Name: James Tsai
Section: MSDS6371-401
Date: 10/31/15

1. **Find the least squares regression line for using payroll to predict number of wins. Interpret the slope and the intercept in the context of the problem.**

Assumptions:

1) The subpopulation of responses for each value of the explanatory variable is normally distributed.
2) The subpopulation has equal standard deviations.
3) The means of the subpopulation fall on a straight line function of the explanatory variable.
4) The observations from any given subpopulation is independent of the other observations.

First we solve for the B_1 and B_0 given the following formula:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}, \qquad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}$$

We are given the summary statistic:

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

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

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

We can solve for x-bar and y-bar:

x-bar = 2707/30 = 90.23333
y-bar = 2430/30 = 81

We can also use the last two summary statistics to solve for B_1 and B_0:

B_1 = 4461/42247.37 = 0.1055924
B_0 = 81 – (.1055924)(90.23333) = 71.47205

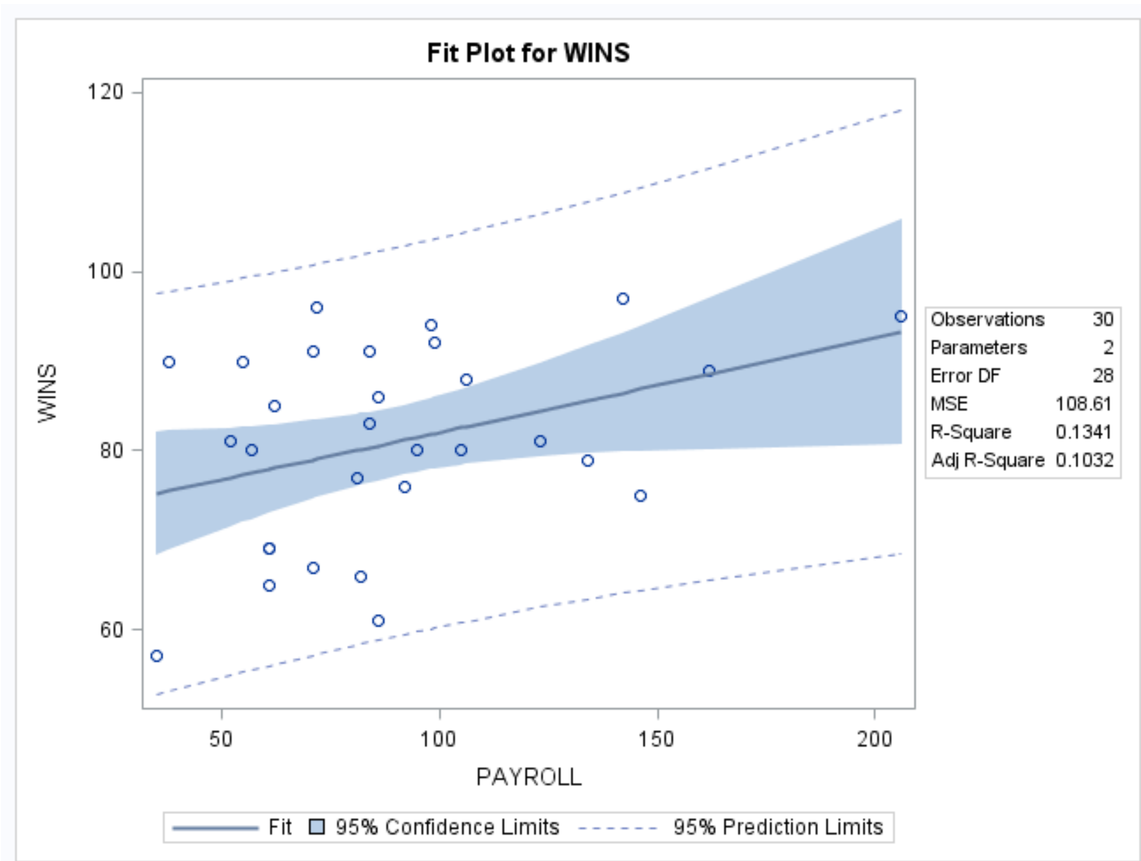
The least squares regression line is:

y = 0.1055924x + 71.47205

Interpretation of slope and intercept:

For each increase of 10 units in payroll, we can expect an increase of 1 win. When the payroll is zero, the estimated win is 71. In this case, payroll of zero has no practical meaning.

SAS Output:



2. Is the slope of the regression line significantly different from zero? Carry out the appropriate test and interpret the results.

