Department of Statistics

COURSE STATS 330/762

Model answers for Assignment 2, 2010

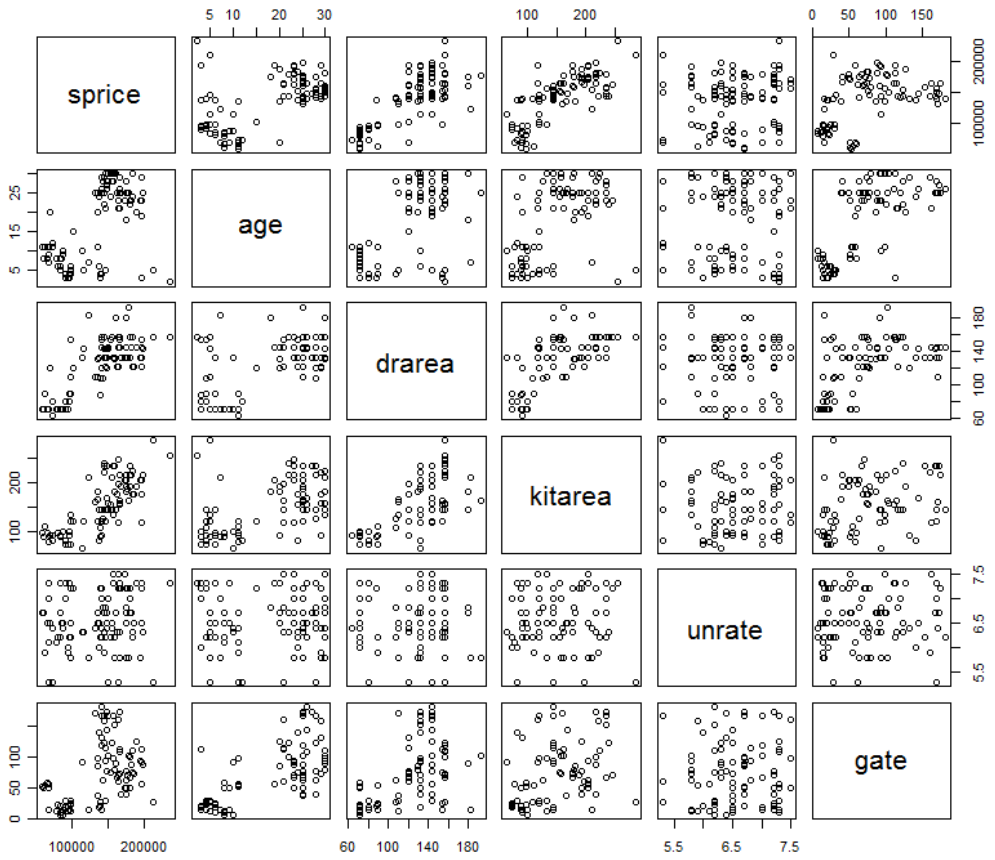
1. *Read the data into R and do the usual checks. Print out the first 10 lines. Note that some of the data are recorded as NA. This the R code for a missing value. [3 marks]*

   The data can be read in using the code

```
> njgolf.df=read.table(file.choose(), header=TRUE )
```

   A check of the data using a pairs plot (not shown) revealed no obvious outliers. We only looked at the continuous variables:

```
pairs(njgolf.df[,c(1,2,7,8,10,11)])
```



The pairs plot shows no obvious outliers. There are several variables (age, drarea, kitarea, gate) that have a relationship with sale price. The relationship with gate seems curved.

The first 10 lines are

```
> njgolf.df[1:10,]
   sprice age stories firepl garage beds drarea kitarea golf unrate gate cond
1  145000  30       2      1      1    4    132     144    0    6.8   94    0
2  175000  18       2      1      1    5    180     180    0    6.8   NA    0
3   68000   7       1      1      0    2     72      88    0    7.0   16    1
4  142000  28       2      2      1    3    108     126    1    6.8   63    0
5  144750  25       2      1      1    4    143     234    0    6.7  167    0
6   90200   3       1      1      1    2     72      72    0    7.0   20    1
7  154000  30       2      1      1    4    156     210    0    6.4   NA    0
8  150000  30       2      1      1    4    143     156    0    7.0  103    0
9   96700   3       3      1      0    1     80      99    1    7.3   15    1
10 137900   3       2      1      1    2     88      88    1    7.2   22    1
```

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

We first fit a model using all the variables, and **sprice** as the response:

> njgolf.lm=lm(sprice~., data=njgolf.df))

