Department of Statistics

COURSE STATS 330

Model answers for Assignment 2, 2005

The file `houseprices.csv` contains data on 199 houses sold in Auckland in April and May of 2005. The variables in the data set are

`SalePrice:`	Sale price in $$,
`DaysOnMarket:`	Number of days between listing and sale,
`Bedrooms:`	Number of bedrooms,
`Valuation:`	Valuation New Zealand valuation in $$,
`YearValued:`	Year valuation was done, either 2002 or unknown.

1. *Load the data into R and create a data frame* ***`houseprices.df`***.

   Use the code

   ```
   houseprices.df<-read.csv(file.choose(), header=T)
   ```

2. *Using the 2002 valuation data only, fit a regression model to the data using SalePrice as the response. Are there any problems with the model? If so, fix them. Can some variables be eliminated from the model? If so, eliminate them.*

First let's look at a pairs plot of the data, shown below.

Use the code

```
use.data<-(1:199)[housing.df$YearValued=="2002"]
pairs(houseprices.df[use.data,1:4])
```

Seems that there is not much relationship betweem price and days on market, a moderate relationship between price and bedrooms, and a very strong relationship between price and valuation. There also seems to be an outlier. To identify this, we will fit the model, and look at the outlier diagnostics.

```
housing.lm<-lm(SalePrice~DaysOnMarket + Bedrooms + Valuation,
subset=use.data, data=housing.df)
summary(housing.lm)
```

The result is

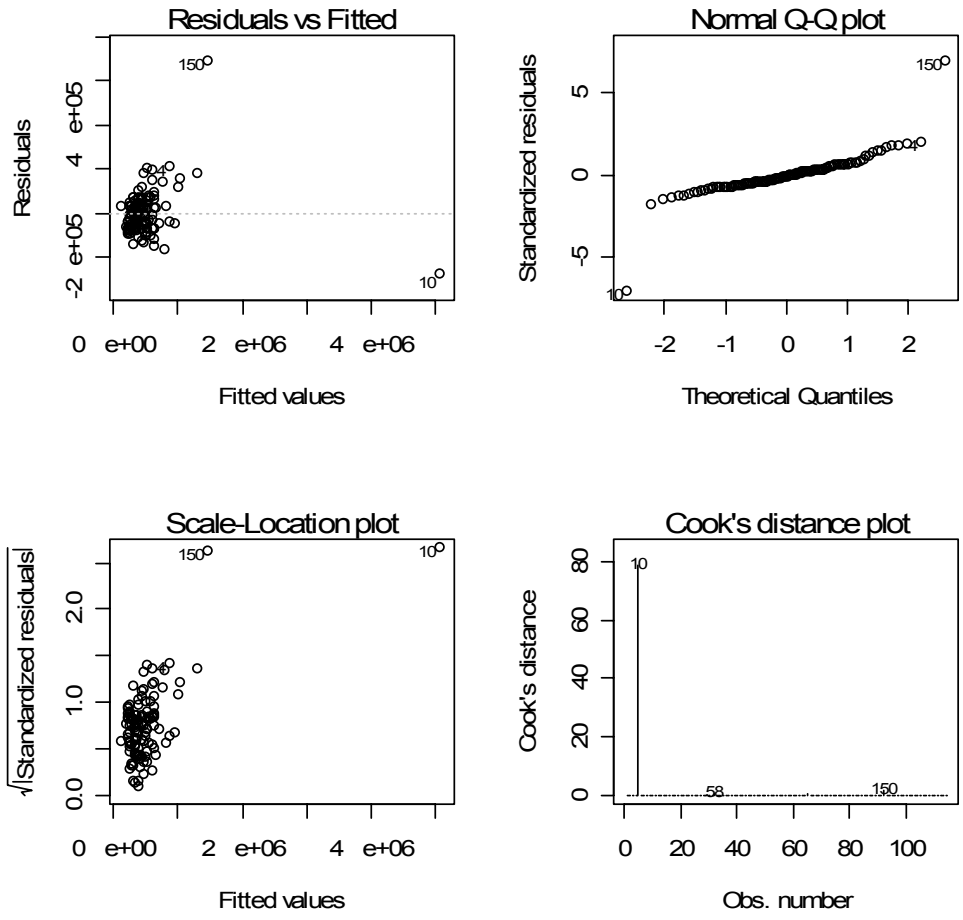
```
Call:
lm(formula = SalePrice ~ DaysOnMarket + Bedrooms + Valuation,
    data = housing.df, subset = use.data)
Residuals:
    Min      1Q  Median      3Q     Max
-274540  -61068   -9485   44257  686087

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -9.677e+03  4.077e+04  -0.237  0.81280
DaysOnMarket -4.384e+02  3.332e+02  -1.316  0.19097
Bedrooms      3.893e+04  1.199e+04   3.248  0.00154 **
Valuation     1.161e+00  2.475e-02  46.902  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 106000 on 110 degrees of freedom
Multiple R-Squared: 0.9552,     Adjusted R-squared: 0.954
F-statistic: 781.8 on 3 and 110 DF,  p-value: < 2.2e-16
```

