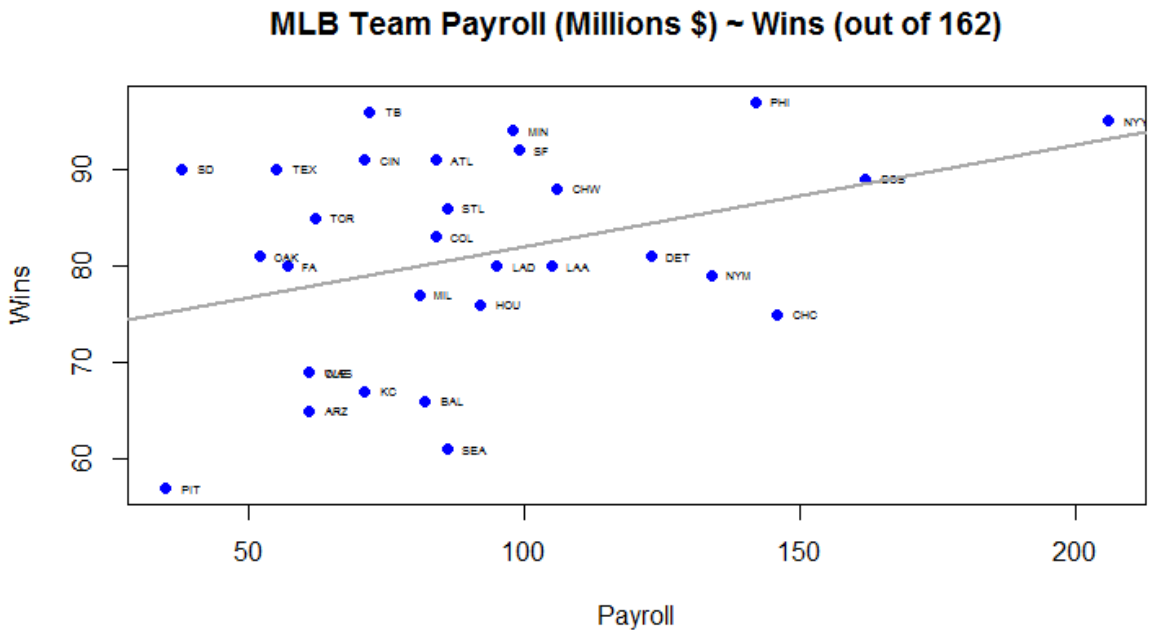
Question 1)

Scatterplots and figures (R and SAS) for the full data set are included below (Figure 1), which indicate a **moderate, positive linear correlation coefficient of roughly 0.4**, with higher values of Payroll (X) correlated with higher numbers of Wins (Y).

Question 2)

