Business Statistics - Spring 2013 **Homework V Solution**

Regression analysis of used Mercedes cars: Data taken from the advertising pages of the Sunday Times a few years ago, presenting cars for sale in the UK (mainly in and around London). The asking prices (in pounds sterling) are classified according to type/model of car, age of car (in six-month units based on date of registration), recorded mileage, and vendor. The data in `usedcars.xls` are: 1. Case number 1:54; 2. `price`: Asking price in pounds; 3. `type`: Type/Model: 0=model 500, 1=450, 2=380, 3=280, 4=200; 4. `age`: Age of car in six-month units, based on registration; 5. `mileage`: Recorded mileage (in thousands); and 6. `vendor`: Vendor (0,1,2,3 are dealerships, 4="sale by owner").

a) Run the simple linear regression of `price` on `mileage`.

```
#  Coefficients     Estimate Std. Error t value Pr(>|t|)
#  (Intercept)       19302.0     1235.3  15.625  < 2e-16
#  mileage            -209.4       52.8  -3.966 0.000225
#
# Residual standard error: 4554 on 52 degrees of freedom
# Adjusted R-squared: 0.2175
```

b) Run the simple linear regression of `price` on `age`.

```
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)  19409.5     1421.6  13.653  < 2e-16
# age          -1128.2      329.7  -3.422  0.00122
#
# Residual standard error: 4695 on 52 degrees of freedom
# Adjusted R-squared: 0.1681
```

c) Which of the linear regressions fits `price` better in terms of R^2?

Regressing price on mileage produces a larger R^2.

d) Run the multiple linear regression of `price` on `mileage` and `age`.

```
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept) 20142.37    1413.22  14.253   <2e-16
# mileage      -153.23      70.18  -2.183   0.0336
# age          -513.21     425.08  -1.207   0.2329
#
# Residual standard error: 4534 on 51 degrees of freedom
# Adjusted R-squared: 0.2243
```

e) Based on R^2, is the multiple linear regression in d) better than the ones in a) and b)?

Regressing price on mileage and age is slightly better than regressing it on mileage alone.

f) Do the residuals of the previous three linear regressions look i.i.d. normal?

Visually they look iid normal. See Figure 1.

The variables `type` and `vendor` are both categorical and need special consideration. Remember, from class, that in order to run regressions with these variables we need first to create dummy variables (0/1 variables) to account for the different categories. For example, `type` has 5 categories, so 4 dummy variables are necessary.

g) Run the regression of `price` on `type`.

Since type is in $\{0, 1, 2, 3, 4\}$ (5 categories), we need 4 dummy variables. Let M0=1 when type=0 and M0=0 otherwise, M1=1 when type=1 and M1=0 otherwise, the same for M2 and M3. In this case, the intercept of the regression corresponds to type=4.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    9235.6      617.6  14.953  < 2e-16
M0            12843.1     1069.8  12.005 3.33e-16
M1             5610.4     1265.8   4.432 5.25e-05
M2             9921.9      995.9   9.963 2.28e-13
M3             5647.7      887.9   6.361 6.49e-08

Residual standard error: 2471 on 49 degrees of freedom
Adjusted R-squared: 0.7697
```

The average price of type=4 car is 9235.6, while the average price of type=0 car is (9235.6+12843.1)=22078.7. Similarly for type=1,2,3.

h) Run the regression of `price` on `vendor`.

Since vendor is in $\{0, 1, 2, 3, 4\}$ (5 categories), we need 4 dummy variables. Let V0=1 when vendor=0 and V0=0 otherwise, V1=1 when vendor=1 and V1=0 otherwise, the same for V2 and V3. In this case, the intercept of the regression corresponds to vendor=4 (or "sale by owner").

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   13503.1     1369.7   9.859 3.22e-13
V0             3015.1     2023.1   1.490   0.1425
V1             5054.4     2219.1   2.278   0.0271
V2             1925.3     2141.4   0.899   0.3730
V3             -510.8     1937.0  -0.264   0.7931

Residual standard error: 4938 on 49 degrees of freedom
Adjusted R-squared: 0.07975
```

i) Run the regression of `price` on `mileage`, `age`, `type` and `vendor`.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13079.94     930.74  14.053  < 2e-16
mileage       -30.02      30.76  -0.976    0.335
age          -916.31     189.58  -4.833 1.74e-05
M0          11822.16     723.35  16.344  < 2e-16
M1           8652.82    1075.65   8.044 4.10e-10
M2           9109.26     670.22  13.592  < 2e-16
M3           4930.18     641.60   7.684 1.33e-09
V0            856.77     706.61   1.213    0.232
V1           1232.46     860.62   1.432    0.159
V2           -334.57     819.06  -0.408    0.685
V3            875.92     643.86   1.360    0.181

