Homework Assignment 6: Solutions

STA 371G, Statistics and Modeling

Problem 1: Housing Price Structure

The file **MidCity.csv**, available in the course website, contains data on 128 recent sales of houses in a town. For each sale, the file shows the neighborhood in which the house is located, the number of offers made on the house, the square footage, whether the house is made out of brick, the number of bathrooms, the number of bedrooms, and the selling price. Neighborhoods 1 and 2 are more traditional whereas 3 is a more modern, newer and more prestigious part of town. Let N_2 be 1 if the house is in neighborhood 2 and be 0 otherwise, and let N_3 be 1 if the house is in neighborhood 3 and be 0 otherwise.

To estimate the house pricing structure in this town, we consider a regression model as

$$Y = \beta_0 + \beta_1\, Brick + \beta_2\, N_2 + \beta_3\, N_3 + \beta_4\, Bids \\ + \beta_5\, SqFt + \beta_6\, Bed + \beta_7 Bath + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma^2).$$

Consider, in particular, the following questions and be specific in your answers:

(a) Is there a premium for brick houses?

(b) Is there a premium for houses in neighborhood 3?

(c) For the purposes of prediction could you combine the neighborhoods 1 and 2 into a single "older" neighborhood?

(d) Is there an extra premium for brick houses in neighborhood 3?
(Hint: include $N_3 * Brick$ into your regression)

(e) Based on this model, explain the relationship between the selling price of a house and the number of offers made on the house. Does it make sense to use this model to predict the selling price of a house before putting it on the market? If no, provide a suggestion to improve the model.

There may be more than one way to answer these questions.

(a) To begin we create dummy variable $Brick$ to indicate if a house is made of brick and N_2 and N_3 to indicate if a house came from neighborhood two and neighborhood three respectively. Using these dummy variables and the other covariates, we ran a regression for the model

$$Y = \beta_0 + \beta_1\, Brick + \beta_2\, N_2 + \beta_3\, N_3 + \beta_4\, Bids \\ + \beta_5\, SqFt + \beta_6\, Bed + \beta_7 Bath + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma^2).$$

and got the following regression output.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2159.498   8877.810   0.243 0.808230
BrickYes    17297.350   1981.616   8.729 1.78e-14 ***
N2          -1560.579   2396.765  -0.651 0.516215
N3          20681.037   3148.954   6.568 1.38e-09 ***
Offers      -8267.488   1084.777  -7.621 6.47e-12 ***
SqFt           52.994      5.734   9.242 1.10e-15 ***
Bedrooms     4246.794   1597.911   2.658 0.008939 **
Bathrooms    7883.278   2117.035   3.724 0.000300 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 10020 on 120 degrees of freedom
Multiple R-squared: 0.8686,     Adjusted R-squared: 0.861
F-statistic: 113.3 on 7 and 120 DF,  p-value: < 2.2e-16
```

