UNIT 9 HW

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

Team	Payroll	Wins	Team	Payroll	Wins	Team	Payroll	Wins
NYY	206	95	LAD	95	80	KC	71	67
BOS	162	89	HOU	92	76	TOR	62	85
CHC	146	75	SEA	86	61	ARZ	61	65
PHI	142	97	STL	86	86	CLE	61	69
NYM	134	79	ATL	84	91	WAS	61	69
DET	123	81	COL	84	83	FA	57	80
CHW	106	88	BAL	82	66	TEX	55	90
LAA	105	80	MIL	81	77	OAK	52	81
SF	99	92	TB	72	96	SD	38	90
MIN	98	94	CIN	71	91	PIT	35	57

Here are some summary statistics for these data to make doing this by hand a little easier:

$$\sum_{i=1}^{30} x_i = 2707 \qquad \sum_{i=1}^{30} x_i^2 = 286509 \qquad \sum_{i=1}^{30} x_i y_i = 223728 \qquad \sum_{i=1}^{30} (x_i - \bar{x})^2 = 42247.37$$

$$\sum_{i=1}^{30} y_i = 2430 \qquad \sum_{i=1}^{30} y_i^2 = 200342 \qquad \sum_{i=1}^{30} (y_i - \bar{y})^2 = 3512 \qquad \sum_{i=1}^{30} (x_i - \bar{x})(y_i - \bar{y}) = 4461$$

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 ($\widehat{\beta_1}$ and $\widehat{\beta_0}$). Find one that utilizes the series (sums) above.

$\widehat{\beta_1}$ = *SP/SS$_X$* = 4461/42247.37 = 0.10559

$\widehat{\beta_0}$= 2430/30 - (0.11*(2707/30)) = 81- (.11*90.23)= 71.47205

wins =71.47+ 0.10559*(payroll)*

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

Slope: for each 25 million dollars spent on payroll, we expect on average that the wins will increase by .10559.

Y-Intercept: the predicted wins for a team with no payroll is 71.47.

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. (One of the first 4 choices is suggested. ☺) Use α = 0.05.

$H_o: \beta_1 = 0 \;\; vs. H_A: \beta_1 \neq 0$

Critical Value: qt(0.975, 28, lower.tail=T) = 2.048

t =2.08

p=0.0465

There is evidence at the α=0.05 level of significance (p=0.0465) to suggest that wins and payroll are linearly correlated.

The estimated regression line is wins =71.47+ 0.10559(payroll)

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.

71.47 ±4.9549*2.048= [61.32, 81.62]

ii. Interpret the interval.

For a team with no payroll it is predicted that it will win (71.47 wins from regression equation.) A 95% confidence interval is (61.32, 81.62).

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 import datafile = '/home/chec0/New Folder/Baseball_Data.csv'

out = baseball

```
dbms = CSV

 ;

 proc print data = baseball;

run;

proc reg data= baseball;

model Wins = Payroll / clb;

run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: Wins

Number of Observations Read	30
Number of Observations Used	30

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	471.04761	471.04761	4.34	0.0465
Error	28	3040.95239	108.60544		
Corrected Total	29	3512.00000			

Root MSE	10.42139	R-Square	0.1341
Dependent Mean	81.00000	Adj R-Sq	0.1032
Coeff Var	12.86592		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	71.47205	4.95490	14.42	<.0001	61.32240	81.62169
Payroll	1	0.10559	0.05070	2.08	0.0465	0.00173	0.20945

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.

scores <- read.xlsx('/TEST DATA.xlsx')

scores.lm <- lm(scores$math ~ scores$science)

summary(scores.lm)

```
Call:
lm(formula = scores$math ~ scores$science)

