Practice Test

The test will be open note, open device.

1. This problem uses multiple regression. A data set includes mortality and pollution data for various cities. The variables include (among others). We'll only analyze a subset of the variables.

- MORT (mortality per 100,000 adjusted for age–don't worry about this adjustment, just think of it as your response variable, and higher numbers mean a higher proportion of people die per year).
- JANT average temperature in January
- JULT average temperature in July
- DENS a measure of population density
- HC relative hydrocarbon potential
- NOX measure of nitrous oxide pollution
- SO measure of sulphur dioxide
- POPN average household size

Here is what the data looks like

```
> head(x)
  PREC JANT JULT OVR65 POPN EDUC HOUS DENS NONW WWORK POOR HC NOX  SO HUMID
1   36   27   71   8.1 3.34 11.4 81.5 3243  8.8  42.6 11.7 21  15  59    59
2   35   23   72  11.1 3.14 11.0 78.8 4281  3.5  50.7 14.4  8  10  39    57
3   44   29   74  10.4 3.21  9.8 81.6 4260  0.8  39.4 12.4  6   6  33    54
4   47   45   79   6.5 3.41 11.1 77.5 3125 27.1  50.2 20.6 18   8  24    56
5   43   35   77   7.6 3.44  9.6 84.6 6441 24.4  43.7 14.3 43  38 206    55
6   53   45   80   7.7 3.45 10.2 66.8 3325 38.5  43.1 25.5 30  32  72    54
      MORT
1  921.870
2  997.875
3  962.354
4  982.291
5 1071.289
6 1030.380
```

Suppose the full model uses these variables, and interaction terms won't be considered. Here is the summary of this full model:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 317.234007 196.456766   1.615   0.1124
JANT          0.740021   0.837653   0.883   0.3811
JULT          1.463714   1.762884   0.830   0.4102
DENS          0.008582   0.004999   1.717   0.0920 .
HC           -1.412308   0.585429  -2.412   0.0194 *
NOX           2.673728   1.204168   2.220   0.0308 *
SO            0.136196   0.175738   0.775   0.4419
POPN        135.180426  54.051238   2.501   0.0156 *
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 48.65 on 52 degrees of freedom
Multiple R-squared:  0.4608,Adjusted R-squared:  0.3882
F-statistic: 6.349 on 7 and 52 DF,  p-value: 2.078e-05
```

(a) If you were to do backward elimination using p-values, which variable would you eliminate from the model at this point to fit a slightly smaller model?

(b) If you were to do forward selection using p-values for doing model selection, which variable would you want to include first?

(c). Consider a smaller model

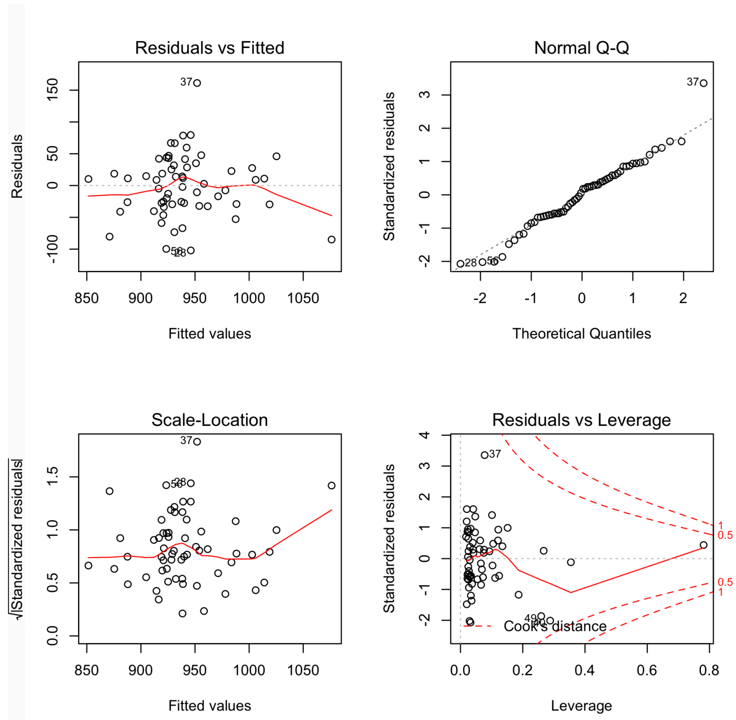
```
> m2 <- lm(MORT ~ HC + SO + NOX+POPN)
> summary(m2)