```
                   2.5 %      97.5 %
(Intercept) -15417.94711 19736.94349
BrickYes     13373.88702 21220.81203
N2           -6306.00785  3184.84961
N3           14446.32799 26915.74671
Offers      -10415.27089 -6119.70575
SqFt            41.64034    64.34714
Bedrooms      1083.04162  7410.54616
Bathrooms     3691.69572 12074.86126
```

To check if there is a premium for brick houses given everything else being equal we test the hypothesis that $\beta_1 = 0$ at the 95% confidence level. Using the regression output we see that the 95% confidence interval for β_1 is $[13373.89, 21220.91]$. Since this does not include zero we conclude that brick is a significant factor when pricing a house. Further, since the entire confidence interval is greater than zero we conclude that people pay a premium for a brick house.

(b) To check that there is a premium for houses in Neighborhood three, given everything else we repeat the procedure from part (1), this time looking at β_3. The regression output tells us that the confidence interval for β_3 is $[14446.33, 26915.75]$. Since the entire confidence interval is greater than zero we conclude that people pay a premium to live in neighborhood three.

(c) We want to determine if Neighborhood 2 plays a significant role in the pricing of a house. If it does not, then it will be reasonable to combine neighborhoods one and two into one "old" neighborhood. To check if Neighborhood 2 is important, we perform a hypothesis test on $\beta_2 = 0$. The null hypothesis $\beta_2 = 0$ corresponds to the dummy variable N_2 being unimportant. Looking at the confidence interval from the regression output we see that the 95% confidence interval for β_2 is $[-6306, 3184]$, which includes zero. Thus we can conclude that it is reasonable to let β_2 be zero and that neighborhood 2 may be combined with neighborhood 1.

(d) To check that there is a premium for brick houses in neighborhood three we need to alter our model slightly. In particular, we need to add an interaction term $Brick \times N3$. This more complicated model is

$$\begin{aligned} Y = \beta_0 + \beta_1\, Brick + \beta_2\, N_2 + \beta_3\, N_3 + \beta_4\, Bids \\ + \beta_5\, SqFt + \beta_6\, Bed + \beta_7 Bath + \beta_8\, Brick \cdot N_3 + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma^2). \end{aligned}$$

To see what this interaction term does, observe that

$$\frac{\partial E[Y|Brick, N_3]}{\partial N_3} = \beta_3 + \beta_8\, Brick.$$

Thus if β_8 is non-zero we can conclude that consumers pay a premium to buy a brick house when shopping in neighborhood three. The output of the regression which includes the interaction term is below.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3009.993   8706.264   0.346  0.73016
BrickYes    13826.465   2405.556   5.748 7.11e-08 ***
N2           -673.028   2376.477  -0.283  0.77751
N3          17241.413   3391.347   5.084 1.39e-06 ***
Offers      -8401.088   1064.370  -7.893 1.62e-12 ***
SqFt           54.065      5.636   9.593  < 2e-16 ***
Bedrooms     4718.163   1577.613   2.991  0.00338 **
Bathrooms    6463.365   2154.264   3.000  0.00329 **
BrickYes:N3 10181.577   4165.274   2.444  0.01598 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 9817 on 119 degrees of freedom
Multiple R-squared: 0.8749,     Adjusted R-squared: 0.8665
F-statistic:   104 on 8 and 119 DF,  p-value: < 2.2e-16
```

```
                   0.5 %      99.5 %
(Intercept) -19781.05615 25801.04303
BrickYes      7529.25747 20123.67244
N2           -6894.11333  5548.05681
N3            8363.62557 26119.20030
Offers      -11187.37034 -5614.80551
SqFt            39.31099    68.81858
Bedrooms       588.32720  8847.99967
Bathrooms      823.98555 12102.74436
BrickYes:N3   -722.17781 21085.33248
```


```
                   0.5 %      99.5 %
(Intercept) -19781.05615 25801.04303
BrickYes      7529.25747 20123.67244
N2           -6894.11333  5548.05681
N3            8363.62557 26119.20030
Offers      -11187.37034 -5614.80551
SqFt            39.31099    68.81858
Bedrooms       588.32720  8847.99967
Bathrooms      823.98555 12102.74436
BrickYes:N3   -722.17781 21085.33248
```


To see if there is a premium for brick houses in neighborhood three we check that the 95% confidence interval is greater than zero. Indeed, we calculate that the 95% confidence interval is $[1933, 18429]$. Hence we conclude that there is a premium at the 95% confidence level. Notice however, that the confidence interval at the 99% includes zero. Thus if one was very stringent about drawing conclusions from statistical data, they may accept the claim that there is no premium for brick houses in neighborhood three.

(e) Based on this model, controlling all the other explanatory variables, the increase of the number of bids by one corresponds to an expected decrease of over $8000 for the selling price. This model is not suitable to predict the selling price of a house that is to be put on the market, since the number of bids for the house would be unknown at that time. One may consider excluding "Bids" from the regression model.

