STA371G Homework Assignment 5

Problem 1 (5 points)

Suppose we are modeling house price as depending on house size, the number of bedrooms in the house and the number of bathrooms in the house. Price is measured in thousands of dollars and size is measured in thousands of square feet.

Suppose our model is:

$$P = 20 + 50\,\text{size} + 10\,\text{nbed} + 15\,\text{nbath} + \epsilon, \;\; \epsilon \sim N(0, 10^2).$$

(a) Suppose you know that a house has size =1.6, nbed = 3, and nbath =2.

What is the distribution of its price given the values for size, nbed, and nbath.

(hint: it is normal with mean = ?? and variance = ??)

$20 + 50 \times 1.6 + 10 \times 3 + 15 \times 2 = 160$
$P = 160 + \epsilon$ so that $P \sim N(160, 10^2)$

(b) Given the values for the explanatory variables from part (a), give the 95% predictive interval for the price of the house.

160 ± 20

(c) Suppose you know that a house has size =2.6, nbed = 4, and nbath =3. Give the 95% predictive interval for the price of the house.

$20 + 50 \times 2.6 + 10 \times 4 + 15 \times 3 = 235$
$P = 235 + \epsilon$ so that $P \sim N(235, 10^2)$ and the 95% predictive interval is
235 ± 20

(d) In our model the slope for the variable nbath is 15. What are the units of this number?

Thousands of dollars per bathroom.

(e) What are the units of the intercept 20? What are the units of the the error standard deviation 10?

The intercept has the same units as P... in this case, thousands of dollars. The error std deviation is also in the same units as P, ie, thousands of dollars.

Problem 2 (5 points)

The data for this question is in the file **Profits.csv**, which can be found in the course website.

There are 18 observations.
Each observation corresponds to a project developed by a firm.
y = Profit: profit on the project in thousands of dollars.
x1= RD: expenditure on research and development for the project in thousands of dollars.
x2=Risk: a measure of risk assigned to the project at the outset.

We want to see how profit on a project relates to research and development expenditure and "risk".

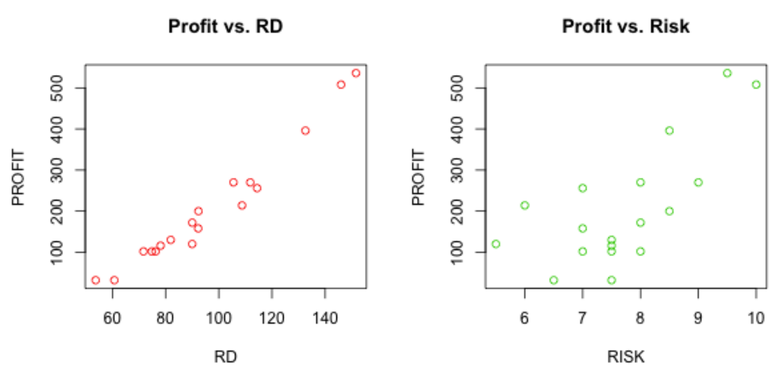
(a) Plot profit vs. each of the two x variables. That is, do two plots y vs. x1 and y vs x2. You can't really understand the full three-dimensional relationship from these two plots, but it is still a good idea to look at them. Does it seem like the y is related to the x's?

(b) Suppose all you knew was risk=7. Run the simple linear regression of profit on risk and get the 68% plug-in predictive interval for profit.

(c) Suppose a project has risk=7 and research and development = 76. Give the 95% plug-in predictive interval for the profit on the project. Compare that to the correct, predictive interval (using the predict function in R).

(d) How does the size of your interval in (c) compare with the size of your interval in (b)? What does this tell us about our variables?

(a) It seems like there is some relationship, especially between RD and profit.



(b) Using the model $PROFIT = \beta_0 + \beta_1 RISK + \epsilon$, the least squares estimates of regression coefficients are $b_0 = \hat{\beta}_0 = -489.53$ and $b_1 = \hat{\beta}_1 = 90.45$, and the regression standard error is $s = \hat{\sigma} = 106.1$. Thus the 68% plug-in prediction interval for when $RISK = 7$ is $143.6 \pm 106.1 = [37.5, 249.7]$.

