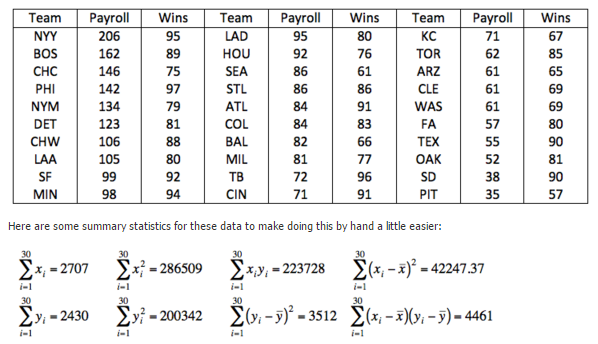
Unit 9 HW Solutions

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



## Question 1 (55 points total)

### Part A (15 points)

1. 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.
2. Interpret the slope AND the intercept in the context of the problem.

*Note: if only the results are given without showing work, 10 points should be subtracted. The slope is worth 5 points, the intercept is worth 5 points, and each interpretation is worth 2.5 points*

**Slope: for every $1 million increase in payroll, the estimated mean number of wins increases by 0.1056 wins.**

**Intercept: for a team with zero payroll, the expected mean number of wins is about 71. This is, of course, extrapolation, as $0 payrolls are not realistic in MLB.**

### Part B (15 points)

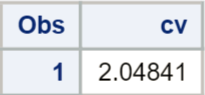
Is the slope (only concerned with the slope here) of the regression line significantly different from zero? Carry out a 6-step hypothesis test to address this question. Use the above calculations to find the relevant statistics for this test. You will need to use SAS, R, the internet, a calculator, or 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 .

**(2 points) Step 1 - Hypotheses:**

**(2 points) Step 2 - Identification of Critical Value:**

data mycritval;  
cv = quantile(“t”, 0.975, 30-2);  
run;

proc print data = mycritval;  
run;



qt(0.975, 30-2)

## [1] 2.048407

**(4 points) Step 3 - Value of Test Statistic: First, calculate the standard error of :**

**So, the test statistic is:**

**(2 points) Step 4 - Give p-value:**

data mypval;  
pv = 2\*(1-cdf(“t”, 2.08, 30-2));  
run;

proc print data = mypval;  
run;



2\*pt(2.08, 30-2, lower.tail=F)

## [1] 0.04679611

**(2 points) Step 5 - Decision: Reject**

**(2 points for the conclusion, 1 point for the scope) Step 6 - Conclusion: There is sufficient evidence at the level of significance () that there is a linear relationship between the payroll of a baseball team in this season and the mean number of games won. Note that this is a census of the entire season, and it cannot be generalized to other seasons. Because the payrolls were not randomly assigned to teams, no causality can be inferred.**

### Part C (15 points)

1. 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.
2. Interpret the interval.

**A 95% confidence interval for is:**

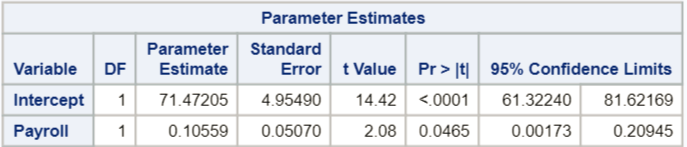
**Point Estimate Multiplier Standard Error**

**Interpretation: We are 95% confident that the slope parameter () is contained in the interval . This interval is the range of plausible values for the true slope, .**

### Part D (10 points)

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 reg data = baseball;  
model wins = payroll / clb;  
run;



**Parameter estimates, t-value, p-value, and confidence interval are all the same, give or take rounding error in the hand calculations.**

## Question 2 (45 points total)

### Part A (9 points)

1. 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.
2. Interpret the slope and the intercept in the context of the math and science scores.

*Note: 3 points for the equation and relevant statistics/output, 3 points for the slope interpretation, and 3 points for the intercept interpretation.*

test.data <- read.csv('C:/Users/Charles/Documents/SMU/Online Teaching/MSDS 6371 - Statistical Foundations for Data Science/UNIT 9/TEST DATA.csv')  
  
math.lm <- lm(math ~ science, data=test.data)  
summary(math.lm)

##   
## Call:  
## lm(formula = math ~ 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 \*\*\*  
## 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

**Slope: it is estimated that for each 1-point increase in the science score, the estimated (predicted) mean math score increases by 0.5968 points.**

**It is estimated that for a science score of zero, the predicted math score is 21.7002. Of course, this is extrapolation, as no science scores of 0 are in the data set.**

### Part B (16 points)

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.

*Note: the numbers below are taken from the relevant R output given in part A. Each hypothesis test is worth 8 points and has the same point structure.*

**First, the slope (though order doesn’t matter).**

**(2 points) Step 1 - Hypotheses:**

**(1 point) Step 2 - Identification of Critical Value:**

qt(0.995, 200-2)

## [1] 2.600887

**(1 point) Step 3 - Value of Test Statistic: :**

**(1 point) Step 4 - Give p-value:**

2\*pt(11.437, 200-2, lower.tail=F)

## [1] 1.389772e-23

**(1 point) Step 5 - Decision: Reject**

**(1 point for the conclusion, 1 point for the scope) Step 6 - Conclusion: There is overwhelming evidence at the level of significance () that there is a linear relationship between the science score of student and the mean math score. In other words, there is overwhelming evidence that the slope of the regression equation is not equal to zero. We do not have sufficient information on the data collection methodology to generalize the results to a broader population. Because this is an observational study, the results are limited to revealing the association between math and science scores, rather than causality.**