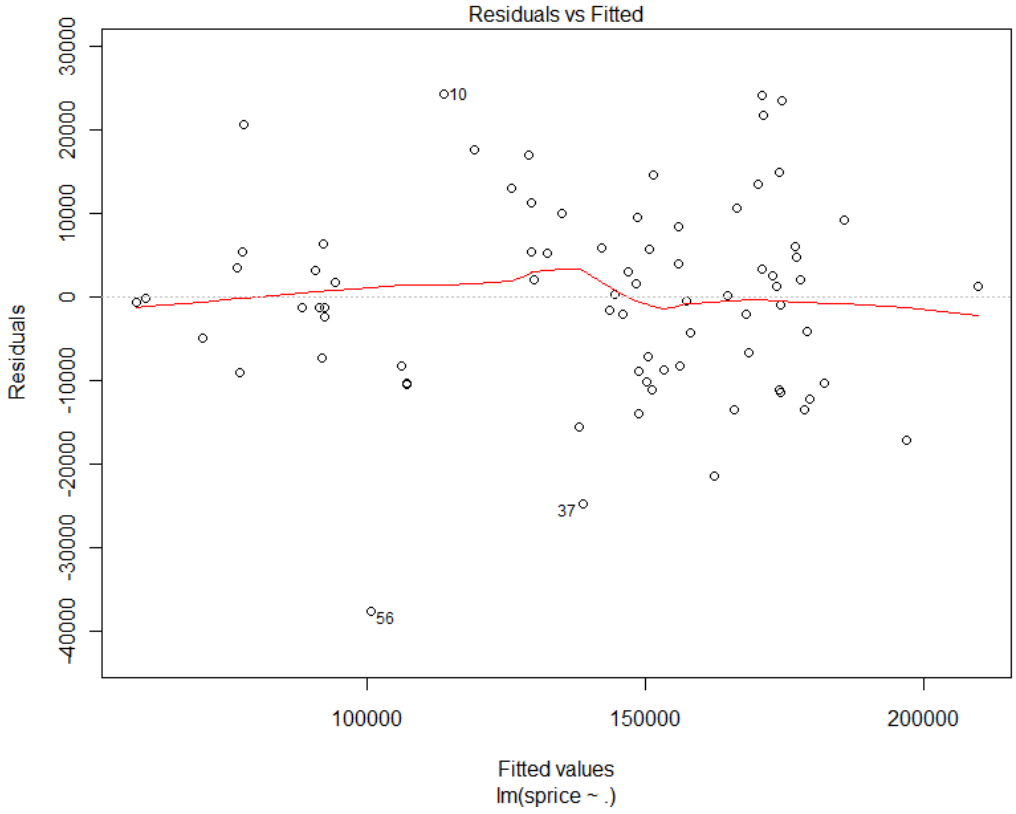
Note the use of . This means that all the variables (other than sprice) are included as explanatory variables.

The summary is

```
> summary(njgolf.lm)
Call:
lm(formula = sprice ~ ., data = njgolf.df)

Residuals:
     Min       1Q   Median       3Q      Max
-37667.3  -8273.4    145.3   5883.6  24159.4

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  36006.27   25037.76   1.438 0.155209
age           -872.58     389.81  -2.238 0.028622 *
stories      16151.02    3924.05   4.116 0.000111 ***
firepl        7510.37    5046.04   1.488 0.141490
garage       13772.66    2904.54   4.742 1.20e-05 ***
beds         10705.42    2589.73   4.134 0.000104 ***
drarea         120.23      83.60   1.438 0.155159
```

```
kitarea        146.74      38.88   3.774 0.000350 ***
golf          1651.95    3537.23   0.467 0.642050
unrate        2088.15    2614.39   0.799 0.427363
gate          -135.41      46.98  -2.882 0.005348 **
cond        -31845.47   10075.10  -3.161 0.002390 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12440 on 65 degrees of freedom
  (24 observations deleted due to missingness)
Multiple R-squared: 0.9073,     Adjusted R-squared: 0.8916
F-statistic: 57.82 on 11 and 65 DF,  p-value: < 2.2e-16
```

The model seems to fit well, with an R2 of over 90%. The residual plot below shows no major problems.



Point 56 is not too extreme. The variables contributing to the explanation of price seem to be **age, stories, garage, beds, kitarea, gate** and con**d**. Most of these are related to size of the house, but **gate** is also seems to have an influence on price – the closer the golf course, the higher the price. Being right next door seems not to matter, but this effect could be confounded with gate. The variables **firepl, drarea, golf, unrate** are candidates

for dropping, due to their small p-values, but dropping them all might not be a good idea due to possible multicollinearity.

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

According to the output, 24 observations have been deleted. Looking at the data, we see that 24 houses have at least one NA. Thus, the lm function includes only houses with complete data. The number of observations used in the regression is the degrees of freedom + the number of independent variables+1, or 65+12=77. The number of houses in the data set is 101, so 101-77=24 have been deleted.

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]

```
> confint(njgolf.lm, c("golf","gate"))
          2.5 %     97.5 %
golf -5412.3994 8716.29140
gate  -229.2423  -41.57924
```

Thus, having a golf course next door has an effect on the price between -$5412 and $8716. This variable does not have a significant effect on price. The effect of the distance to the gate is to drop the price by between $229 and $41 for each extra unit of distance.

In practical terms, distance does matter, the further from the course the lower the price. The effect of being right next door is inconclusive, but this effect may be already explained by the gate variable, so including golf contributes nothing further.

