  - $Y_i = \beta_1^{(Group\ 1)} + \beta_3^{Group1} X_{i2} + E_i$
  - Shows different slopes and intercepts ^