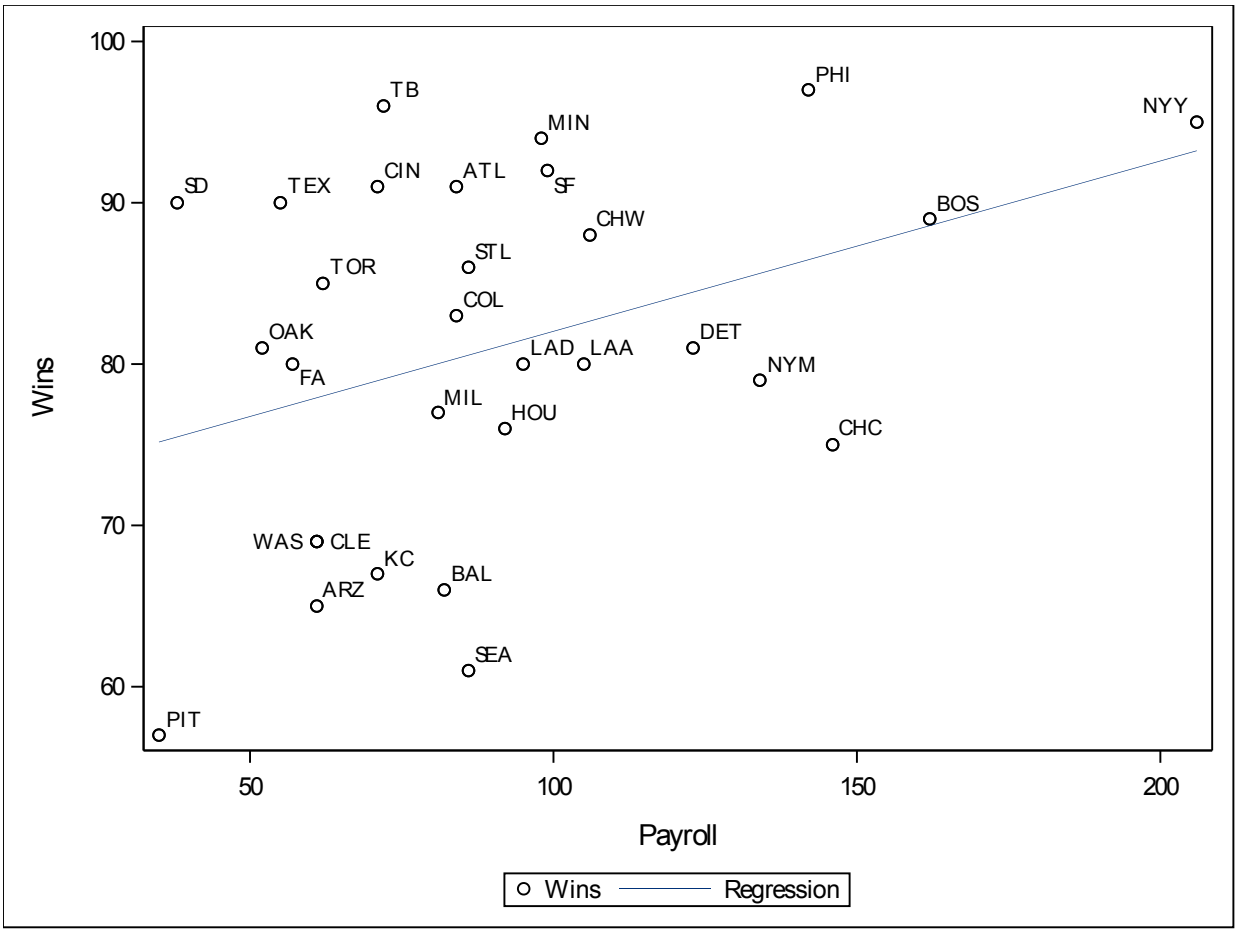
Statistical software (R and SAS) was used to verify this estimate of the **correlation coefficient**, which had an actual value of **0.36623**.

Figure 1:



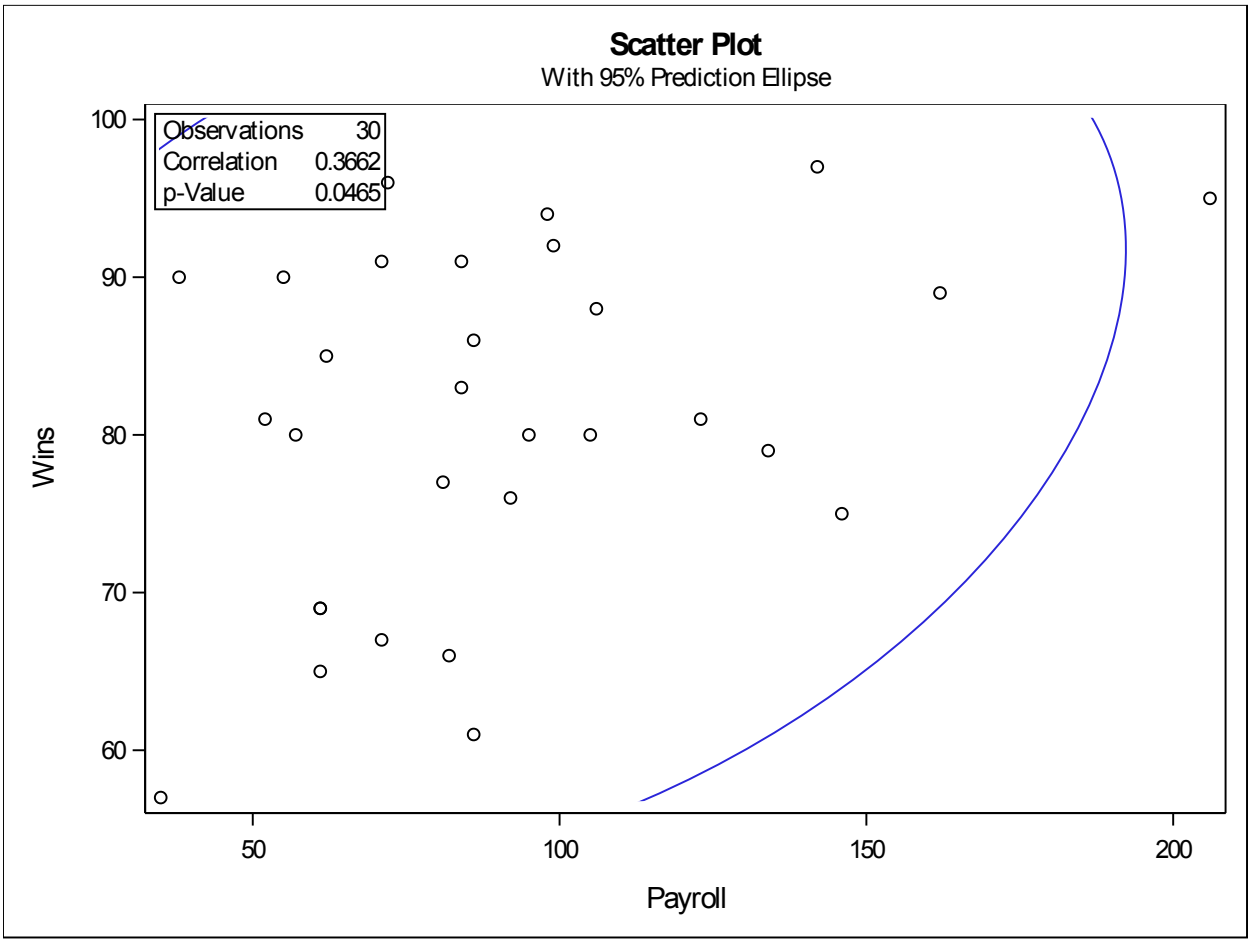

```
> cor(baseball4$Payroll, baseball4$Wins)
[1] 0.366231

> cor.test(baseball4$Payroll, baseball4$Wins)

        Pearson's product-moment correlation

data:  baseball4$Payroll and baseball4$Wins
t = 2.0826, df = 28, p-value = 0.04654
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.00686799 0.64181770
sample estimates:
      cor
0.366231
```

Simple Statistics							
Variable	**N**	**Mean**	**Std Dev**	**Sum**	**Minimum**	**Maximum**	**Label**
Payroll	30	90.23333	38.16812	2707	35.00000	206.00000	Payroll
Wins	30	81.00000	11.00470	2430	57.00000	97.00000	Wins



Pearson Correlation Coefficients, N = 30 Prob > |r| under H0: Rho=0		
	Payroll	**Wins**
Payroll Payroll	1.00000	0.36623 0.0465
Wins Wins	0.36623 0.0465	1.00000

Pearson Correlation Statistics (Fisher's z Transformation)									
Variable	With Variable	N	Sample Correlation	Fisher's z	Bias Adjustment	Correlation Estimate	95% Confidence Limits		p Value for H0:Rho=0
Payroll	Wins	30	0.36623	0.38406	0.00631	0.36075	0.000554	0.638089	0.0460