Problem 2

The data file **Profits.csv** is available in the course website. It contains information on 18 projects developed at a firm. The variables included in the file are:

- **Profit**: profit of the project in thousands of dollars
- **RD**: expenditure on research and development for the project in thousands of dollars
- **Risk**: a measure of risk assigned at the outset of the project

(a) Regress **Profit** on **RD** and **Risk**. Does there appear to be a relationship between **Profit** and **RD** after the risk of the project has been controlled for?

(b) Plot residuals versus fitted values. Do any of the regression assumptions appear to be violated? If yes, state which assumptions and justify your answer.

(c) Plot residuals versus **RD** and residuals versus **Risk**. Based on these plots suggest a correction for any violation detected in (b)? Try implementing your suggested correction. Does your new model appear to be an improvement over the original model? Justify your answer.

(d) What does your new model suggest is the expected change in **Profit** when **RD** changes? What is the expected change in **Profit** when **Risk** changes?

(a) We want to know if there is a relationship between Profit and RD after Risk has been controlled for. To figure this out we use the model

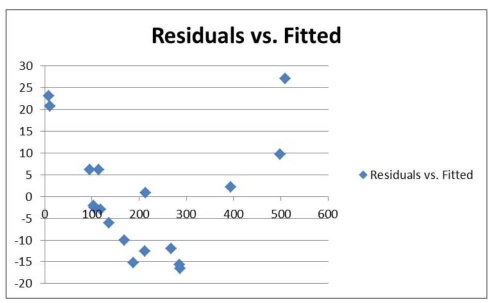
$$Profit = \beta_0 + \beta_1 Risk + \beta_2 RD + \epsilon,\ \epsilon \sim N(0, \sigma^2).$$

Risk is included in the model because we want to take that into account. If we did not include Risk in the model, it would not be controlled for. The table below shows the output of the regression.

	Coefficients	Standard Err	t Stat	P-value	Lower 95%	Upper 95%	Lower 99.0%	Upper 99.0%
Intercept	-453.176	23.50614	-19.2791	5.37E-12	-503.278	-403.074	-522.442	-383.91
RISK	29.30904	3.668586	7.989194	8.76E-07	21.48964	37.12845	18.49877	40.11931
RD	4.510005	0.15375	29.33333	1.16E-14	4.182294	4.837715	4.056947	4.963062

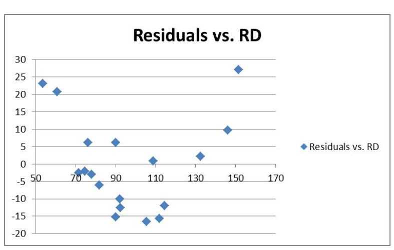
Profit is not linearly related to RD if $\beta_2 = 0$. To test the hypothesis that $\beta_2 = 0$ at the 95% confidence level we check if zero is in the 95% confidence interval. From the table we see that the interval for β_2 is $[4.2, 4.8]$, which does not include zero. Hence we can conclude that there appears to be a linear relationship between Profit and RD after the risk has been controlled for.

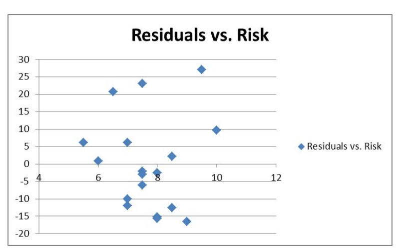
(b) The plot of the residuals verse the fitted values is below.



It appears that the smallest and largest fitted values are consistently above zero, while the fitted values in the middle are consistently below zero. This is evidence against randomness we expect to see in the residuals.

(c) The plots of residuals versus RD and residuals versus Risk are plotted below.



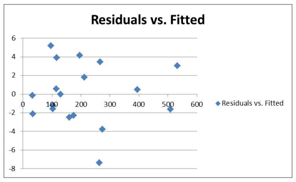


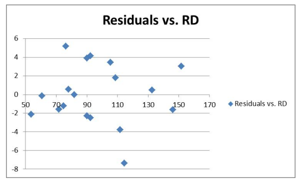
The plot of residuals versus Risk appears to be random. However, the plot of residuals vs. RD has the appearance of a parabola. This suggests that we should include the term RD^2 in our regression.