Residual standard error: 1608 on 43 degrees of freedom
Adjusted R-squared: 0.9025
```

j) Compare all 6 models based on R^2.

Based on R^2 the complete model is the best model. For illustration let us go one step further (this was not part of the homework). It seems that both mileage and vendor become obsolete variables in this complete model. Running the regression of price on age and type leads to $R^2 = 0.8965$ and $s = 1656$, both of with are very close to their counterparts in the complete model.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   13485.8      683.9  19.720  < 2e-16
age           -1079.4      138.2  -7.810 4.27e-10
M0            11966.1      726.0  16.482  < 2e-16
M1             8916.1      948.4   9.401 1.84e-12
M2             9233.7      673.5  13.709  < 2e-16
M3             5139.5      598.9   8.582 2.95e-11

Residual standard error: 1656 on 48 degrees of freedom
Adjusted R-squared: 0.8965
```

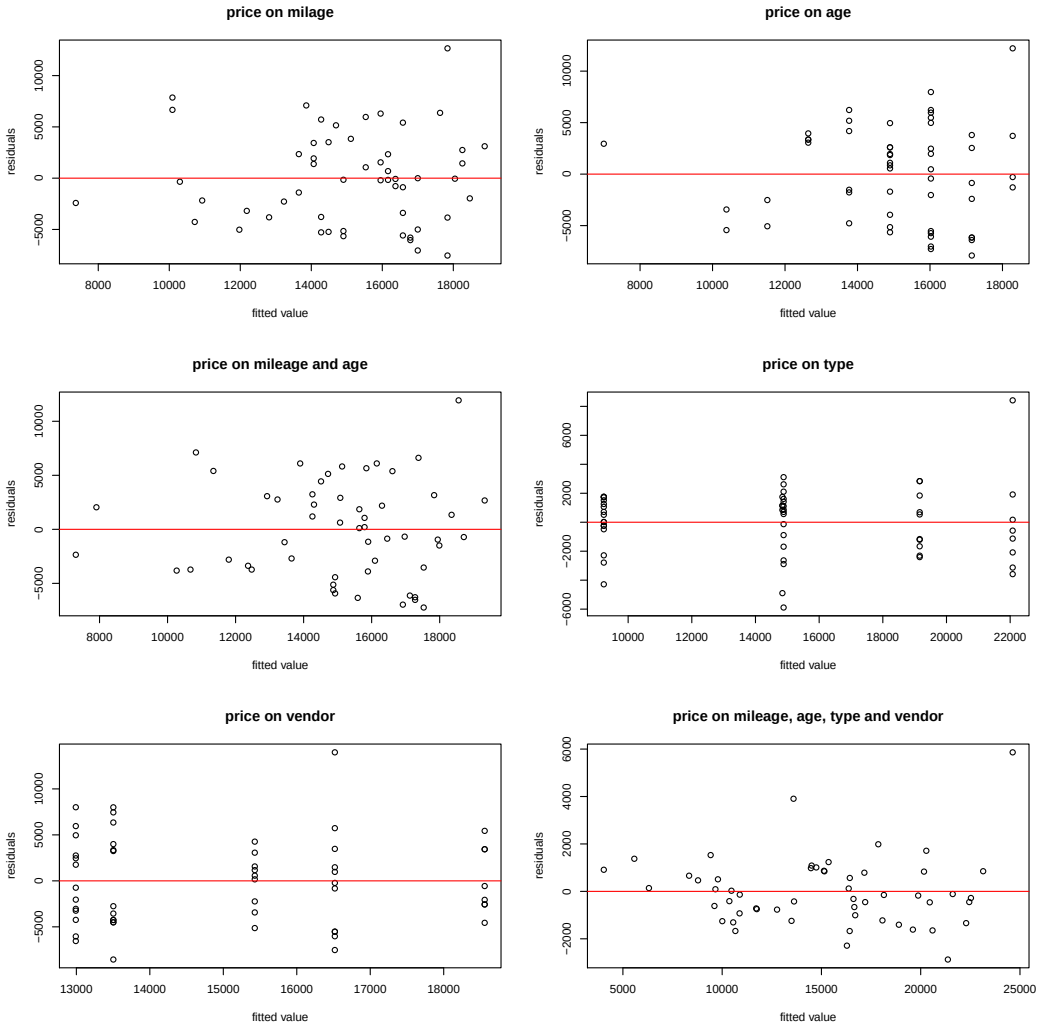

Figure 1: *Residual analysis.*