(c) Using the model $PROFIT = \beta_0 + \beta_1 RISK + \beta_2 RD + \epsilon$, the least squares estimates of regression coefficients are $b_0 = \hat{\beta}_0 = -453.18$, $b_1 = \hat{\beta}_1 = 29.31$ and $b_2 = \hat{\beta}_2 = 4.51$, and the regression standard error is $s = \hat{\sigma} = 14.34$. The 95% plug-in predictive interval, when $RD = 76$ and $RISK = 7$ is $94.75 \pm 2 * 14.34 = [66.1, 123.4]$.

You may use the following R code to find the correct prediction interval:

```
setwd("~/yourfolder")
##change this to your working directly, where the CSV file is stored.
data = read.csv("Profits.csv", header=TRUE)
attach(data)
Fit = lm(PROFIT~RISK+RD)
new=data.frame(RISK=7,RD=76)
predict(Fit, new, interval = "prediction")
```

(d) Our interval in (b) is bigger than the interval in (c) despite the fact that it is a "weaker" confidence interval. In essence (b) says that we predict Y will be in $[38, 250]$ 68% of the time when $RISK = 7$. In contrast, (c) says that Y will be in $[63, 127]$ 95% of the time when $RISK = 7$ and $RD = 76$. Using RD in our regression narrows our prediction interval by quite a bit.

Problem 3 (10 points)

The data for this question is in the file **zagat.xls**, which can be found in the course web-site. The data is from the Zagat restaurant guide. There are 114 observations and each observation corresponds to a restaurant.
There are 4 variables:
price: the price of a typical meal
food: the zagat rating for the quality of food.
service: the zagat rating for the quality of service.
decor: the zagat rating for the quality of the decor.

We want to see how the price of a meal relates the quality characteristics of the restaurant experience as measured by the variables food, service, and decor.

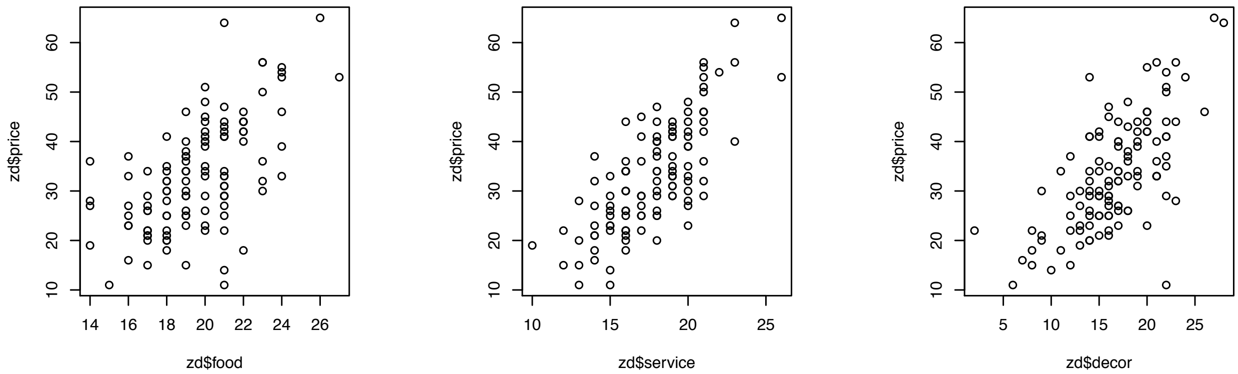
(a) Plot price vs. each of the three x's. Does it seem like our y (price) is related to the x's (food, service, and decor) ?

(b) Suppose a restaurant has food = 18, service=14, and decor=16. Run the regression of price on food, decor, and service and give the 95% predictive interval for the price of a meal.

(c) What is the interpretation of the coefficient estimate for the explanatory variable food in the multiple regression from part (b) ?