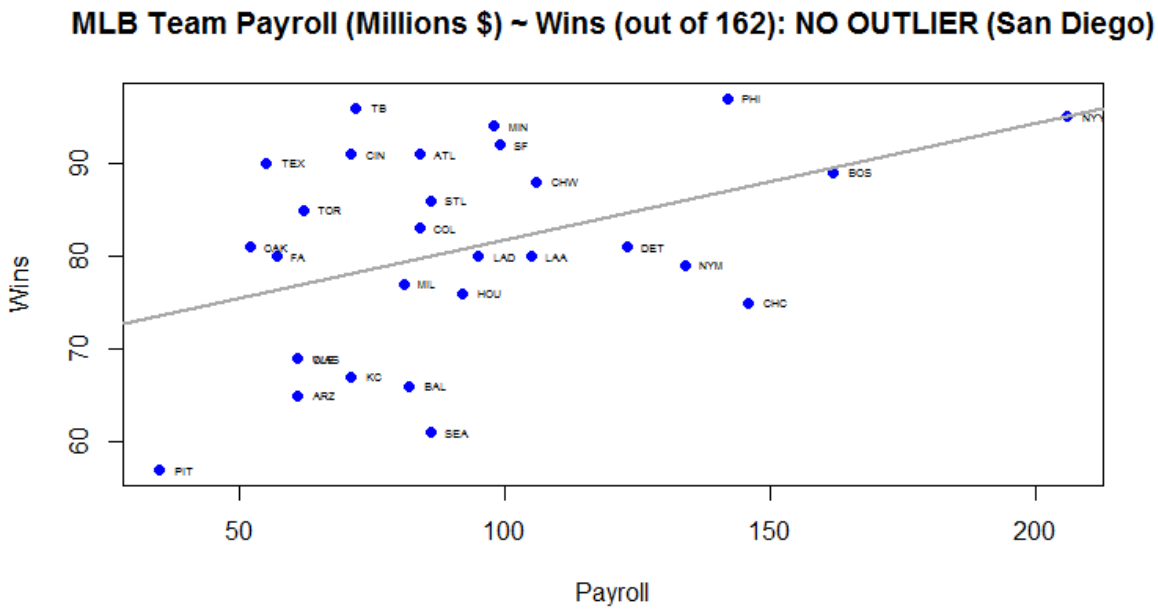

```
proc sgplot data = baseball;

scatter x = Payroll y = Wins / datalabel = ID;

reg y = Wins x = Payroll;

run;
```

```
proc corr data = baseball fisher plots = scatter;

var Payroll Wins;

run;
```

Question 3)

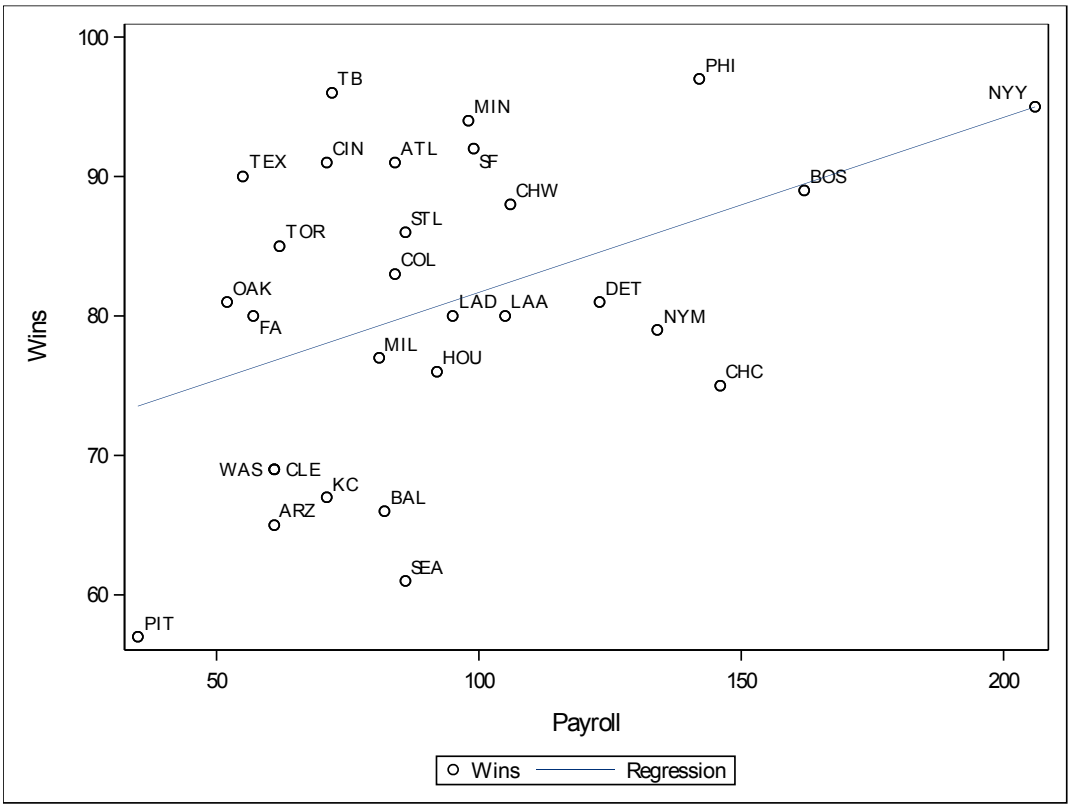
Another analysis of correlation (in R and SAS) was run, **removing the outlier** associated with the team from **San Diego (SD)**, which has an unusually high number of wins despite having a very low payroll. As shown below (Figure 2), this results in a **correlation coefficient of 0.42555**.