<u>We use the 6-step hypothesis test (t-test) for the slope:</u>

1) <u>Set up H_0 and H_A</u>

H_0: $B_1 = 0$

H_A: $B_1 \neq 0$

2) Identify alpha and critical value

$\alpha = 0.05$

$t_{28}(.975) = 2.048$

3) Identify the test statistic

t-statistic = $(B_1 - 0)/SE$

$$SE(\hat{\beta}_1) = \hat{\sigma}\sqrt{\frac{1}{(n-1)s_X^2}}, \quad \text{d.f.} = n - 2$$

$$\hat{\sigma} = \sqrt{\frac{\textit{Sum of all squared residuals}}{\textit{Degrees of freedom}}},$$

PAYROLL	WINS	SQUARED RESIDUALS
206	95	3.153902832
162	89	0.178073409
146	75	141.3372549
142	97	110.9616774
134	79	43.84328411
123	81	11.97097755
106	88	28.46393512
105	80	6.549747033
99	92	101.491659
98	94	148.3499456
95	80	2.259981963
92	76	26.90026484
86	61	382.3195076
86	86	29.66989296
84	91	113.5970663
84	83	7.065987105
82	66	199.6745009
81	77	9.150809126
72	96	286.465811
71	91	144.7423932
71	67	143.2595156
62	85	48.73749734
61	65	166.7502951
61	69	79.44483116
61	69	79.44483116
57	80	6.296016837

55	90	161.8078443
52	81	16.29856678
38	90	210.6980435
35	57	330.0682785
SSR		**3040.952**
S_x		**38.168**

Root MSE = σ = (3040.952/28)^0.5 = 10.42139
SE(B_1) = (10.42139)((1/(29)(38.168^2))^0.5) = 0.0507
t-statistic = (B_1-0)/SE(B_1) = (0.1055924–0)/0.0507 = 2.08269

4) <u>Find p-value</u>

p-value = 0.0466 (two-tail test)

5) <u>Reject H_0 if the p-value is less than the significance level (alpha). Fail to reject if it is not</u>

Reject H_0 since 0.0466 < 0.05

6) <u>Conclusion</u>

There is sufficient evident to suggest at alpha = 0.05 level of significance (p-value = 0.0466) to suggest that slope B_1 is significantly different than zero.

SAS Output:

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|
Intercept	1	71.47205	4.95490	14.42	<.0001
PAYROLL	1	0.10559	0.05070	2.08	0.0465

3. Calculate a confidence interval for the slope and interpret this interval.

CI for slope = B_1 ± (SE(B_1)*(critical value)) = 0.1055924 ± (0.0507)(2.048) =
CI for slope = (0.0017588,0.209426)

A 95% confidence interval is (0.0017588, 0.209426). There is a 95% confidence that for every additional increase of 10 units in payroll, the wins increase between 0.017588 and 2.09426. Since our confidence interval does not include zero, it is consistent with our conclusion of rejecting H_0.

4. Give a 95% CI for the expected number of wins for a team with $100 million payroll. Give a 95% PI for the number of wins for a team with $100 million payroll. Explain the difference between these two intervals.

We use the following formula to calculate the SE:

$$\mathrm{SE}[\hat{\mu}\{Y|X_0\}] = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{(n-1)s_X{}^2}}, \quad \text{d.f.} = n = 2.$$

We plug in the values we calculated earlier:
σ = 10.42139
n = 30
s_x = 38.168
x-bar = 90.23333

SE($ß_1$) = 10.42139*((1/30) + ((100-90.23333)^2/(29)*(38.168)^2))^0.5 =
SE($ß_1$) = 10.42139*(0.033333 + .0022578559)^0.5 = 1.966

For the mean CI, we use the following formula:
CI = $ß_1$ ± (SE($ß_1$)*t)

To solve for $ß_1$ plug 100 into y = 0.1055924x + 71.47205, we get y = 0.1055924(100)+71.47205 = 82.03129.

CI = 82.03129 ± (4.0263)

$\boxed{\text{CI = (78.0049, 86.0577)}}$

For a specific payroll amount, we use the following formula to calculate standard error:

$$\mathrm{SE}[\mathrm{Pred}\{Y|X_0\}] = \sqrt{\hat{\sigma}^2 + \mathrm{SE}[\hat{\mu}\{Y|X_0\}]^2}.$$

$SE(ß_1) = (10.42139^2+1.966^2)$^0.5 = 10.6052

<u>For the PI, we use the following formula:</u>
PI = $ß_1$ ± ($SE(ß_1)$*t) = 82.03 ± (10.6052*2.048) = 82.03 ± 21.7194

$\boxed{PI = (60.3106, 103.7494)}$

The 95% confidence interval for the number of wins (78.0049, 86.0577) for a team with a 100 million dollar payroll pertains to the average of the payroll data; in this case it is 90.23333 million dollars.

The 95% predicted interval for the number of wins (60.3106, 103.7494) for a team pertains to a specific amount of payroll; in this case it is 100 million dollars.