

Thirty five people took the exam. The mean score was 69.4 and the median score was 73. The standard deviation of the scores was 20.5.

1. There are 30 major league baseball teams whose payrolls in 2003 ranged from 19.63 to 152.75 million dollars. The question of whether money can buy success has been much debated. Each year 8 of the 30 teams make the playoffs based on their records in the 162 games of the regular season. The playoffs end with the world series. In 2003 the Florida Marlins with a payroll of 49.05 defeated the New York Yankees with a payroll of 152.75 to win the championship.

The 8 playoff teams ranked 1, 3, 6, 9, 11, 18, 23, and 25 in size of their payrolls. They had an average payroll of 84.6 million with a standard deviation of 35.2 while the remaining 22 teams had a average payroll of 66.1 million with a standard deviation of 23.9.

Below is some of the results from the linear regression

$$Wins = \beta_0 + \beta_1 Payroll + Z$$

> Payroll

```
[1] 152.75 117.18 106.24 105.87 103.49 99.95 86.96 83.49 82.85 80.64
[11] 79.87 79.03 73.88 71.04 70.78 67.18 59.36 55.51 54.81 51.95
[21] 51.27 51.01 50.26 49.17 49.05 48.59 47.93 40.63 40.52 19.63
```

> Wins

```
[1] 101 66 101 85 71 95 93 85 100 84 88 77 71 87 86 74 69 90
[19] 75 83 86 86 96 43 91 68 64 68 83 63
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	66.90062	6.26135	10.685	2.19e-11 ***
Payroll	0.19803	0.08221	2.409	0.0228 *

---

Residual standard error: 12.38 on 28 degrees of freedom

Multiple R-Squared: 0.1717,

Analysis of Variance Table

Response: Wins

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Payroll	1	889.4	889.4	5.8029	0.02283 *
Residuals	28	4291.6	153.3		

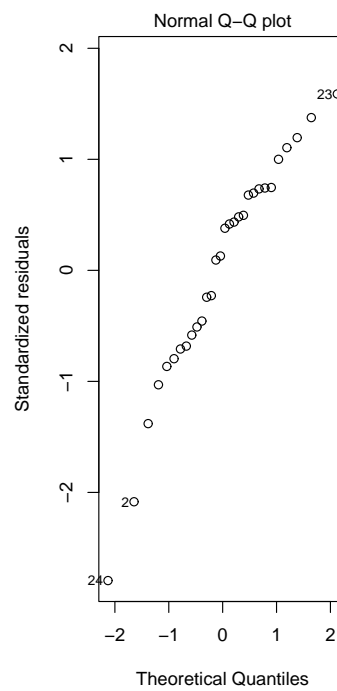
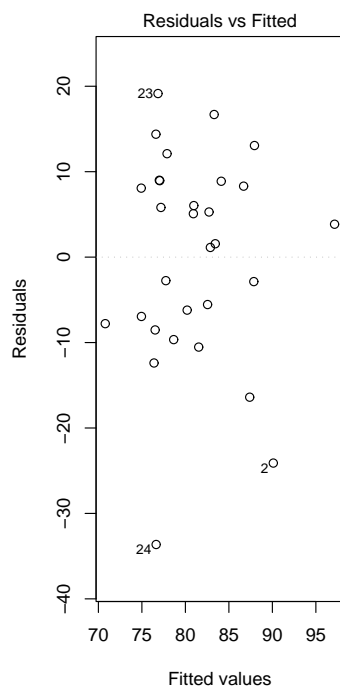
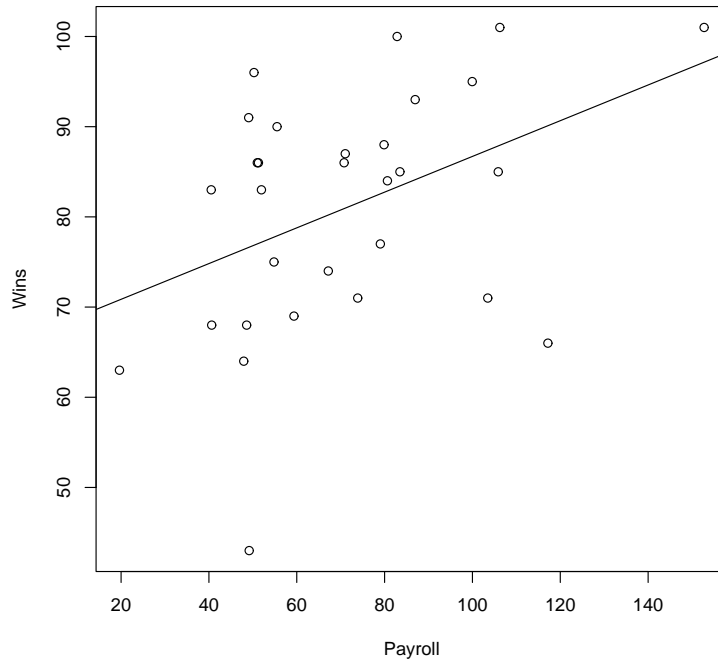
```
predict(bball.lm,data.frame(Payroll=c(53,62,69)),se.fit=T)
```