Figure 2:



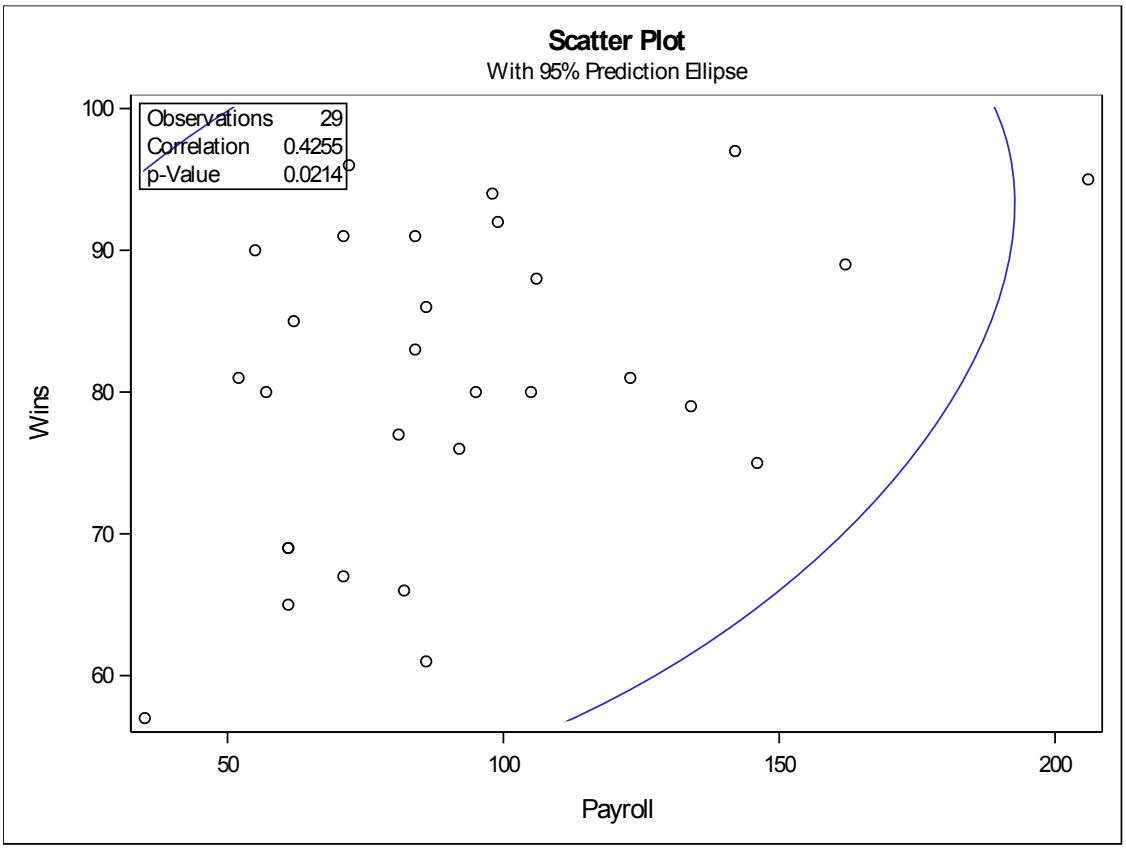

```
> cor(baseball_noSD$Payroll, baseball_noSD$Wins)
[1] 0.4255494

> cor.test(baseball_noSD$Payroll, baseball_noSD$Wins)

        Pearson's product-moment correlation

data:  baseball_noSD$Payroll and baseball_noSD$Wins
t = 2.4435, df = 27, p-value = 0.02136
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.06995422 0.68518874
sample estimates:
      cor
0.4255494
```

Simple Statistics							
Variable	**N**	**Mean**	**Std Dev**	**Sum**	**Minimum**	**Maximum**	**Label**
Payroll	29	92.03448	37.52379	2669	35.00000	206.00000	Payroll
Wins	29	80.68966	11.06508	2340	57.00000	97.00000	Wins



Pearson Correlation Coefficients, N = 29 Prob > |r| under H0: Rho=0		
	Payroll	**Wins**
Payroll Payroll	1.00000	0.42555 0.0214
Wins Wins	0.42555 0.0214	1.00000