(d) Suppose you were to regress price on the one variable food in a simple linear regression? What would be the interpretation of the slope? Plot food vs. service. Is there a relationship? Does it make sense? What is your prediction for how the estimated coefficient for the variable food in the regression of price on food will compare to the estimated coefficient for food in the regression of price on food, service, and decor? Run the simple linear regression of price on food and see if you are right! Why are the coefficients different in the two regressions?

(e) Suppose I asked you to use the multiple regression results to predict the price of a meal at a restaurant with food = 20, service = 3, and decor =17. How would you feel about it?



Solutions.

(a) Check out the figure above... definitely looks like price is related to each of the 3 X's.

(b) The regression output is

Regression Statistics	
Multiple R	0.829
R Square	0.687
Adjusted R	0.679
Standard E	6.298
Observatio	114.000

ANOVA

	df	SS	MS	F	Significance F
Regression	3.000	9598.887	3199.629	80.655	0.000
Residual	110.000	4363.745	39.670		
Total	113.000	13962.632			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-30.664	4.787	-6.405	0.000	-40.151	-21.177
food	1.380	0.353	3.904	0.000	0.679	2.080
decor	1.104	0.176	6.272	0.000	0.755	1.453
service	1.048	0.381	2.750	0.007	0.293	1.803

so that $-30.66 + 1.38 \times 18 + 1.1 \times 16 + 1.05 \times 14 = 26.476$ and the 95% plug-in prediction interval is 26.476 ± 12.6

(c) If you hold service and decor constant and increase food by 1, then price goes up (on average) by 1.38.

(d) If food goes up by 1 price goes up by the slope (on average)... from the plot in item (a) we know that it looks like food and price are related in a positive way. Now, you would think that these four variables are somewhat related to each other, right? A better restaurant tend to have good food, service and decor... and also a higher price. By running the regression with only food as a explanatory variable I would guess the coefficient for food would be higher... let's see:

Regression Statistics	
Multiple R	0.599
R Square	0.359
Adjusted R Squ	0.353
Standard Error	8.939
Observations	114.000

ANOVA

	df	SS	MS	F	Significance F
Regression	1.000	5012.239	5012.239	62.720	0.000
Residual	112.000	8950.393	79.914		
Total	113.000	13962.632			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-18.154	6.553	-2.770	0.007	-31.137	-5.170
food	2.625	0.331	7.920	0.000	1.968	3.282

I was right! In the simple linear, regression food works as a proxy for the overall quality of a restaurant. When food goes up service and decor tend to go up as well but since they are not in the regression, the coefficient for food has to reflect the other factors. Once decor and service are in the regression, the coefficient for food just has to reflect the impact associated with food but not with the other variables.

(e) Very bad! We just dont see in our data restaurants with that low of a service rating given food equal to 20 and decor equal to 17. This would be a extreme extrapolation from what we have seen so far and the model might not be appropriate.

Problem 4: Baseball (10 points)

Using our baseball data (**RunsPerGame.xls**), regress R/G on a binary variable for league membership (League = 0 if National and League = 1 if American) and OBP.

$$R/G = \beta_0 + \beta_1 League + \beta_2 OBP + \epsilon$$

(a) Based on the model assumptions, what is the expected value of R/G given OBP for teams in the AL? How about the NL?

(b) Interpret β_0, β_1 and β_2.

(c) After running the regression and obtaining the results, can you conclude with 95% probability that the marginal effect of OBP on R/G (after taking into account the League effect) is positive?

(d) Test the hypothesis that $\beta_1 = 0$ (with 99% probability). What do you conclude?

(a) The expected value of R/G given OBP is

$$E\Big[R/G|OBP, League = 0\Big] = \beta_0 + \beta_2 OBP$$

for the NL and

$$E[R/G|OBP, League = 1] = (\beta_0 + \beta_1) + \beta_2 OBP$$

for the AL.

(b) β_0 is the number of runs per game we expect a team from the National League to score if their OBP is zero.

We expect a team in the American League to score β_1 more runs per game on average than a team in the National League with the same OBP.

β_2 tells us how R/G scales with OBP. For every unit increase in OBP there will be a β_2 increase in R/G.

