Department of Statistics

COURSE STATS 330/762

Assignment 2, 2010

Instructions: Hand in your completed assignment to the Student Resource Centre by **4pm August 19th**

The data set for this assignment is in the file **njgolf.txt** which is available on the course web page.

In a 1996 study, researchers studied the impact on house prices of having a golf course nearby. The data set consists of measurements on 101 houses and condominiums sold between March 1992 and September 1994 in the township of Mt Laurel, New Jersey. As well as data on the individual houses, the proximity of each house to the nearby golf course at the Ramblewood Country Club was also measured. The data set contains the following variables:

sprice: The sale price of the house in $US,
age: age of the house in years,
stories: the number of stories in the house,
firepl: number of fireplaces,
garage: number of garages,
beds: number of bedrooms,
drarea: the area of the dining area,
kitarea: the area of the kitchen,
golf: 1 if the house shares a common boundary with the golf course, otherwise 0,
unrate: the US unemployment rate as the time of sale,
gate: the distance to the gate of the country club (units unknown),
cond: 1 if a condominium, 0 if a house.

The price of a house depends on several factors, such as the size and age of the house, and the prevailing economic conditions. (The variable **unrate** is a proxy for the economic conditions). The idea here is to investigate whether, if these things are held constant, the proximity to the golf course has an extra effect on prices. This is measured by two variables: the **golf** variable is designed to measure the benefit or otherwise of having a golf course next door. The variable **gate** is designed to measure the distance to the golf course entrance.

1. Read the data into R and do the usual checks. Print out the first 10 lines. Note that some of the data are recorded as NA. This the R code for a missing value. [*3 marks*]

2. Fit a regression model to the data using price as the response. Make a comment on how well the regression model fits. Does a residual plot reveal anything wrong? What are the main factors that affect the price of a house? Do you think any variables could be dropped?
   [*15 marks*]

3. How do you think the missing values are treated when fitting the regression model? Hint: look at the degrees of freedom. [*2 marks*]

4. Calculate a confidence interval for the coefficients of golf and gate. Give a careful interpretation of these intervals. Write a concise paragraph describing in lay terms what effect a golf course will have on the price of a house. [*10 marks*]

5. I have deleted the data on one house from the data set. The values for the explanatory variables for this house are

age	stories	firepl	garage	beds	drarea	kitarea	golf	unrate	gate	cond
26	2	1	1	4	156	169	1	7.5	70	0

   Using your model, predict the price of this house using a prediction interval. [*5 marks*]

6. In a plot of **sprice** versus **gate** there is a hint that the relationship might be curved. Is the model improved by fitting a quadratic in gate? [10 marks]

Extra question for 762 students

In lecture 11 we discussed studentised residuals. We are often led to consider the largest studentised residual. How big should we expect this to be if the model is in fact OK and there are no outliers?

A suitable way to answer this is to calculate the 95% point (and possibly 97.5%, 99% as well) of the distribution of the biggest residual in a sample of n, for a range of values of n.

Hint: the largest standardized residual will be approximately like the largest observation in a standard normal sample. Simulate say 10,000 standard normal samples of size n, and for each sample record the maximum. Calculate the 95% quantile of the 10,000 maxima. Repeat for different values of n and different quantiles.

Useful functions: rnorm, max, quantile.