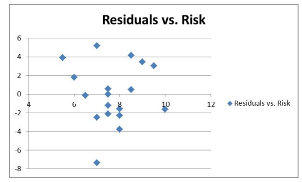
The improved model we will use is

$$Profit = \beta_0 + \beta_1 Risk + \beta_2 RD + \beta_3 RD^2 + \epsilon,\ \epsilon \sim N(0, \sigma^2).$$

We have re-plotted the residuals versus the fitted values, RD, and Risk under this model.







It now appears that the residuals versus the fitted values and the residuals versus RD are random. Since it appears that the new model has removed some systematic errors in our previous model, we may cautiously conclude that it is better.

(d) To figure out how expected Profit changes when RD or Risk changes we look at the derivative of the expected profit with respect to RD and with respect to Risk. The expected Profit given RD and Risk is

$$E[Profit|RD, Risk] = \beta_0 + \beta_1 Risk + \beta_2 RD + \beta_3 RD^2.$$

Taking the derivative with respect to RD we have that

$$\frac{\partial E[Profit|RD, Risk]}{\partial RD} = \beta_2 + 2\beta_3 RD.$$

Thus the expected change in Profit given a small change in RD, Δ_{RD}, is $(\beta_2 + 2\beta_3 RD)\ \Delta_{RD}$.

This implies that the impact of RD onto $Profits$ depends on the level of RD. Given that we estimate β_3 to be positive, it means that for large levels of RD the impact on $Profits$ is going to be larger.

Taking the derivative with respect to Risk we have that

$$\frac{\partial E[Profit|RD, Risk]}{\partial Risk} = \beta_1.$$

Thus the expected change in Profit given a small change in Risk, Δ_{Risk}, is $\beta_1\ \Delta_{Risk}$.

Problem 3

The data file **AutoMPG.csv** is available in the course website. It contains information on **MPG** (miles per gallon) and **Weight** (in pounds) of 392 cars.

(a) Regress **MPG** on **Weight**. What does your model suggest is the expected change in **MPG** when **Weight** changes?

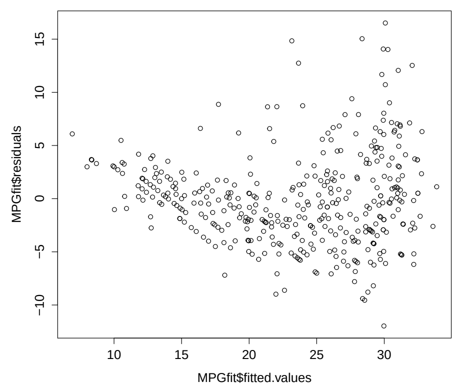
(b) Plot residuals versus fitted values. Do any of the regression assumptions appear to be violated? If yes, state which assumptions and justify your answer.

(c) Regress **log(MPG)** on **log(Weight)**. What does your model suggest is the expected change in **MPG** when **Weight** changes?

(d) Plot residuals versus fitted values. Do any of the regression assumptions appear to be violated? If yes, state which assumptions and justify your answer.

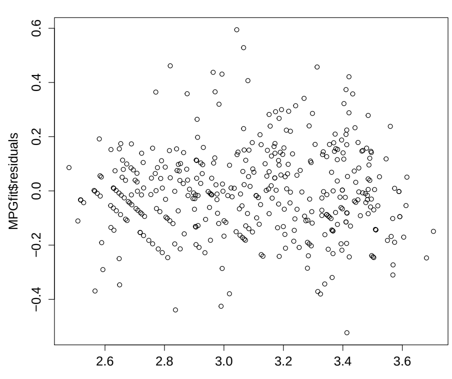
(a) MPG = 46.22 − 0.00765*Weight + ϵ. MPG decrease by 7.65 when weight increases by 1000 pounds.