(c) The 95% confidence interval for β_2 is 37.26±2*2.72 = (31.82;42.70) hence, yes, it is greater than zero.

(d) The best guess of β_1 is $b_1 = 0.01615$ with standard error 0.06560. Thus the 99% confidence interval is $b_1 \pm 3 * s_{b_1} = [-0.18, 0.21]$, which includes zero. Since zero is in our interval of reasonable values we cannot conclude that $\beta_1 \neq 0$.

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -7.72065    0.93031  -8.299 6.59e-09 ***
LeagueAmerican  0.01615    0.06560   0.246    0.807
OBP            37.26060    2.72081  13.695 1.14e-13 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1712 on 27 degrees of freedom
Multiple R-squared: 0.8851,     Adjusted R-squared: 0.8765
F-statistic: 103.9 on 2 and 27 DF,  p-value: 2.073e-13
```

Problem 5 (10 points)

Read the case "Orion Bus Industries: Contract Bidding Strategy" in the course packet. Orion Bus Industries wants to develop a method for determining how to bid on specific bus contracts to maximize expected profits. In order to do this, it needs to develop a model of winning bids that takes into account such factors as the number of buses in the contract, the estimated cost of the buses and the type of bus (e.g. length, type of fuel used, etc.). The data set is available in the course website. This data set only includes the bus contracts from Exhibit 1 in the case where Orion did not win the contract. This eliminates 28 of the 69 observations and leaves a sample of size n = 41 observations.

(a) Run a regression of $WinningBid$ against $NumberOfBusesInContract$, $OrionsEstimatedCost$, $Length$, $Diesel$ and $HighFloor$, ie, the following regression model:

$$WinningBid_i = \beta_0 + \beta_1 NumberOfBusesInContract_i + \beta_2 OrionsEstimatedCost_i + \beta_3 Length_i + \beta_4 Diesel_i + \beta_5 HighFloor_i + \epsilon_i$$

What is the estimated regression model? How would you interpret the estimated coefficient associated with the dummy variable Diesel?

(b) What is the estimate of σ^2 in the model in part (a)?

The city of Louisville, Kentucky is putting out a contract for bid for five 30-foot, low-floor, diesel-fuelled buses. Orion estimates their cost to manufacture these buses to be $234,229 per bus.

(c) Using the model in part (a), what is the distribution representing the uncertainty about the amount of the winning bid per bus for this contract? In particular, what are the mean and standard deviation of the distribution?

(d) Given the distribution in part (c), what is the probability that Orion wins the contract if it bids $240,000 per bus? If it wins the contract, what is its profit per bus per bus?

(e) What is the probability that Orion loses the contract if it bids $240,000 per bus? If it loses the contract, what is its profit per bus? (You do not need to take into account the cost of putting the bid together when determining the profit for a lost contract.)

(f) Why is there uncertainty about the profit per bus that Orion will obtain if it bids $240,000 per bus? What is the probability distribution representing this uncertainty? In particular, what is the mean of the distribution (i.e. what is the expected profit per bus if it bids $240,000 per bus)?

We now want to develop an Excel spreadsheet (or program in R) that will allow ExpectedProfit to be plotted against different possible bid amounts (i.e. $240,000; $241,000; ...; $260,000). The maximum of this graph will give Orion the bid amount that will maximize expected profit.

