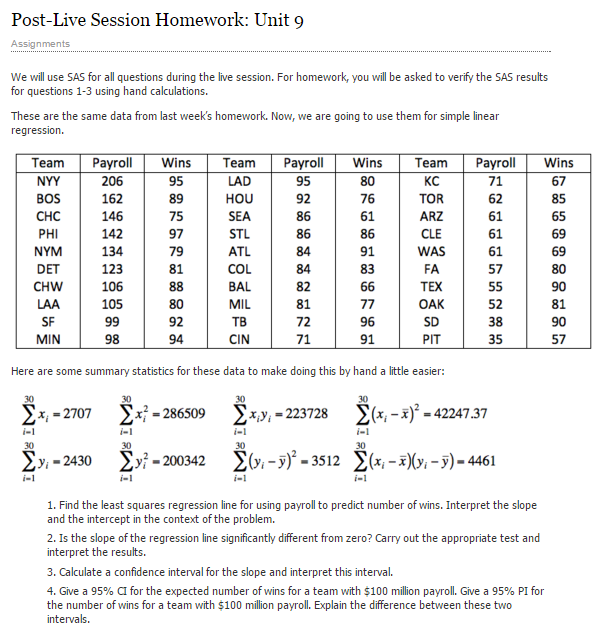
UNIT 9 HW

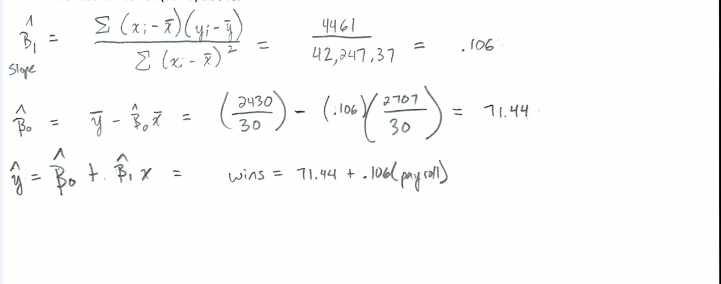
These are the same data from last week’s HW. Now, we are going to use them for simple linear regression.



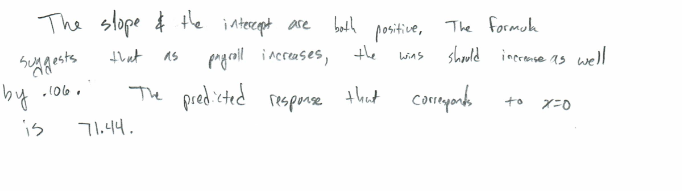
1)

a.

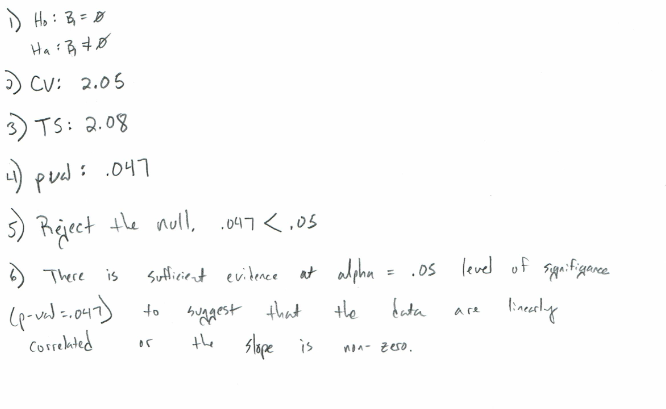
i. Find the least squares regression line using payroll to predict the number of wins. Interpret the slope and the intercept in the context of the problem. Show your work in finding the slope and intercept. You will need the above calculations. Do this by hand or using a basic calculator, but **NOT** by uploading the data into software. There are several equivalent formulations for the elements of the least squares regression line ( and ). Find one that utilizes the series (sums) above.



ii. Interpret the slope **AND** the intercept in the context of the problem.