**Next, the intercept.**

**(2 points) Step 1 - Hypotheses:**

**(1 point) Step 2 - Identification of Critical Value:**

qt(0.995, 200-2)

## [1] 2.600887

**(1 point) Step 3 - Value of Test Statistic: :**

**(1 point) Step 4 - Give p-value:**

2\*pt(7.879, 200-2, lower.tail=F)

## [1] 2.147623e-13

**(1 point) Step 5 - Decision: Reject**

**(1 point for the conclusion, 1 point for the scope) Step 6 - Conclusion: There is overwhelming evidence at the level of significance () that the predicted math score is not zero when the science score is 0. In other words, there is overwhelming evidence that the regression equation does not pass through the origin. Obviously, this is extrapolation, as no science scores were zero in the data set. We do not have sufficient information on the data collection methodology to generalize the results to a broader population. Because this is an observational study, causality cannot be inferred.**

### Part C (10 points)

1. 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).
2. Interpret these intervals.

*Note: 3 points for the slope interval, 3 points for the intercept interval, 2 points for the slope interpretation, 2 points for the intercept interpretation.*

**A 99% confidence interval for is:**

**Point Estimate Multiplier Standard Error**

**Interpretation: We are 99% confident that the increase in the mean math score for a 1 unit increase in the science score is contained in the interval (0.4611, 0.7325) points.**

**A 99% confidence interval for is:**

**Interpretation: We are 99% confident that for a science score of 0, the mean math score is contained in the interval (14.539, 28.861) points. This is an extrapolated interval.**

### Part D (10 points)

Verify your confidence intervals with R and paste your code and relevant output below.

test.data <- read.csv('C:/Users/Charles/Documents/SMU/Online Teaching/MSDS 6371 - Statistical Foundations for Data Science/UNIT 9/TEST DATA.csv')  
  
math.lm <- lm(math ~ science, data=test.data)  
confint(math.lm, level=0.99)

## 0.5 % 99.5 %  
## (Intercept) 14.536591 28.8637921  
## science 0.461094 0.7325341

## BONUS: (+6 points, 1 for each part)

1. Repeat 1(d) using R.

baseball <- read.csv('C:/Users/Charles/Documents/SMU/Online Teaching/MSDS 6371 - Statistical Foundations for Data Science/UNIT 8/Baseball\_Data.csv')  
  
payroll.lm <- lm(Wins ~ Payroll, data=baseball)  
summary(payroll.lm)

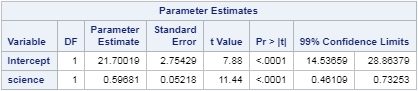
##   
## Call:  
## lm(formula = Wins ~ Payroll, data = baseball)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19.553 -8.340 1.099 9.301 16.925   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 71.4720 4.9549 14.425 1.73e-14 \*\*\*  
## Payroll 0.1056 0.0507 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(payroll.lm)

## 2.5 % 97.5 %  
## (Intercept) 61.32240470 81.6216904  
## Payroll 0.00173383 0.2094509

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

proc reg data = test\_data;  
model math = science / clb;  
run;



**Same as with R:**

1. 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.

**A 95% confidence interval for the mean number of wins when a team has a payroll equal to $100m is (78 wins, 86 wins).**

1. Give a 95% PI (prediction interval) for the number of wins for a team with $100 million payroll. Use SAS or R.

**A 95% prediction interval for the wins for an individual team when the team has a payroll equal to $100m is (60 wins, 104 wins).**

1. Explain the difference between these two intervals.

**The confidence interval is the set of plausible values for the mean number of wins for all possible teams if they had a payroll of $100m. The prediction interval is the set of the plausible number of wins for any individual team that has a payroll of $100m. Naturally, the prediction interval, where we are finding the range of one value, is much larger than the confidence interval, which finds the range of the mean of the population.**

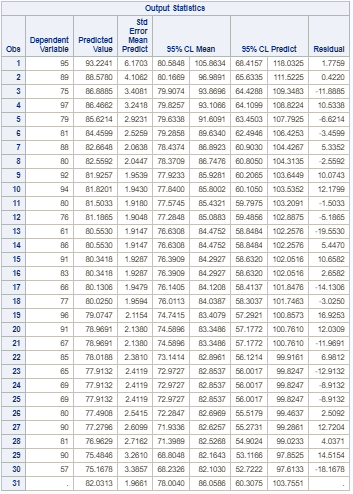
\*To create a record with payroll=100 but missing wins;  
data baseball2;  
input Team $ Wins Payroll;  
datalines;  
dummy . 100  
;  
run;

proc print data = baseball2;  
run;

\*To add the dummy record to the original baseball set;  
data combined;  
set baseball baseball2;  
run;

proc print data = combined;  
run;

\*To get confidence intervals and prediction intervals at every value of payroll (especially 100, which is what we are looking for);  
proc reg data = combined alpha = 0.05;  
model Wins=payroll / clm cli;  
run;



baseball <- read.csv('C:/Users/Charles/Documents/SMU/Online Teaching/MSDS 6371 - Statistical Foundations for Data Science/UNIT 8/Baseball\_Data.csv')  
  
payroll.lm <- lm(Wins ~ Payroll, data=baseball)  
newpoint <- data.frame(team=NA, Payroll=100, Wins=NA)  
predict(payroll.lm, newdata = newpoint, interval='confidence')

## fit lwr upr  
## 1 82.03129 78.00399 86.05858

predict(payroll.lm, newdata = newpoint, interval='prediction')

## fit lwr upr  
## 1 82.03129 60.30746 103.7551