Seems like DaysOn Market is not required. To identify the outlier, we draw a residual plot.

```
plot(housing.lm)
```



Clearly points 150 and 10 are outliers.We delete these. Note that the labels on the plot are the rowlabels:

```
row.names(housing.df[use.data,])
```

```
row.names(housing.df[use.data,])
  [1] "1"   "4"   "5"   "7"   "10"  "11"  "12"  "13"  "15"  "17"  "18"  "19"
 [13] "20"  "21"  "22"  "23"  "28"  "29"  "30"  "35"  "37"  "39"  "43"  "45"
 [25] "47"  "49"  "50"  "51"  "53"  "54"  "57"  "58"  "61"  "62"  "63"  "64"
 [37] "65"  "67"  "68"  "69"  "71"  "72"  "73"  "74"  "75"  "76"  "78"  "80"
 [49] "83"  "84"  "85"  "87"  "93"  "94"  "95"  "96"  "97"  "98"  "99"  "100"
 [61] "101" "102" "103" "104" "105" "107" "108" "110" "113" "115" "116" "118"
 [73] "119" "120" "121" "123" "124" "126" "128" "130" "132" "134" "136" "137"
 [85] "139" "140" "141" "142" "143" "146" "147" "150" "152" "153" "155" "157"
 [97] "161" "163" "166" "167" "168" "169" "170" "173" "175" "180" "185" "186"
[109] "191" "192" "193" "194" "196" "197"
```

So in fact we have to delete the 5th and 92nd observations.

```
newuse.data<-use.data[-c(5,92)]
```

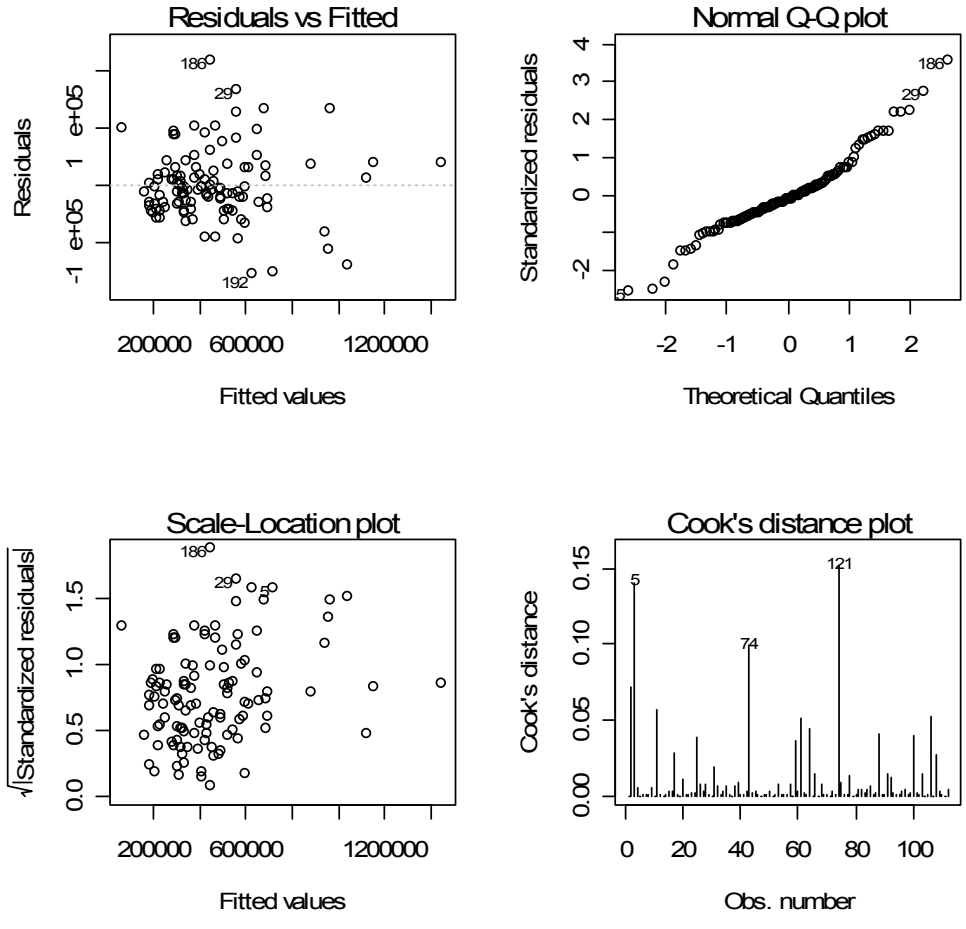
```
newhousing.lm<-lm(SalePrice~DaysOnMarket + Bedrooms + Valuation,
subset=newuse.data, data=housing.df)
```