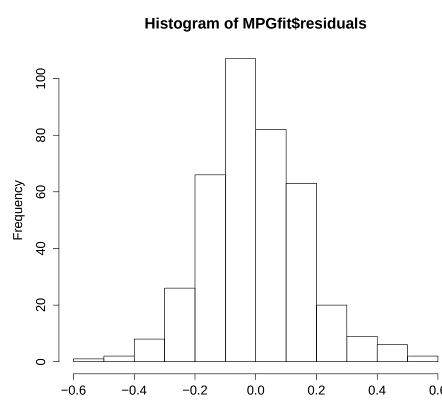
(b) Yes, the constant variance assumption is clearly violated. Note that we assume that the errors are independent, and identically distributed as $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.



(c) log(MPG) = 11.52 − 1.06*log(Weight) + ϵ. MPG decrease by 1.06% when weight increases by 1%.

(d) The regression assumptions appear to be well satisfied after doing the log transformations for both MPG and Weight.





Problem 4

Read the "Oakland A's (A)" case in the course pack. The data is available in the course website. The tab in the spreadsheet labeled Full Data Set contains the data in Exhibit 1 of the case while the tab labeled Nobel Data contains the attendance figures for the games Nobel pitched in and those he did not pitch in.

(a) Compute the descriptive statistics for the attendance at the games Nobel pitched in and those he did not pitch in. What is the difference in the average attendance for these two sets of games? Does this provide meaningful evidence that Nobel should be paid more because attendance was higher in the games he pitched in?

(b) Plot Ticket against Time (i.e. create a time series plot of Ticket). Do you see any patterns in the data?

(c) Run the regression

$$Ticket_t = \beta_0 + \beta_1 Nobel_t + \epsilon_t$$

where $Nobel$ is a dummy variable that takes the value 1 when Nobel starts on day t.

What are the estimates of β_0 and β_1? How do these relate to the average attendance figures computed in part (a)?

(d) Do the residuals from the regression in part (c) appear to be independent? Why or why not? If they are not independent, what factors might explain the pattern?

(e) Run the regression

$$Ticket_t = \beta_0 + \beta_1 Pos_t + \beta_2 GB_t + \beta_3 Temp_t + \beta_4 Prec_t + \beta_5 TOG_t + \beta_6 TV_t + \\ + \beta_7 Promo_t + \beta_8 Nobel_t + \beta_9 Yanks_t + \beta_{10} Weekend_t + \beta_{11} OD_t + \beta_{12} DH_t + \epsilon_t$$

Do the residuals from this regression appear to be independent? (It is a close call but assume they are independent.) Why would these residuals be independent while the residuals from the model in part (c) are dependent?

(f) What evidence is there about Nobel pitching in a game being related to the attendance at the game? Do you have more confidence in drawing a conclusion from the model in part (c) or the model in part (e) to answer this question? Why?

(g) Do you think Nobel's agent has a legitimate case that Nobel should be paid more because he brings fans to the games?

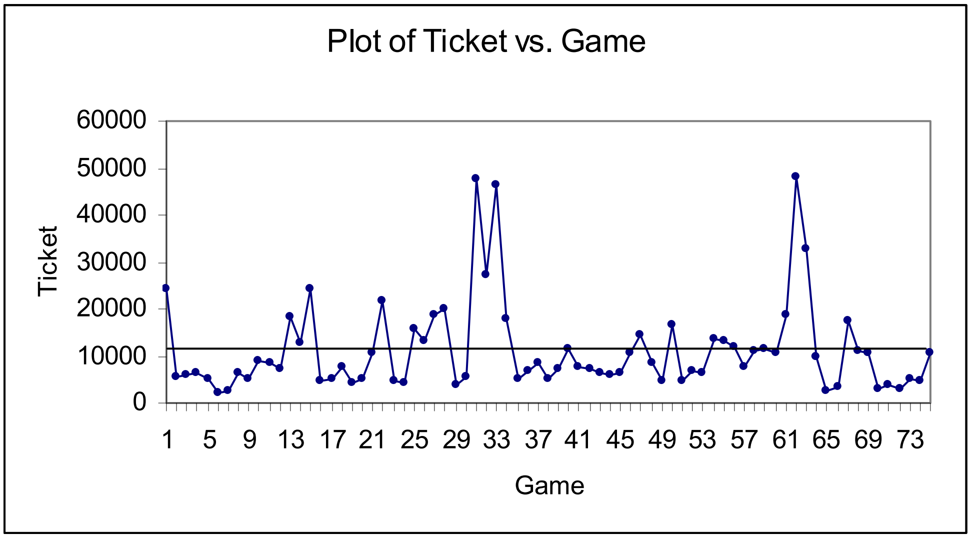
(a)