Pearson Correlation Statistics (Fisher's z Transformation)									
Variable	With Variable	N	Sample Correlation	Fisher's z	Bias Adjustment	Correlation Estimate	95% Confidence Limits		p Value for H0:Rho=0
Payroll	Wins	29	0.42555	0.45445	0.00760	0.41931	0.062388	0.681136	0.0205




```
proc sgplot data = baseball_noSD;

scatter x = Payroll y = Wins / datalabel = ID;

reg y = Wins x = Payroll;

run;


proc corr data = baseball_noSD fisher plots = scatter;

var Payroll Wins;

run;
```

Question 4)

Pointing to two outliers and claiming that there is no winning advantage for teams with a higher payroll, as the league commissioner does, would be tantamount to assessing the strength of the correlation on the basis of single (and not particularly representative) observations. **The statement would only hold true if the value of the estimated correlation coefficient were approximately zero**. Since there is, however, a fairly moderate, positive linear correlation, there is a correspondingly moderate advantage for teams having a higher payroll. By comparison, if one were constructing a linear model using only Payroll (X) as the explanatory variable for the number of Wins (Y), one would find a coefficient of determination of **approximately $R^2 = 0.1341$** (Figure 3), suggesting that roughly 13% of the variance in the response variable (Y) can be attributed to changes in the explanatory variable (X). In a very rough sense, one might say that the statement made by the league commissioner is ~13% wrong!

Figure 3:

```
> baseball4_model <- lm(Wins~Payroll)
> summary(baseball4_model)

Call:
lm(formula = Wins ~ 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 ***
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
```

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	71.47205	4.95490	14.42	<.0001
Payroll	Payroll	1	0.10559	0.05070	2.08	0.0465

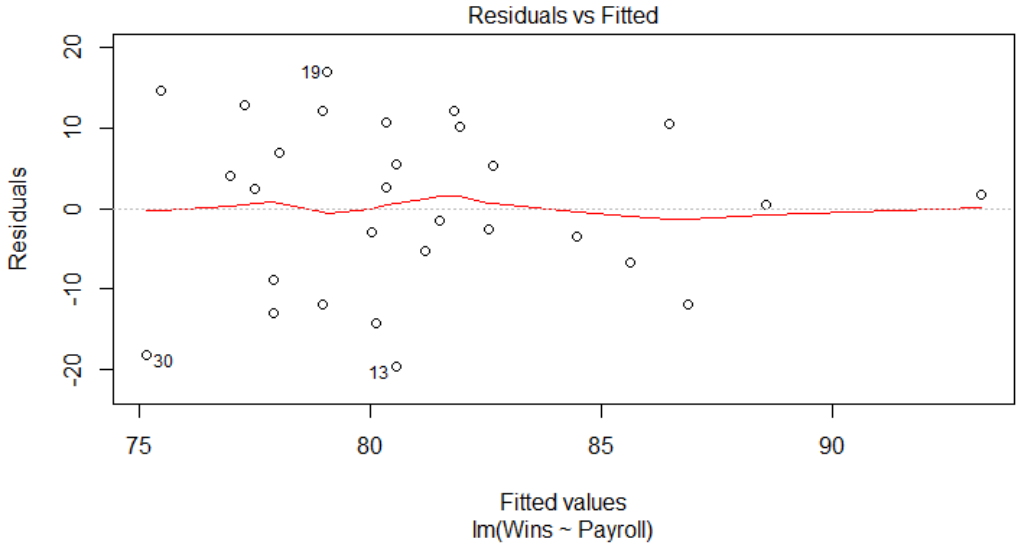
Root MSE	10.42139	**R-Square**	0.1341
Dependent Mean	81.00000	**Adj R-Sq**	0.1032
Coeff Var	12.86592		