PTO for the golf data

Golf data

```
sprice age stories firepl garage beds drarea kitarea golf unrate gate cond
145000 30 2 1 1 4 132 144 0 6.8 94 0
175000 18 2 1 1 5 180 180 0 6.8 NA 0
68000 7 1 1 0 2 72 88 0 7 16 1
142000 28 2 2 1 3 108 126 1 6.8 63 0
144750 25 2 1 1 4 143 234 0 6.7 167 0
90200 3 1 1 1 2 72 72 0 7 20 1
154000 30 2 1 1 4 156 210 0 6.4 NA 0
150000 30 2 1 1 4 143 156 0 7 103 0
96700 3 3 1 0 1 80 99 1 7.3 15 1
137900 3 2 1 1 2 88 88 1 7.2 22 1
57900 11 1 1 0 1 72 96 0 6.7 53 1
195000 29 2 1 2 4 121 176 1 7.2 NA 0
157000 30 2 1 2 4 132 132 0 7.5 99 0
172000 23 2 1 2 4 132 204 1 7.5 51 0
141000 25 2 1 2 4 156 143 0 7 120 0
164500 21 2 1 2 4 144 117 0 7.5 162 0
90000 3 1 1 1 2 80 72 0 7 20 1
195000 19 2 1 2 4 144 204 1 7.3 58 0
84500 6 1 1 0 3 72 90 1 7.3 12 1
145000 21 1 1 2 3 121 143 0 6.8 116 0
180000 23 2 1 2 5 156 234 1 7 67 0
195000 3 2 1 2 4 154 NA 0 7.3 113 0
156300 28 2 1 1 4 132 234 0 7.3 168 0
135500 11 2 1 1 2 NA 120 1 6.7 99 0
175000 28 2 1 2 4 144 216 0 6.7 41 0
83000 5 1 1 0 2 72 88 0 6.7 21 1
68000 11 1 0 0 2 NA NA 0 5.3 NA 1
101000 15 2 1 0 3 120 120 0 7.2 NA 0
140000 25 2 1 2 4 144 120 0 7 168 0
167500 25 2 1 2 4 132 192 1 7.2 40 0
60000 8 1 1 0 1 72 88 0 6.7 51 1
67500 9 1 1 0 1 72 NA 0 7.3 51 1
160000 30 3 1 1 3 180 144 0 6.7 91 0
182000 23 2 1 2 4 144 216 0 6.7 88 0
91000 4 1 1 1 2 81 81 0 7.3 30 1
135000 20 2 1 1 3 130 182 0 6.7 86 0
114000 10 2 1 1 3 132 64 0 6.3 93 0
94000 3 1 1 1 2 72 72 1 6 22 1
166000 26 2 1 1 4 156 154 0 7.3 110 0
97000 4 1 1 1 2 154 120 0 7.2 25 1
173500 25 2 1 2 4 144 192 0 7.2 77 0
165000 25 2 1 2 4 144 216 0 7.2 174 0
189000 20 2 1 2 4 144 192 1 7.3 126 0
85000 8 1 1 0 2 72 NA 0 6.5 11 1
143500 22 1 1 2 3 156 240 0 6.3 123 0
140900 4 2 1 1 2 154 143 1 7.2 16 1
163000 21 2 1 2 4 121 198 0 5.3 62 0
```

```
163000 23 2 1 1 4 156 247 1 7.2 71 0
235000 2 2 1 2 4 156 255 0 7.3 NA 0
175500 25 2 1 2 4 132 187 0 7.2 72 0
174500 25 2 0 2 4 132 204 1 6.5 50 0
148000 27 2 0 1 4 132 234 0 6.7 173 0
179900 22 2 1 2 4 156 216 0 6.7 100 0
73000 12 1 1 0 2 81 81 0 5.3 NA 1
211000 5 2 1 2 4 156 286 0 5.3 28 0
63000 11 1 1 2 2 90 110 1 5.9 55 1
114900 5 1 1 1 2 143 120 0 6.3 NA 1
140000 29 2 0 2 5 120 90 1 6.4 132 0
179900 NA 2 1 2 4 132 192 0 7.2 50 0
132000 25 2 0 1 4 110 160 0 6.4 171 0
135000 25 1 1 1 3 110 165 0 6.5 38 0
162000 25 2 1 2 4 121 176 0 6.3 68 0
165000 21 2 1 2 5 154 162 0 6.3 114 0
150000 28 2 1 2 4 144 144 0 5.3 168 0
193000 23 2 1 2 4 132 204 0 5.8 94 0
146000 5 2 1 1 3 143 120 0 6.2 31 1
97900 5 1 1 2 1 110 132 1 5.8 14 1
98500 4 1 1 1 2 90 72 1 6.2 24 1
139000 4 2 1 1 3 108 108 1 6.2 28 1
147000 28 3 2 0 4 120 NA 0 6.4 77 0
84000 8 1 1 0 2 72 NA 1 6.2 7 1
80000 6 1 1 0 2 72 96 0 6.2 23 1
137000 24 1 1 2 3 132 80 0 6 141 0
95000 NA 3 1 2 2 NA NA 0 5.9 14 1
144000 29 2 1 1 4 132 221 0 5.9 160 0
79000 8 1 1 2 2 72 NA 0 6.1 14 1
137500 6 2 1 1 3 132 143 0 6.5 20 1
73000 11 1 1 0 2 64 90 0 6.4 NA 1
140000 26 2 1 2 4 144 144 0 6.2 182 0
96000 4 1 1 1 2 90 90 0 6.2 16 1
87000 10 1 1 0 3 72 99 1 6.4 8 1
195000 23 2 1 2 5 132 117 0 6.5 92 0
158000 NA 2 1 1 4 132 234 0 6.2 153 0
183000 30 2 2 2 4 156 224 0 6.2 112 0
183750 24 2 1 2 4 121 176 1 6.3 75 0
166000 29 2.5 2 1 4 156 156 1 5.8 74 0
158000 30 2.5 1 1 3 130 180 1 5.8 80 0
65500 8 1 1 0 2 72 96 0 6.5 60 1
175000 25 2 1 2 4 132 204 0 5.8 63 0
184900 NA 2 1 2 5 121 176 0 6.5 74 0
87500 9 1 1 2 3 72 NA 0 6.5 15 1
152500 24 2 1 1 5 154 154 0 6.2 76 0
177000 25 2 1 1 5 192 162 0 5.8 102 0
68500 11 1 0 0 2 NA 77 0 6.1 58 1
148000 26 2 1 1 5 143 168 0 6.5 145 0
98500 5 1 1 0 1 110 132 1 6.5 30 1
154000 28 2 1 2 4 156 143 0 6.5 124 0
69000 20 1 1 0 2 120 90 0 6.5 NA 1
122500 7 3 1 0 2 182 210 0 5.8 16 1
198000 25 2 1 2 4 143 216 0 6.3 88 0
97500 6 1 1 1 2 90 72 0 NA 25 1
```