Now do a summary: seems like bedrooms is not required in the model:

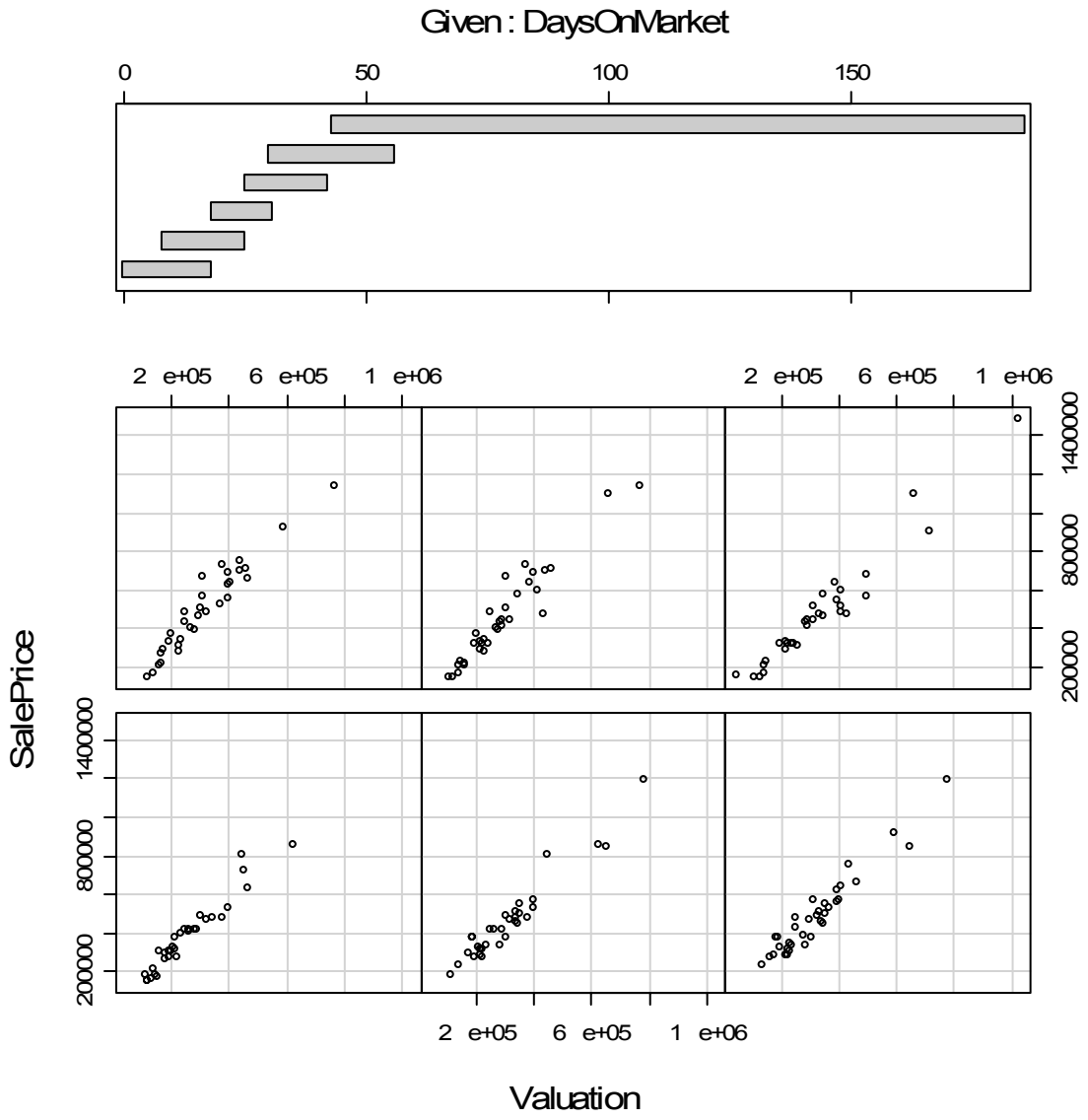
```
Call:
lm(formula = SalePrice ~ DaysOnMarket + Bedrooms + Valuation,
    data = housing.df, subset = newuse.data)

Residuals:
    Min      1Q  Median      3Q     Max
-153044  -36333   -7132   29873  217495

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.127e+04  2.458e+04   2.493   0.0142 *
DaysOnMarket -4.288e+02  1.981e+02  -2.165   0.0326 *
Bedrooms     -1.216e+04  7.887e+03  -1.541   0.1262
Valuation     1.474e+00  4.141e-02  35.597   <2e-16 ***
```