	Tickets when Nobel Pitches	*Tickets* when Nobel Doesn't Pitch
Average	12663	10859
Std. Dev	11211	9357

The difference in average attendance is 1804. On average, the attendance is 1804 fans higher in games that Nobel pitches.

This does not provide meaningful evidence that Nobel should be paid more because the increase in attendance may be explained by factors other than whether Nobel is pitching or not (for example, it may be that he tends to pitch on days when the Yankees are in town and it is the Yankees that are drawing the extra fans).

(b) There is a pattern in the plot because there are several runs of three or four games where attendance is considerably higher than the mean. The pattern can be partially explained by factors such as the team the A's are playing, whether the games are played on a weekend (Friday, Saturday and Sunday), etc. For example, the A's played the Yankees in games 31-33 and it was also a weekend series.



(c) The estimates of β_0 and β_1 are 10,859.4 and 1,804.2, respectively. The estimate of β_0 is the sample mean attendance for games Nobel did not pitch in. The estimate of β_1 is the difference between the sample mean attendance for games that Nobel pitched in (which is 12,663.6) and the sample mean attendance for games that he did not pitch in (which is 10,859.4), i.e. the estimate of β_1 is 12,663.6 $-$ 10,859.4 = 1,804.2.

SUMMARY OUTPUT

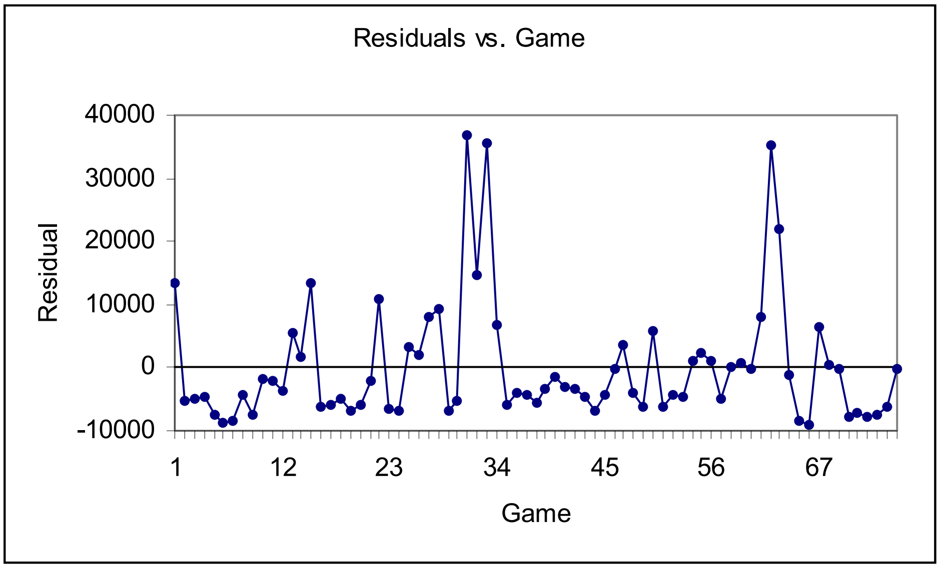
Regression Statistics	
Multiple R	0.076474915
R Square	0.005848413
Adjusted R Square	-0.007770102
Standard Error	9767.591743
Observations	75

ANOVA

	df	*SS*	*MS*
Regression	1	40971630.72	40971630.72
Residual	73	6964626937	95405848.46
Total	74	7005598568	

	Coefficients	*Standard Error*	*t Stat*
Intercept	10859.35593	1271.632132	8.539699226
NOBEL	1804.206568	2753.164326	0.655321061

(d) The residuals do not appear to be independent because there are persistent runs above and below zero. Factors that might explain the pattern are factors that are not included in the regression such as whether the A's are playing the Yankees, whether the games are played on a weekend, whether the weather conditions are good or bad, etc.