5. I have deleted the data on one house from the original 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]

First make a new data frame for the new data:

```
 > new.house.df = data.frame(age=26, stories=2, firepl=1, garage=1, beds=4,
drarea=156,
+ kitarea=169, golf=1,  unrate=7.5, gate=70,  cond=0)
> new.house.df
  age stories firepl garage beds drarea kitarea golf unrate gate cond
1  26       2      1      1    4    156     169    1    7.5   70    0
```

Then do the prediction:

```
> predict(njgolf.lm, new.house.df, interval="p")
       fit      lwr      upr
1 161116.4 134351.1 187881.8
```

The point prediction is about $161,000 and the prediction interval is between $134,000 and $187,000. In actual fact the value was quite close so the prediction is pretty good!

6. *In the pairs plot there was a hint that the relationship with gate might be curved. Is the model improved by fitting a quadratic in gate? [10 marks]*

There is a complication in using poly to fit the quadratic – poly can't handle missing values. The solution is to make a new data frame that has all the rows with missing values deleted. You can do this using an editor or Excel, but here is a trick to do it in R:

```
# make a function that returns TRUE if no elements of x are NA
no.na = function(x)(all(!is.na(x)))

# make a vector having value TRUE if no variable in the corresponding row of
# njgolf.df has a missing value
use = apply(njgolf.df,1,no.na)

njgolf.no.na.df = njgolf.df[use,]
```

Now we can fit the quadratic. Here is a short way to specify the model: we want to include all variables (use .), exclude gate (-gate), and include the polynomial in gate (+poly(gate,2))

```
quad.golf.lm = lm(sprice~ .  -gate + poly(gate,2), data=njgolf.no.na.df)
summary(quad.golf.lm)
Call:
lm(formula = sprice ~ . - gate + poly(gate, 2), data = njgolf.no.na.df)

Residuals:
      Min        1Q    Median        3Q       Max
-38688.37  -8338.43     96.92   6176.42  24093.95

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     25237.83   24932.22   1.012 0.315227
age              -901.30     402.29  -2.240 0.028543 *
stories         16288.58    3973.79   4.099 0.000119 ***
firepl           7316.74    5115.71   1.430 0.157510
garage          13747.59    2925.73   4.699 1.43e-05 ***
beds            10787.49    2619.86   4.118 0.000112 ***
drarea            115.69      85.32   1.356 0.179897
kitarea           150.24      40.60   3.701 0.000449 ***
golf             1751.10    3574.78   0.490 0.625916
unrate           2101.84    2632.88   0.798 0.427646
cond           -30646.26   10791.72  -2.840 0.006045 **
poly(gate, 2)1 -54911.72   25442.68  -2.158 0.034666 *
poly(gate, 2)2  -5908.24   18127.95  -0.326 0.745549
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12520 on 64 degrees of freedom
Multiple R-squared: 0.9074,     Adjusted R-squared: 0.8901
```

```
F-statistic: 52.28 on 12 and 64 DF,  p-value: < 2.2e-16
```

Since the p-value for poly(gate,2) is large, we conclude that a quadratic is not justified.

Extra question for 762 students [10 marks]

In lecture 11 we discussed studentised and standardized residuals. We are often led to consider the largest 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 normal sample. Simulate say 10000 normal samples of size n, and for each sample record the maximum. Calculate the 95% quantile of the 10000 maxima. Repeat for different values of n.

Useful functions: rnorm, max, quantile.

Here is my version. The code prints out a table of quantiles for n=5,10,15,…,100.

```
# n.samp is sample size
# N is number of simulations
N = 10000
n.samp= seq(5, 100, by=5)
quantile.table = matrix(0,20,3)  # reserve space for results
for(i in 1:20){
  xmax = numeric(N)
  for( j in 1:N)xmax[j] = max(abs(rnorm(n.samp[i])))
  quantile.table[i,] = quantile(xmax, c(0.95, 0.975,0.99))
}
dimnames(quantile.table) = list(paste("n=",n.samp,sep=""),
c("95%", "97.5%","99%"))
round(quantile.table,4)
```

```
         95%  97.5%    99%
n=5   2.5551 2.7979 3.1036
n=10  2.7935 2.9927 3.2458
n=15  2.9484 3.1675 3.4200
n=20  3.0155 3.2301 3.4775
n=25  3.0892 3.3028 3.5586
n=30  3.1168 3.3013 3.5254
n=35  3.1727 3.3727 3.6245
n=40  3.2320 3.4293 3.6808
n=45  3.2594 3.4631 3.7144
n=50  3.2777 3.4638 3.6661
n=55  3.3011 3.4782 3.7074
n=60  3.3373 3.5231 3.7754
```

```
n=65  3.3471 3.5384 3.7116
n=70  3.3749 3.5877 3.8288
n=75  3.4076 3.5889 3.8198
n=80  3.4243 3.6249 3.8404
n=85  3.4367 3.6130 3.8465
n=90  3.4351 3.6144 3.8695
n=95  3.4479 3.6357 3.8755
n=100 3.4698 3.6641 3.8864
```

Thus, for example, when n=100, the maximum residual will have an absolute value bigger than 3.47 about 5% of the time., and bigger than 3.89 about 1% of the time.