(g) Using the plot, what should Orion bid if it wants to maximize expected profit per bus?

(a) The Excel output for this regression model is:

Regression Statistics	
Multiple R	0.902304784
R Square	0.814153923
Adjusted R Square	0.787604483
Standard Error	11721.15707
Observations	41

ANOVA

	df	*SS*	*MS*
Regression	5	21065032698	4213006540
Residual	35	4808493307	137385523.1
Total	40	25873526005	

	Coefficients	*Standard Error*	*t Stat*
Intercept	-13872.72734	26062.71483	-0.532282513
NumberOfBusesInContract	42.32044997	219.3318353	0.192951698
OrionsEstimatedCost	0.813616165	0.073356177	11.09131092
Length	1949.968943	456.482292	4.271729652
Diesel	11240.97951	6172.434639	1.821158128
HighFloor	8175.562414	4353.019803	1.878135819

The interpretation of the estimated coefficient for *Diesel* is the following:

First, the true coefficient β_4 is the expected increase, on average, in the winning bid when the buses specified in the contract run on diesel fuel rather than natural gas, holding all other variables constant.

11241.0 is the estimate for β_4 (ie b_4 in our notation) so that 11241.0 is the estimate of the expected increase, on average, in the winning bid when the buses specified in the contract run on diesel fuel rather than natural gas, holding all other variables constant.

But we should also notice that zero is within the 95% confidence interval of β_4 and hence there is no strong evidence to suggest that a diesel bus sells at a premium.

(b) $s^2 = 11721.15^2$

(c) For a contract with five 30-foot, low-floor, diesel-fuelled buses and an estimated cost of $234,229 per bus, the explanatory variables take on the following values:

$NumberOfBusesInContract = 5$; $OrionsEstimatedCost = 234,229$; $Length = 30$; $Diesel = 1$ and $HighFloor = 0$.

Given the estimates from (a), the estimated mean of the distribution is
-13872.7 + 42.3204(5) + 0.813616(234229) + 1949.97(30) + 11241.0(1) + 8175.56(0)
= 246651.5.
so that the distribution of the winning bid can be represented by

$$WinningBid \sim N(246651.5, 11721^2)$$

(d) To find the probability that Orion wins the contract if it bids $240,000 per bus we need to compute the following probability (note that LowBid is the same as WinningBid but is a bit more descriptive of what the above regression provides):

$$\begin{aligned}
Pr(\text{Win Contract}) &= Pr(\text{Low Bid} > 240000) \\
&= Pr\left(\frac{\text{Low Bid} - 246651.5}{11721} > \frac{240000 - 246651.5}{11721}\right) \\
&= Pr(Z > -0.57) \\
&= 1 - Pr(Z \leq -0.57) \\
&= 0.7146 \qquad (1)
\end{aligned}$$

If Orion wins the contract, Profit (which is the difference between the bid amount of $240,000 and the cost of $234,229) is $5,771.

(e) The probability that Orion loses the contract is

$$Pr(\text{Lose Contract}) = 1 - Pr(\text{Win Contract}) = 0.2854$$

If Orion loses the contract, then it receives no revenue and has no production costs so its Profit is 0.

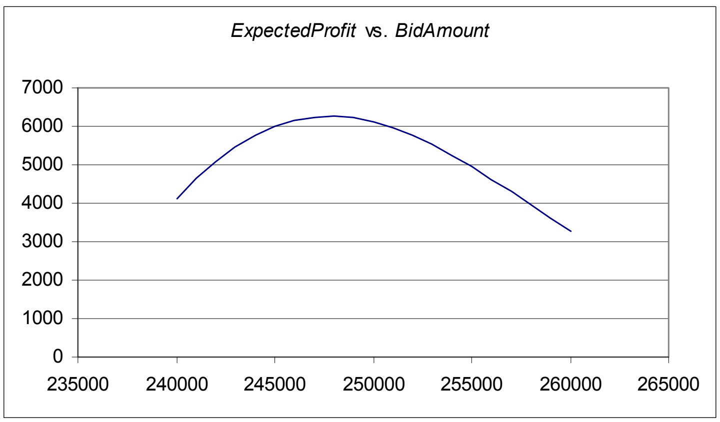
(f) There is uncertainty about the profit that Orion will obtain because there is uncertainty about whether the company will win the contract or not. The probability distribution representing the uncertainty is

Profit	Probability
\$0	0.2854
\$5,771	0.7146

This distribution has a mean of

$$\text{Expected Profit} = E(Profit) = \$0 * (0.2854) + \$5771 * (0.7146) = \$4124$$

(g) The plot of Expected Profit versus Bid Amount is



The maximum Expected Profit in the graph occurs at approximately \$248,000. Therefore, Orion should bid \$248,000 per bus.