The model now is not bad, but there is a hint of a funnel effect. This shows up well in a coplot



A log transformation will cure the funnel effect, but will destroy the planar relationship. The solution is to log valuation as well. This cures the problem, but shows a new outlier, point 25 (row label 49). Removing this results in a good model. With this model, both Days on Market and bedrooms are not required. The fitted model is

```
Log(Price) = 0.40939 + 0.99878 log(Valuation)

                Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.40939    0.34736   1.179    0.241
log(Valuation)   0.99878    0.02772  36.035   <2e-16 ***
```

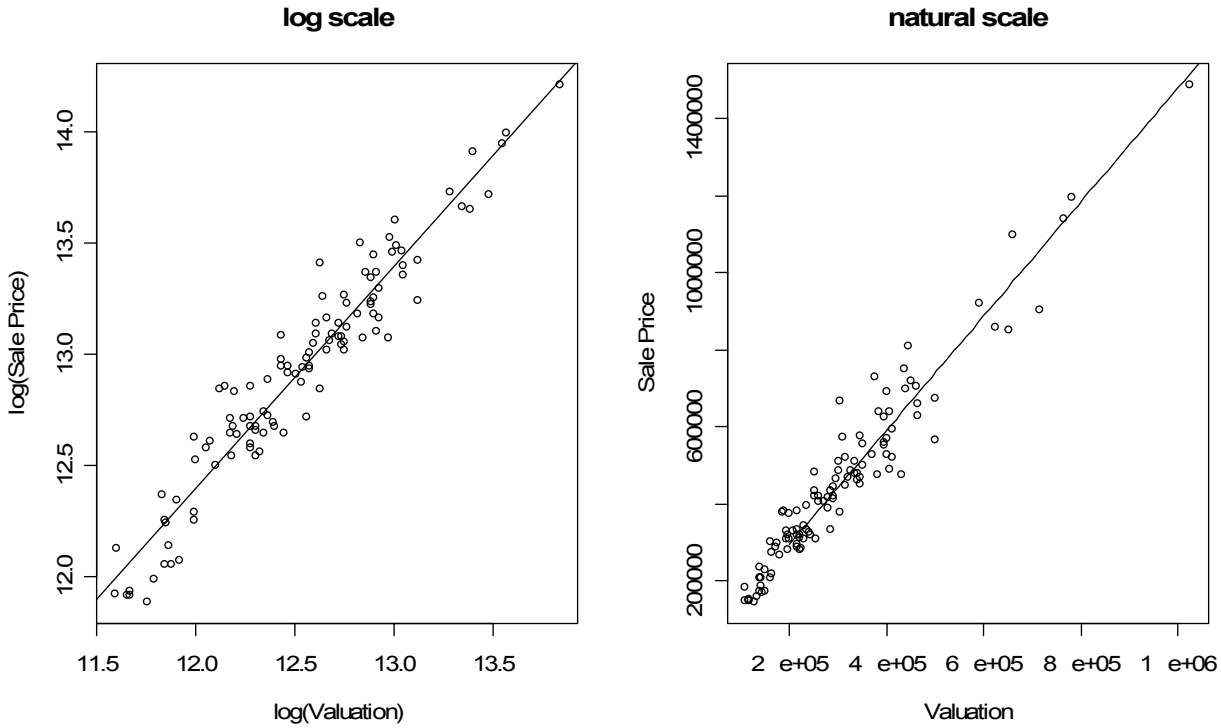
3. What is the relationship between the valuation and the sale price? Does it depend on the other variables? Interpret your model to answer this part, and also draw a suitable graph.