```
$fit
```

	1	2	3
	77.39624	79.17852	80.56473

\$se.fit

	1	2	3
	2.702926	2.379076	2.266458



Using the above information answer the following questions.

- i) At level  $\alpha = .05$  test the hypotheses that there is no difference in the mean payroll of playoff teams and non-playoff teams against the alternative that mean payroll for playoff teams is greater.
- ii) What is the  $p$ -value or level of significance for testing  $H : \beta_1 = 0$  against  $K : \beta_1 > 0$ .
- iii) Find a 95% confidence interval for expected number of wins for a team with a payroll 62 million.
- iv) Find the correlation between Wins and Payroll.
- v) What is the least squares estimate of  $\beta_0$ ? How do you interpret this value?
- vi) Does money buy success? Briefly justify your answer.

**Solution:**

i) Note

$$S_p^2 = (7(35.2)^2 + 21(23.9)^2)/(7 + 21) = 738.2$$

and so

$$\frac{84.6 - 66.1}{S_p(1/8 + 1/22)^{1/2}} = 1.65$$

and we accept the null hypothesis since  $t_{.05,28} = 1.70$ .

- ii)  $P(T_{28} > 2.409) = 0.0114$ .
- iii) Since  $t_{.025,28} = 2.048$  the interval is

$$79.2 \pm 2.048 \times 2.38$$

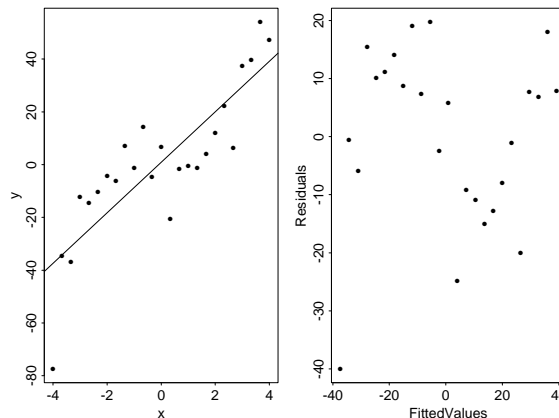
iv) Let  $\rho$  be the correlation. From the plot we see that  $\rho > 0$  and since  $\rho^2 = R\text{-squared} = 0.1717$  we have  $\rho = 0.414$ .

v)  $\hat{\beta}_0 = 66.9$ . Since this is the intercept of the linear regression equation, formally it is the estimate of the number of games a team would win that had a payroll of \$0. Note this does not make much sense. More informally one might think of it as an estimate for the minimum number of games that any team should win. Note however three teams did in fact win less than that.

vi) Although the results of parts i) and ii) are somewhat contradictory the plot and the fact that  $\rho = 0.414$  suggests that the size of the payroll does play a role in determining success. A more careful analysis would need to consider the results from more than one season.



2. In an experiment to study the relationship between an independent variable  $X$  and a dependent variable  $Y$  25 observations were taken where the  $X$  values were equally spaced from -4.0 to 4.0. First a simple linear regression model  $Y = \beta_0 + \beta_1 X + Z$  was fit to the data. The first graph contains a plot of  $y$  against  $X$  and the least squares line. The second graph is a plot for this model of the residuals against the fitted or predicted values.



Next the model  $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + Z$  was fit to the data. Below is the some of the output for this model along with two anova tables for the model. Note xsq is  $X^2$  and xcb is  $X^3$ .