Call:
lm(formula = MORT ~ HC + SO + NOX + POPN)

Residuals:
     Min       1Q   Median       3Q      Max
-102.045  -29.481    5.781   29.330  161.247

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 533.8097   173.2101   3.082  0.00321 **
HC           -1.4346     0.5912  -2.427  0.01854 *
SO            0.1946     0.1668   1.166  0.24861
NOX           2.7138     1.2268   2.212  0.03114 *
POPN        119.1855    52.8191   2.256  0.02804 *
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 50.05 on 55 degrees of freedom
Multiple R-squared:  0.3966,Adjusted R-squared:  0.3527
F-statistic: 9.037 on 4 and 55 DF,  p-value: 1.103e-05
```

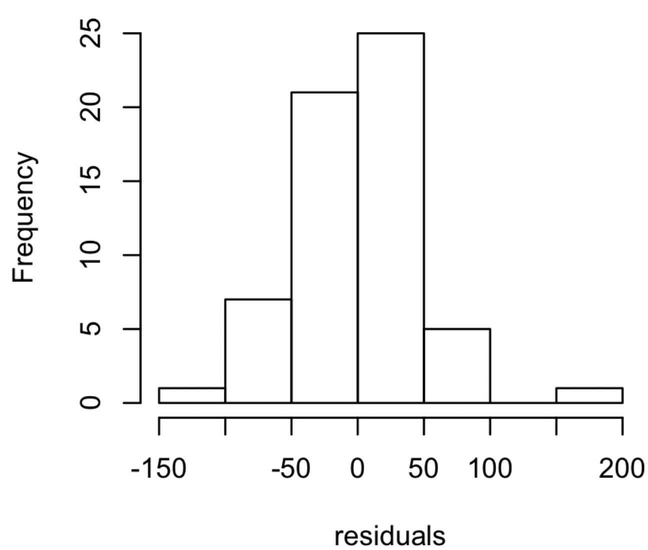
(c) continued. Interpret the coefficient `POPN`. According to this model, what is the effect of having an average household size of 4 versus 3 assuming everything else in the model is equal.

(d) Write the regression equation for model `m2`

(e) Predict the mortality from model `m2` when HC $= 40$, SO $= 50$, NOX $= 25$, and POPN $= 3$ (These are close to the mean values for the data set).

(f) Comment on the histogram of the residuals and diagnostic plots. If you were analyzing this data, would you have any concerns? Is there something you would consider doing to reanalyze the data?


Residuals vs Fitted
Normal Q-Q
Scale-Location
Residuals vs Leverage
Residuals
Fitted values
Standardized residuals
Theoretical Quantiles
√|Standardized residuals|
Leverage
Cook's distance


Frequency
residuals

Recall the Craigslist car data from earlier in the semester. Here we'll use logistic regression to model the probability that a car has a clean versus salvage title as a function of the price. Only the first ten observations were used.

```
> x <- read.table("cars2.txt",header=T)
> x
   year price  miles   title
1  1995  1200 150000   clean
2  2004  4500 184000 salvage
3  1995  3200    NaN   clean
4  1998  1850 152000 salvage
5  1998  3400 136000   clean
6  2004  8500  85500   clean
> title2 <- as.numeric(title=="clean")
> m3 <- glm(title2 ~ price,family="binomial")
> summary(m3)


Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.72342   0.08737   0.29287   0.78568   1.12693

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.2831230  1.5423663  -0.184    0.854
price        0.0003359  0.0003614   0.929    0.353
```

(a) Write the regression equation as a function of the log-odds of the probability of a clean title.

(b) Based on the model, what is the probability that a $5000 car has a salvage title?

(c) Just based on AIC in the following output, is price or year a better predictor of title status?

```
> model1 <- glm(title2[1:10] ~ x$year[1:10],family="binomial")
> model2 <- glm(title2[1:10] ~ x$price[1:10],family="binomial")
> AIC(model1)
[1] 13.98852
> AIC(model2)
[1] 12.04338
```