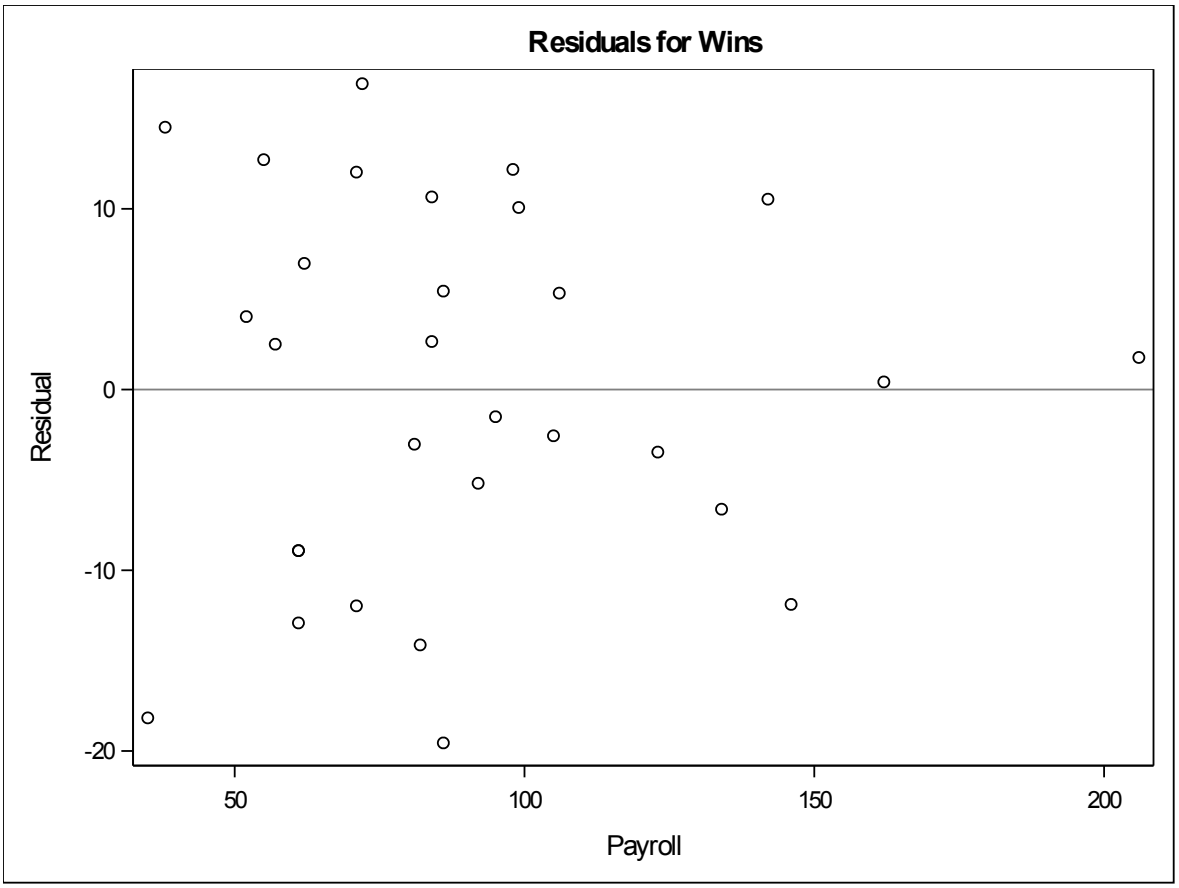
```
proc reg data = baseball;

model Wins = Payroll;

title "Wins|Payroll";

run; quit;
```

Question 5)

Given the meagre description of the study, the **assumptions** should be as follows: 1) the variables are **non-ordinal and continuous** in measurement; 2) the variables appear to be **related pairs**; 3) **outliers** exist, but should not grossly distort the correlation estimation; 4) the relation between the variables appears **linear** enough to warrant an evaluation of their correlation; 5) likewise, the variables appear sufficiently **homoscedastic**; and, in accordance with the Gauss-Markov theorem, 6) their **errors** should be uncorrelated with mean zero and homoscedastic with finite variance (Figure 4).

Figure 4:





However, from the description of this **observational study**, there is **no indication that this data is a random sampling** of any kind. Regarding the **population** being sampled, although the data is from the 2010 regular season, the study description is phrased such that the question of interest seems to be the correlation between Payroll and Wins not just for 2010 or for these 30 teams, but rather for **all American MLB teams throughout the history of the league**. While the analysis did identify a statistically significant correlation for the full data set (**p-value = .0465**), since neither sampling procedure nor allocation was randomized, one **cannot make any causal determinations**, **nor can one confidently make inferences to the population**, though the **results do suggest that further analysis with a random sampling procedure is warranted** for making a more thorough-going assessment of the correlation in question.