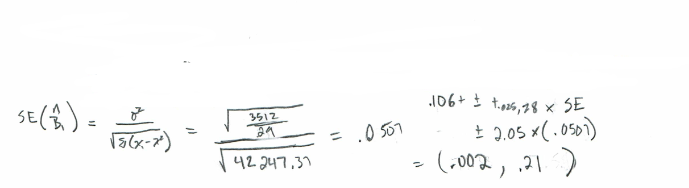


b. Is the slope (only concerned with the slope here) of the regression line significantly different from zero? Carry out a 6-step hypothesis test to address this question. Use the above calculations to find the relevant statistics for this test. You will need to use SAS, R, the internet, a calculator, or integration to find the p-value and critical value, but do NOT upload the data to software to find the p-value. (One of the first 4 choices is suggested. ☺) Use α = 0.05.

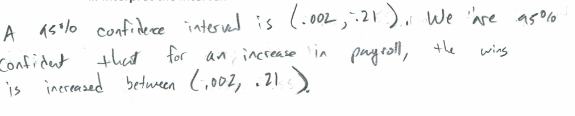


c.

i. **BY HAND (**or basic calculator), calculate a 95% confidence interval for the slope. You should already have the pieces of the confidence interval (point estimate, multiplier, and standard error) from part 1b.



ii. Interpret the interval.

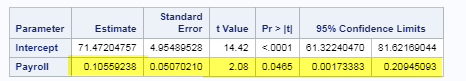


d. Verify your results (parameter estimates, test statistic for the hypothesis test of whether the slope equals zero, p-value for this same hypothesis test, and confidence interval for the slope) with SAS. Paste your code and relevant output below. Note what is the same or different.

proc glm data = work.baseball;

model Wins = Payroll/clparm;

run;



My t-val, p-val and confidence interval is the same as the handwritten results.

2)

a.

i. Find the least squares regression line to assess the relationship between the math and the science score for the Test Data. We would like to be able to estimate a change in the mean math score for a one point change in the mean science score. (This should help identify the response and the independent variables.) Write your regression equation and paste your code and relevant output below. You should obtain the test statistics and other relevant statistics from R.

Code:

library(xlsx)

setwd("C:/Users/Marin Family/Desktop/Statistical Foundations for Data Science/Unit 9")

testdata <- read.xlsx("TestData.xlsx","Test Data")

criticalValue <- as.data.frame(qt(c(.005,.995), df = n-2))

cv <- criticalValue[2,1]

cv

testdatalm <- lm(math~science, data = testdata)

testdatalm

summary(testdatalm)

Output:

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 21.70019 2.75429 7.879 2.15e-13 \*\*\*

science 0.59681 0.05218 11.437 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.288 on 198 degrees of freedom

Multiple R-squared: 0.3978, Adjusted R-squared: 0.3948

F-statistic: 130.8 on 1 and 198 DF, p-value: < 2.2e-16

>

> confint(testdatalm, level = .99)

0.5 % 99.5 %

(Intercept) 14.536591 28.8637921

science 0.461094 0.7325341

Regression Line = MathScore = .60(Science) + 21.7

t-val of Intercept = 7.879

t-val of Slope = 11.437

p-values of Slope and Intercept is less than <.001

ii. Interpret the slope and the intercept in the context of the math and science scores.

The slope and the intercept are both positive. The regression line suggest that as Science Scores increase, the math scores increase as well by .60. The predicted response that corresponds to x=0 is 21.7.

b. Are the slope ***and intercept*** of the regression line significantly different than zero? Carry out a 6-step hypothesis test **for each** regression parameter to address this question (two different hypothesis tests). You should obtain the test statistics and other relevant statistics from R. Paste your code and any relevant output below. Use alpha = 0.01.