(e) The residuals appear to be independent, although it is a close call. While there are some runs above and below zero, there are also some periods where the residuals oscillate above and below zero too fast. This means there is no persistent pattern of runs or oscillations that persists throughout the graph.

The factors inducing the dependence in the residuals for the model in part (c) (i.e. which team the A's are playing, whether or not the game is a weekend game, weather conditions, etc.) are incorporated directly into the model in part (e). This removes the effect of these factors from the residuals and incorporates the explanatory power of the factors directly into the regression model.

(f) There is no statistical evidence that Nobel's pitching in a game is related to an increase in attendance. In fact, the negative estimate of β_8 (the coefficient associated with the dummy variable Nobel in part (e)) indicates that when all other important factors are held constant (i.e. whether the A's are playing the Yankees, whether the games are played on a weekend, etc.) attendance is actually slightly lower for games Nobel pitches in. We have more confidence using the model in part (e) than the one in part (c) because the model in part (e) properly accounts for many of the factors that are related to attendance.

(g) Nobel's agent does not have a legitimate case based on the statistical evidence. When factors related to attendance are accounted for such as the teams the A's are playing, whether the game is a weekend game, weather conditions, etc., average attendance is actually slightly lower in the games Nobel pitches in than in the games he does not pitch in (see the estimated coefficient of -1,159.6 associated with Nobel in the regression in part (e)).

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.873815476
R Square	0.763553486
Adjusted R Square	0.717789645
Standard Error	5168.841761
Observations	75

ANOVA

	df	SS	MS
Regression	12	5349149209	445762434.1
Residual	62	1656449359	26716925.15
Total	74	7005598568	

	Coefficients	Standard Error	t Stat
Intercept	28672.81397	15224.26613	1.883362635
POS	-790.9956716	483.9146844	-1.634576707
GB	98.44232763	139.4133348	0.706118448
TEMP	-363.6261197	259.4137036	-1.401722865
PREC	-5763.004275	3336.141385	-1.727446055
TOG	1950.4875	1316.683459	1.481364018
TV	339.3057503	1970.590748	0.172184788
PROMO	4694.284338	1644.49703	2.854541086
NOBEL	-1159.639474	1521.793262	-0.76202169
YANKS	28909.46388	2504.344094	11.54372674
WKEND	1254.340404	1313.960474	0.954625675
OD	16424.43589	5490.190773	2.991596571
DH	5735.29646	2335.944448	2.455236667

Problem 5

Read the "Oakland A's (B)" case in the course packet. The data file is available in the course website.

(a) Run a regression of Attendance against Wins. What is the interpretation of the coefficient associated with Wins? What is the interpretation of R^2 in this regression? What is the practical problem associated with using this model to forecast Attendance for the next season (i.e. to forecast attendance in the 1981 season)?

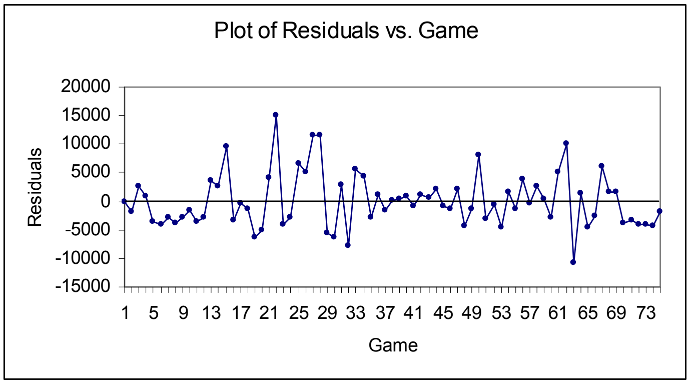
(b) Now run a regression of Attendance against Roddey's forecast of the number of wins for that season. Why is the R^2 value obtained from this regression so much lower than the R^2 obtained from the regression in part (a)?