The relationship is

Log(Price) = 0.40939 + 0.99878 log(Valuation)

Or Price = $e^{0.40939}$ Valuation$^{0.99878}$

This is shown on the log-log scale and the original scale below. The relationship is very close to Price = 1.505899 x Valuation.



4. Suppose you live in a house with a 2002 valuation of $350,000 and 3 bedrooms. What range of prices should you expect to sell for?

```
> predict.int<-predict(model4.lm, data.frame(Valuation=350000),
interval="p")
> predict.int
          fit      lwr      upr
[1,] 13.15951 12.88520 13.43382
> exp(predict.int)
          fit      lwr      upr
[1,] 518924.1 394431.7 682709.2
```

Note that the predict function returns a vector whose elements are the prediction (on the log scale) and the end points of the interval. To get an interval on the natural scale we transform using the exp function. The price is expected to be between $394,000 and $683,000.

5. *Now repeat the analysis for the unknown year of valuation data. Would you expect to get the same result? In what way would you expect the model to change? Has it in fact done so?*

On the next page we have plotted the data on a log – log scale, with the 2002 valuation data in red, and the unknown valuation date data in blue. We can see that the relationship for the latter data is not as strong. We can imagine that the "unknown" data is an approximation to the 2002 valuation. There are several outliers. Deleting the worst 5, and refitting lines separately, we get the following:

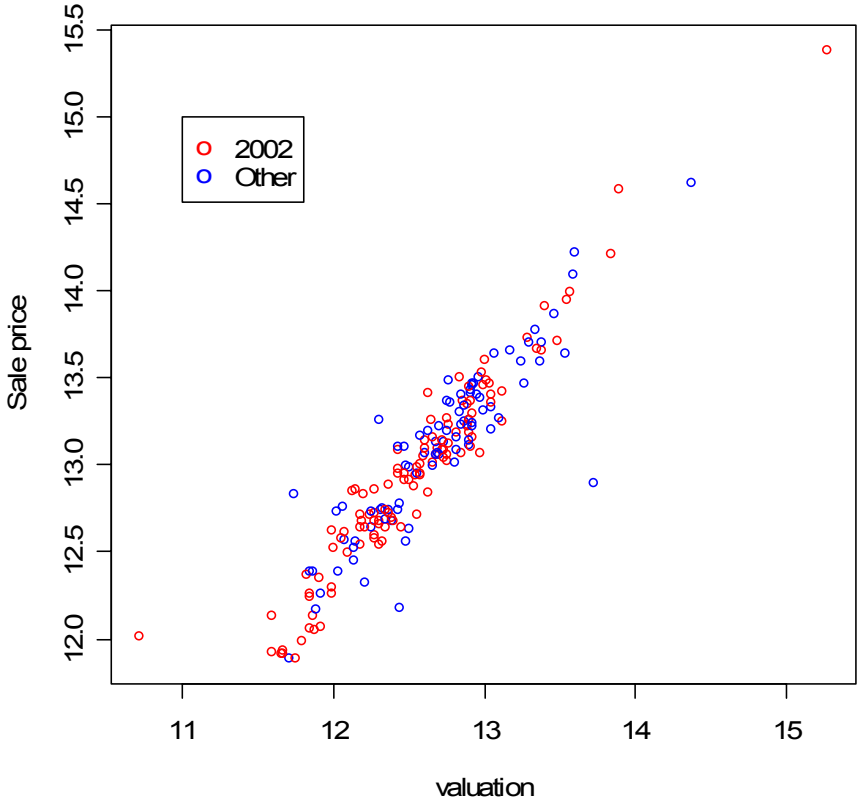
2002 data

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.40939    0.34736   1.179    0.241
log(Valuation)  0.99878    0.02772  36.035   <2e-16 ***
```

Other data

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.51829    0.55301   2.746  0.00744 **
log(Valuation)  0.91238    0.04356  20.943  < 2e-16 ***
```

Standard errors for the "other data" regression are almost twice those of the 2002 regression, reflecting the approximate nature of the valuation data. The other regression has a slightly lower slope, possibly due to the other valuations being later on average than 2002, and hence higher. There is also a general tendency for regressions when the x-values are only approximate to have lower slopes.


2002
Other
Sale price
valuation
12.0
12.5
13.0
13.5
14.0
14.5
15.0
15.5
11
12
13
14
15