Code:

library(xlsx)

setwd("C:/Users/Marin Family/Desktop/Statistical Foundations for Data Science/Unit 9")

testdata <- read.xlsx("TestData.xlsx","Test Data")

testdatalm <- lm(math~science, data = testdata)

testdatalm

summary(testdatalm)

confint(testdatalm, level = .99)

n <- as.numeric(nrow(testdata))

criticalValue <- as.data.frame(qt(c(.005,.995), df = n-2))

cv <- criticalValue[2,1]

cv

cortest <- cor.test(testdata$science, testdata$math, conf.level = .99)

r <- as.numeric(cortest$estimate)

testStat <- (r\*sqrt(n-2))/(sqrt(1-r^2))

pval <- as.numeric(cortest$p.value)

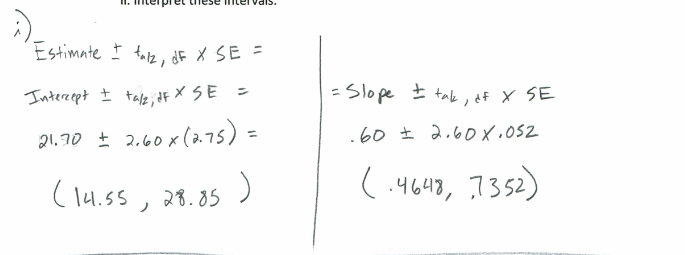
Estimate Std. Error t value Pr(>|t|)

(Intercept) 21.70019 2.75429 7.879 2.15e-13 \*\*\*

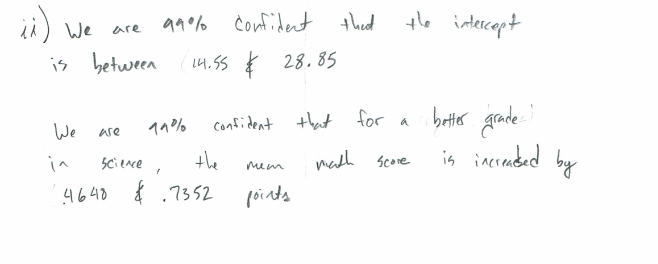
science 0.59681 0.05218 11.437 < 2e-16 \*\*\*

|  |  |
| --- | --- |
| Step 1:  Ho: B1 = 0  Ha: B1 <> 0  Step 2: Crit val = 2.60  Step 3: T-Value =  (Parameter estimate – 0)/StErr =  (.597 -0)/ 0.05218  =11.44  Step 4: P-val = <.001  Step 5: Reject the Null  Step 6: There is sufficient evidence at alpha = .01 level of significance (p-val= < .001) to suggest that the data are linearly correlated or the slope is nonzero. | Ho: B0 = 0  Ha: B0 <> 0  Step 2: Crit val = 2.60  Step 3: T-Value:7.88  Step 4: P-val = <.001  Step 5: Reject the Null  Step 6: There is sufficient evidence at alpha = .01 level of significance (p-val= < .001) to suggest that the data are linearly correlated or the intercept is nonzero. |

c.

**i. BY HAND**, calculate 99% confidence intervals for the slope and intercept (**two** separate confidence intervals). You may use point estimates, multipliers, and standard errors found from software, but put these pieces together to form confidence intervals by hand (or basic calculator). 

ii. Interpret these intervals.



d. Verify your confidence intervals (for and ) with R and paste your code and relevant output below.

Code:

library(xlsx)

setwd("C:/Users/Marin Family/Desktop/Statistical Foundations for Data Science/Unit 9")

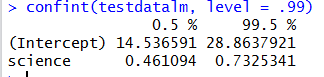
testdata <- read.xlsx("TestData.xlsx","Test Data")

testdatalm <- lm(math~science, data = testdata)

testdatalm

summary(testdatalm)

confint(testdatalm, level = .99)



Confidence intervals are about the same as what I had.

BONUS:

3) Repeat 1(d) using R.

