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:

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

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

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.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{4461}{42247.37} = 0.105592$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 81 - (0.105592 \times 90.23333) = 71.47205$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \;\rightarrow\; wins = 71.47205 + 0.105592(payroll)\,; \quad df = n - parameters = 30 - 2 = 28$$

$$\sum_{i=1}^{30}(y-\hat{y})^2 = 3040.9523 \qquad \hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{30}(y-\hat{y})^2}{df}} = \sqrt{\frac{3040.9523}{28}} = 10.42139$$

$$SE(\hat{\beta}_1) = \hat{\sigma}\sqrt{\frac{1}{(n-1)s_x^2}} = 10.42139\sqrt{\frac{1}{(30-1)1456.806}} = 0.050702$$

$$SE(\hat{\beta}_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}} = 10.42139\sqrt{\frac{1}{30} + \frac{90.233^2}{(30-1)1456.806}} = 4.954895$$

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

The y-intercept of the model indicates that should a team be paid the minimum allotted by the MLB, the mean games won would be 71.47205. The slope indicates that for every additional \$1M on top of this minimum, the mean additional games won will be 0.105592 or for every \$10M more spent in payroll, the mean additional games won will be 1.005592 more.

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.

1. $H_0: \rho = 0 \qquad \beta = 0$
   $H_a: \rho \neq 0 \qquad \beta \neq 0$
2. Critical Value: ± 2.048
3. $t_{statistic} = t_{0.975,28} = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = 2.082604$
4. p value = 0.0465
5. Reject H_0
6. There is sufficient evidence at the alpha = 0.05 level of significance (p-value = 0.0028) to suggest that the data are linearly correlated.

c.

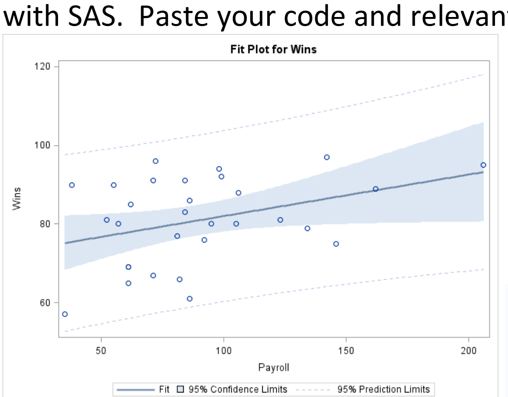
i. **BY HAND (**or basic calculator), calculate a 95% confidence interval for the slope. You should already have the pieces of the confidence interval (point estimate, multiplier, and standard error) from part 1b.

$\hat{\beta}_1 \pm t_{0.025,28} \times SE(\hat{\beta}_1) = 0.105592 \pm 2.048 \times 0.050702 =$ [0.20943, 0.001754]

ii. Interpret the interval.

We are 95% confident that when the payroll is increased by \$1M, the mean games won increases between 0.20943 and 0.001754.

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 = baseball;
  model wins=payroll / clparm;
  run;
```

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	71.47204757	4.95489528	14.42	<.0001	61.32240470	81.62169044
Payroll	0.10559238	0.05070210	2.08	0.0465	0.00173383	0.20945093

In this instance, the numbers match for the most part. Except, there appears to be some rounding error once we get into the thousandth values. The intercept, slope, SEs, t-statistic, and p-value all match. Even the confidence intervals match as well. However, there is slight difference as mentioned prior.

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.

```
> testdatalm <- lm(Test.Data$math ~ Test.Data$science, data = Test.Data)
> testdatalm

Call:
lm(formula = Test.Data$math ~ Test.Data$science, data = Test.Data)

Coefficients:
      (Intercept)  Test.Data$science
          21.7002             0.5968
```

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad \rightarrow \quad mean\ math\ score = 21.7002 + 0.5968(science\ score)$

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

The intercept value indicates that in the event the student gets the lowest score possible on the science test (26 in this data set), the mean math test score is going to be 21.7002. Yet, for every point increased in the science test score, the mean math score will increase by 0.5968 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.

1. $H_0: \rho = 0$
   $H_a: \rho \neq 0$
2. Critical Value: ± 2.763
3. $t_{statistic} = t_{0.995,28} = 11.437$
4. p-value = <0.001
5. Reject null
6. There is sufficient evidence at the alpha = 0.01 level of significance (p-value < 0.001) to suggest that the data are linearly correlated.

1. $H_0: \beta_0 = 0$
   $H_a: \beta_0 \neq 0$
2. Critical Value: ± 2.763
3. $t_{statistic} = t_{0.995,28} = 7.879$
4. p-value = <0.001
5. Reject null
6. There is sufficient evidence at the alpha = 0.01 level of significance (p-value < 0.001) to suggest that the y-intercept is not equal to zero.