Coefficients:

	Value	Std. Error	t value	Pr(> t )
(Intercept)	1.4345	2.9478	0.4866	0.6315
x	-1.5443	2.0528	-0.7523	0.4602
xsq	-0.0945	0.3807	-0.2484	0.8063
xcb	1.0713	0.1815	5.9029	0.0000

Residual standard error: 9.813 on 21 degrees of freedom

Multiple R-Squared: 0.8914

F-statistic: 57.46 on 3 and 21 degrees of freedom, the p-value is 2.7e-10

Terms added sequentially (first to last)

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
x	1	13237.89	13237.89	137.4751	0.0000000
xsq	1	5.94	5.94	0.0617	0.8062701
xcb	1	3355.20	3355.20	34.8437	0.0000074
Residuals	21	2022.15	96.29		

Terms added sequentially (first to last)

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
xcb	1	16538.60	16538.60	171.7528	0.0000000
xsq	1	5.94	5.94	0.0617	0.8062701
x	1	54.50	54.50	0.5659	0.4602276
Residuals	21	2022.15	96.29		

Use this output to answer the following questions. If a question cannot be answered with the output in hand explain what additional output is needed.

i) At level  $\alpha = .05$  test  $H: \beta_2 = \beta_3 = 0$  against  $K: \text{at least one is not zero}$ .

ii) At level  $\alpha = .05$  test  $H: \beta_1 = \beta_3 = 0$  against  $K: \text{at least one is not zero}$ .

iii) Find the value of  $R^2$  for these data if we fit the model  $Y = \beta_0 + \beta_3 X^3 + Z$ .

iv) Based on this information suggest a model that can be used for predicting  $Y$  by using  $X$ .

Briefly justify your answer.

**Solution:** i)

$$\frac{(SSReg\beta_2|\beta_0, \beta_1 + SSReg(\beta_3|\beta_0, \beta_1, \beta_2))/2}{RSS/21} = \frac{(5.94 + 3355.20)/2}{2022.15/21} = 17.45$$

Since  $17.45 > 3.47 = f_{.05;2,21}$  we reject  $H$ .

ii) Cannot answer this question. We need an anova table where  $X$  and  $X^3$  are the last two variables in the table.

iii) Since the total TCSS is the same for all the models and for the full model we have

$$\begin{aligned} TCSS &= SSReg(\beta_3|\beta_0) + SSReg(\beta_2|\beta_0, \beta_3) + SSReg(\beta_1|\beta_0, \beta_3, \beta_2) + RSS \\ &= 16538.60 + 5.94 + 54.50 + 2022.15 \\ &= 18621.19 \end{aligned}$$

and so

$$R^2 = \frac{SSReg(\beta_3|\beta_0)}{TCSS} = \frac{16538.60}{18621.19} = 0.888$$

iv) For the model  $Y = \beta_0 + \beta_1X + Z$  we have  $R^2 = 13237.89/18621.19 = 0.711$  but the two graphs show that there is a pattern to the residuals for this model and this suggest that a polynomial model should be used. Now  $R^2 = 0.8914$  for the full model while for the model  $Y = \beta_0 + \beta_3X^3 + Z$  we have  $R^2 = 0.888$  and so this later model is probably preferred.

■