Residuals:
     Min       1Q   Median       3Q      Max 
-26.0899  -5.0044   0.4671   4.6886  19.2336 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    21.70019    2.75429   7.879 2.15e-13 ***
scores$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
```

math=21.70019+ 0.59681(science)

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

Slope: for each one point change in the mean science score, we expect on average that the math score will increase by .59681.

Y-Intercept: the predicted math score for a student with no science score is 21.70019 points.

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.

Slope

$H_o: \beta_0 = 0$ vs. $H_A: \beta_0 \neq 0$

Critical Value: qt(0.995, 198, lower.tail=T) = 2.60089

t =11.44

p=<.0001

There is sufficient evidence at the α=0.05 level of significance (p=0<.0001) to suggest that science and math means are linearly correlated.

The estimated regression line is math=21.70019+ 0.59681(science).

Intercept

$H_o: \beta_1 = 0$ vs. $H_A: \beta_1 \neq 0$

Critical Value: qt(0.995, 198, lower.tail=T) = 2.60089

t =7.88

p=<.0001

There is sufficient evidence at the α=0.05 level of significance (p=0<.0001) to suggest that math and science means are linearly correlated.

The estimated regression line is math=21.70019+ 0.59681(science).

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).

Slope= .59681±.05218*2.60089 = [.461096, .732524]

Intercept= 21.70019±2.75429*2.60089= [14.53658, 28.8638]

ii. Interpret these intervals.

For every 1 point change in science mean, the estimated math mean increases (on average) .59681. We are 99% confident that this value is between .491096 and .732524.

If the science score is 0, the estimated mean of math score is 21.70019. We are 95% confident that the intercept is between 14.53658 and 28.8638.

d. Verify your confidence intervals (for β_1 and β_0) with R and paste your code and relevant output below.

scores <- read.xlsx('/TEST DATA.xlsx')

scores.lm <- lm(scores$math ~ scores$science)

summary(scores.lm)

confint(scores.lm, level = 0.99)

```
                    0.5 %     99.5 %
(Intercept)     14.536591 28.8637921
scores$science   0.461094  0.7325341
> summary(scores.lm)

Call:
lm(formula = scores$math ~ scores$science)

Residuals:
     Min       1Q   Median       3Q      Max 
-26.0899  -5.0044   0.4671   4.6886  19.2336 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    21.70019    2.75429   7.879 2.15e-13 ***
scores$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
```

BONUS:

3) Repeat 1(d) using R.

baseball <- read.csv('/Baseball_Data.csv')

baseball.lm <-lm(baseball$Wins ~ baseball$Payroll)

summary(baseball.lm)

confint(baseball.lm)

```
Call:
lm(formula = baseball$Wins ~ baseball$Payroll)

Residuals:
    Min      1Q  Median      3Q     Max
-19.553  -8.340   1.099   9.301  16.925

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       71.4720     4.9549  14.425 1.73e-14 ***
baseball$Payroll   0.1056     0.0507   2.083   0.0465 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.42 on 28 degrees of freedom
Multiple R-squared:  0.1341,	Adjusted R-squared:  0.1032
F-statistic: 4.337 on 1 and 28 DF,  p-value: 0.04654

> confint(baseball.lm)
                       2.5 %     97.5 %
(Intercept)      61.32240470 81.6216904
baseball$Payroll  0.00173383  0.2094509
```

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

```
proc import datafile = '/home/chec0/New Folder/TEST DATA.xlsx'

 out = test

 dbms = xlsx

 ;


proc print data = test;

run;

proc reg data= test;

model math=science / clb;

run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: math

Number of Observations Read	200
Number of Observations Used	200

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6948.31801	6948.31801	130.81	<.0001
Error	198	10517	53.11857		
Corrected Total	199	17466			

Root MSE	7.28825	R-Square	0.3978
Dependent Mean	52.64500	Adj R-Sq	0.3948
Coeff Var	13.84414		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	21.70019	2.75429	7.88	<.0001	16.26868	27.13170
science	1	0.59681	0.05218	11.44	<.0001	0.49391	0.69972

proc reg data= test

alpha = .01;

model math=science / clb;

run;

Coeff Var	13.84414		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	99% Confidence Limits	
Intercept	1	21.70019	2.75429	7.88	<.0001	14.53659	28.86379
science	1	0.59681	0.05218	11.44	<.0001	0.46109	0.73253

5) We will cover this in Unit 10 ….

With reference to the baseball data … we will learn how to do the following next week.

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