```
> testdatalm <- lm(Test.Data$math ~ Test.Data$science, data = Test.Data)
> testdatalm

Call:
lm(formula = Test.Data$math ~ Test.Data$science, data = Test.Data)

Coefficients:
      (Intercept)  Test.Data$science
          21.7002             0.5968

> summary(testdatalm)

Call:
lm(formula = Test.Data$math ~ Test.Data$science, data = Test.Data)

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 ***
Test.Data$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
```

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).

$\hat{\beta}_1 \pm t_{0.995,28} \times SE(\hat{\beta}_1) = 0.59681 \pm 2.763 \times 0.05218 = [0.461094, 0.7325341]$

$\hat{\beta}_0 \pm t_{0.995,28} \times SE(\hat{\beta}_0) = 21.70019 \pm 2.763 \times 2.75429 = [14.536591, 28.8637921]$

ii. Interpret these intervals.

We are 99% confidence that for every additional point in the science test, the mean math test score will increase between 0.46 and 0.73 points.

We are 99% confident that the intercept is between 14.5 and 28.9 points.

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

```
> confint(testdatalm, level = 0.99)
                     0.5 %     99.5 %
(Intercept)      14.536591 28.8637921
Test.Data$science  0.461094  0.7325341
```

BONUS:

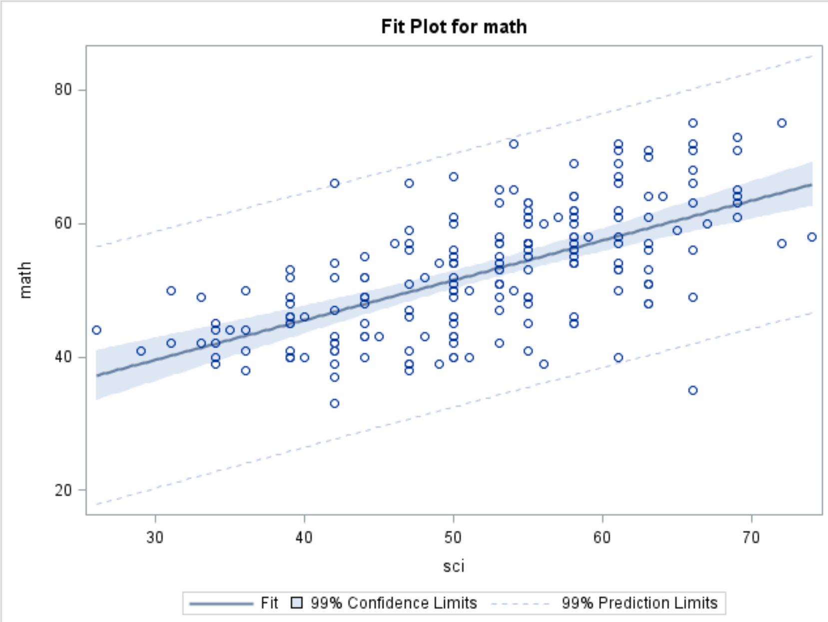
3) Repeat 1(d) using R.

As mentioned in the prior answer, the results are pretty identical with some slight variance in the thousandth decimal values.

```
> baseballlm <- lm(Baseball_Data$Wins..y. ~ Baseball_Data$Payroll..x.,
data = Baseball_Data)
> baseballlm

Call:
lm(formula = Baseball_Data$Wins..y. ~ Baseball_Data$Payroll..x.,
    data = Baseball_Data)

Coefficients:
              (Intercept)  Baseball_Data$Payroll..x.
                  71.4720                     0.1056

> summary(baseballlm)

Call:
lm(formula = Baseball_Data$Wins..y. ~ Baseball_Data$Payroll..x.,
    data = Baseball_Data)

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_Data$Payroll..x.   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(baseballlm)
                               2.5 %     97.5 %
(Intercept)              61.32240470 81.6216904
Baseball_Data$Payroll..x.  0.00173383  0.2094509
```

+1

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



+1

Parameter	Estimate	Standard Error	t Value	Pr > |t|	99% Confidence Limits	
Intercept	21.70019172	2.75429099	7.88	<.0001	14.53659134	28.86379211
sci	0.59681405	0.05218220	11.44	<.0001	0.46109403	0.73253407

```
proc glm data = testdata;
  model math=sci / clparm alpha = 0.01;
  run;
```

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.