(c) Why is it more appropriate to use the model in part (b) for forecasting Attendance than the model in part (a)?

(d) Before the 1981 season starts Roddey forecasts 95 wins for the season. Using the model from part (b), what is the prediction for attendance in the 1981 season? What is the standard deviation associated with the prediction?

(e) Using the prediction and standard deviation for the prediction from the model in part (b), what is the probability associated with a bonus to Nobel of \$0, \$50,000, \$100,000 and \$150,000? What is the mean of this distribution?

(e) Using the probability distribution from part (d), what is the expected cost if the lump-sum incentive plan is used?



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.948332499
R Square	0.899334529
Adjusted R Square	0.890183123
Standard Error	72137.41936
Observations	13

ANOVA

	df	*SS*	*MS*
Regression	1	5.11393E+11	5.11393E+11
Residual	11	57241879985	5203807271
Total	12	5.68635E+11	

	Coefficients	*Standard Error*	*t Stat*
Intercept	-464710.1155	127193.5213	-3.653567498
Wins	14837.47159	1496.729049	9.91326493

(a) The interpretation of the coefficient associated with $Wins$ is that for each additional win the A's can expect attendance to increase, on average, by 14,837.5 fans. The $R^2 2$ value is 0.899 so 89.9% of the variability in annual attendance can be explained by the variability in the number of games the A's win in each season. The practical problem associated with using this model to forecast Attendance for next season is that the number of wins for the season is unknown. This means the number of wins must be forecasted and this is likely to induce considerable error in the forecast of Attendance.

(b) The R^2 value from the regression of $Attendance$ against $WinsForecasted$ is lower than the R^2 value from the regression in part (a) because there is not as much information in $WinsForecasted$ about $Attendance$ as there is in the actual number of $Wins$. If there is not as much information, the R^2 value will be lower.

(c) The model in part (b) is more appropriate because the relationship between *Attendance* and the number of wins forecast by Roddey is the only relevant relationship for prediction purposes. The reason is that only Roddey's forecast of the number of wins is available at the beginning of 1981 when Attendance for the 1981 season needs to be predicted actual wins in 1981 are not available when predicting Attendance before the season starts.

(d) The prediction for attendance when the forecasted wins is 95 is

$$-430,201.8 + 14,413.1(95) = 939,042.7$$

and the estimate of the standard deviation associated with the prediction is 145,907.4

(e) Using the model in part (b) and the prediction in part (d) the probability Nobel will receive a bonus of:

1. \$0 is $Pr(Attendance < 1,000,000) = 0.6628$
2. \$50,000 is $Pr(1,000,000 < Attendance < 1,500,000) = 0.3372$
3. \$100,000 is $Pr(1,500,000 < Attendance < 2,000,000) = 0.000$
4. \$150,000 is $Pr(Attendance > 2,000,000) = 0.000$

Therefore, the expected bonus for Nobel is:

$$(0 \times 0.6628) + (50,000 \times 0.3372) + (100,000 \times 0.000) + (150,000 \times 0.000) = 16,860$$

(f) The expected cost of the lump-sum incentive plan is \$16,860 (i.e. the expected cost is the mean of the probability distribution reflecting the uncertainty in the bonus that will need to be paid to Nobel).

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.766924826
R Square	0.588173689
Adjusted R Square	0.550734933
Standard Error	145907.4216
Observations	13

ANOVA

	df	*SS*	*MS*
Regression	1	3.34456E+11	3.34456E+11
Residual	11	2.34179E+11	21288975668
Total	12	5.68635E+11	

	Coefficients	*Standard Error*	*t Stat*
Intercept	-430201.8462	308121.5075	-1.39620843
WinsForecasted	14413.07143	3636.339678	3.963620757