Code:

setwd("C:/Users/Marin Family/Desktop/Statistical Foundations for Data Science/Unit 8/Download")

library(sqldf)

baseball <- read.csv("Baseball\_Data.csv")

#baseball <- sqldf("select \* from baseball where Team <> 'TEX'")

n <- as.numeric(nrow(baseball))

criticalValue <- as.data.frame(qt(c(.025,.975), df = n-2))

cv <- criticalValue[2,1]

cv

baseballlm <- lm(Wins~Payroll, data = baseball)

baseballlm

summary(baseballlm)

Step 1:

Ho: B1 = 0

Ha: B1 <> 0

Step 2: Crit val = 2.04

Step 3: T-Value = 2.083

Step 4: P-val = .04654

Step 5: Reject the Null

Step 6: There is sufficient evidence at alpha = .05 level of significance (p-val= < .05) to suggest that the data are linearly correlated or the slope is nonzero.

4) Repeat 2(a)(i) and 2(d) using SAS.

2a)

i)

Code:

PROC IMPORT OUT= WORK.TestData

DATAFILE= "/home/marinfamily1010/sasuser.v94/Data/TestData.xlsx"

DBMS=xlsx REPLACE;

GETNAMES=YES;

DATAROW=2;

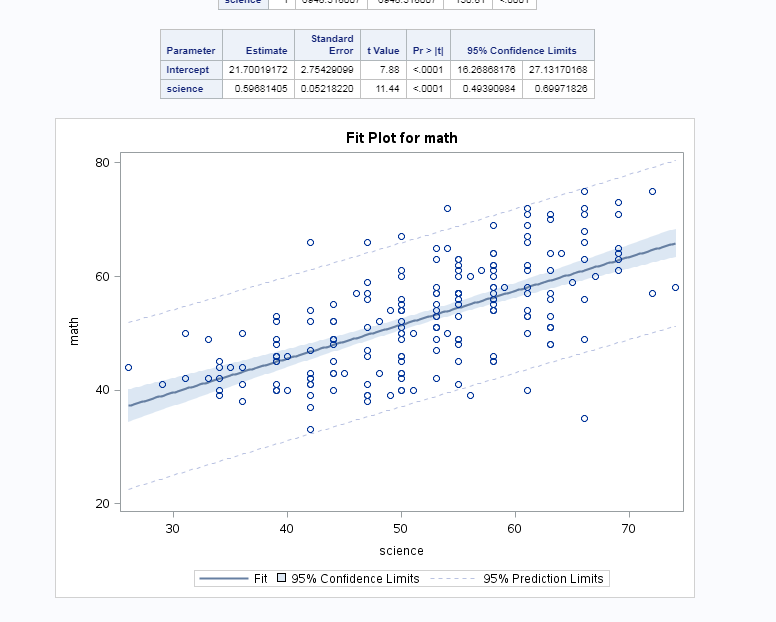
RUN;

proc glm data = work.TestData;

model math = science/clparm;

run;

Output:



Regression Line = MathScore = .60(Science) + 21.7

T-val = 1.44

P-val = <.0001

95% conf int: (.49,.69)

Standard Error: .0521

2d:

Code

PROC IMPORT OUT= WORK.TestData

DATAFILE= "/home/marinfamily1010/sasuser.v94/Data/TestData.xlsx"

DBMS=xlsx REPLACE;

GETNAMES=YES;

DATAROW=2;

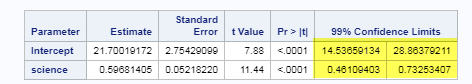
RUN;

proc glm data = work.TestData;

model math = science/clparm alpha= .01;

run;

Output:



99% Confidence Intervals highlighted above.

5) We will cover this in Unit 10 ….

With reference to the baseball data …

1. Give a 95% CI (confidence interval) for the expected number of wins for a team with $100 million payroll. Use SAS or R.
2. Give a 95% PI (prediction interval) for the number of wins for a team with $100 million payroll. Use SAS or R.
3. Explain the difference between these two intervals.