# STAT 420 Homework 10


### Name: Kimmy Liu

**Netid: zl32**

**Section: 4UG**


## Problem 1

```
library(faraway)
```

```
## Warning: package 'faraway' was built under R version 4.0.3
```

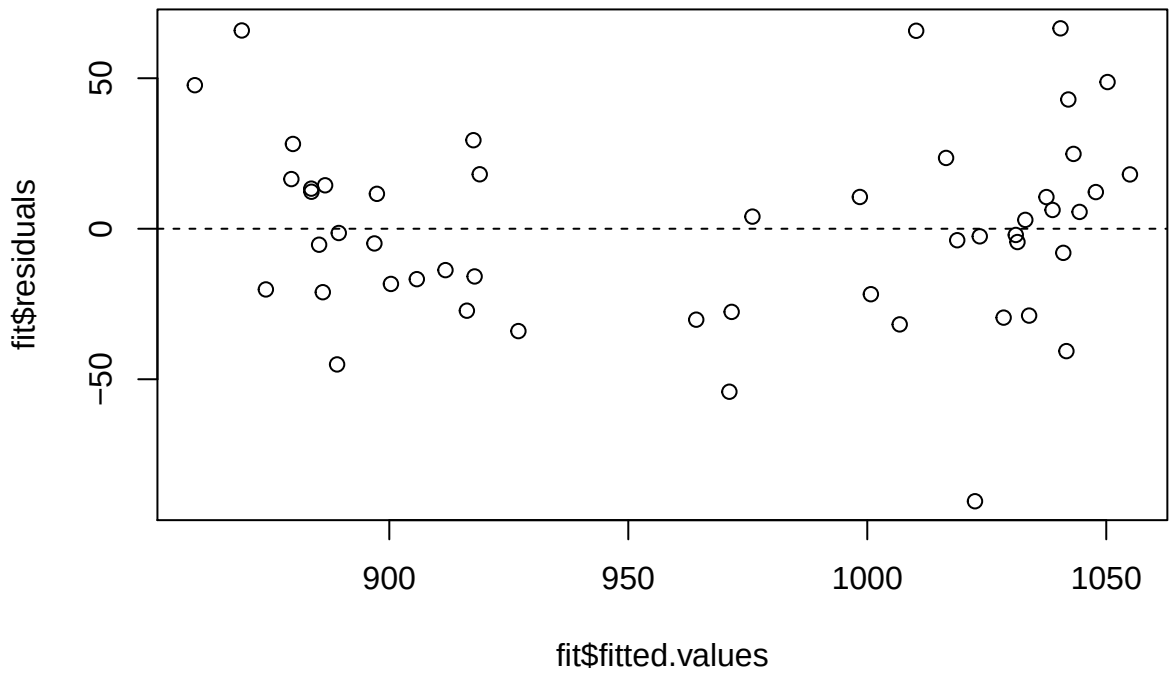
```
data(sat)
head(sat)
```

```
##            expend ratio salary takers verbal math total
## Alabama     4.405  17.2 31.144      8    491  538  1029
## Alaska      8.963  17.6 47.951     47    445  489   934
## Arizona     4.778  19.3 32.175     27    448  496   944
## Arkansas    4.459  17.1 28.934      6    482  523  1005
## California  4.992  24.0 41.078     45    417  485   902
## Colorado    5.443  18.4 34.571     29    462  518   980
```

### a)

```
fit = lm(total~expend+ratio+salary+takers, data = sat)
```

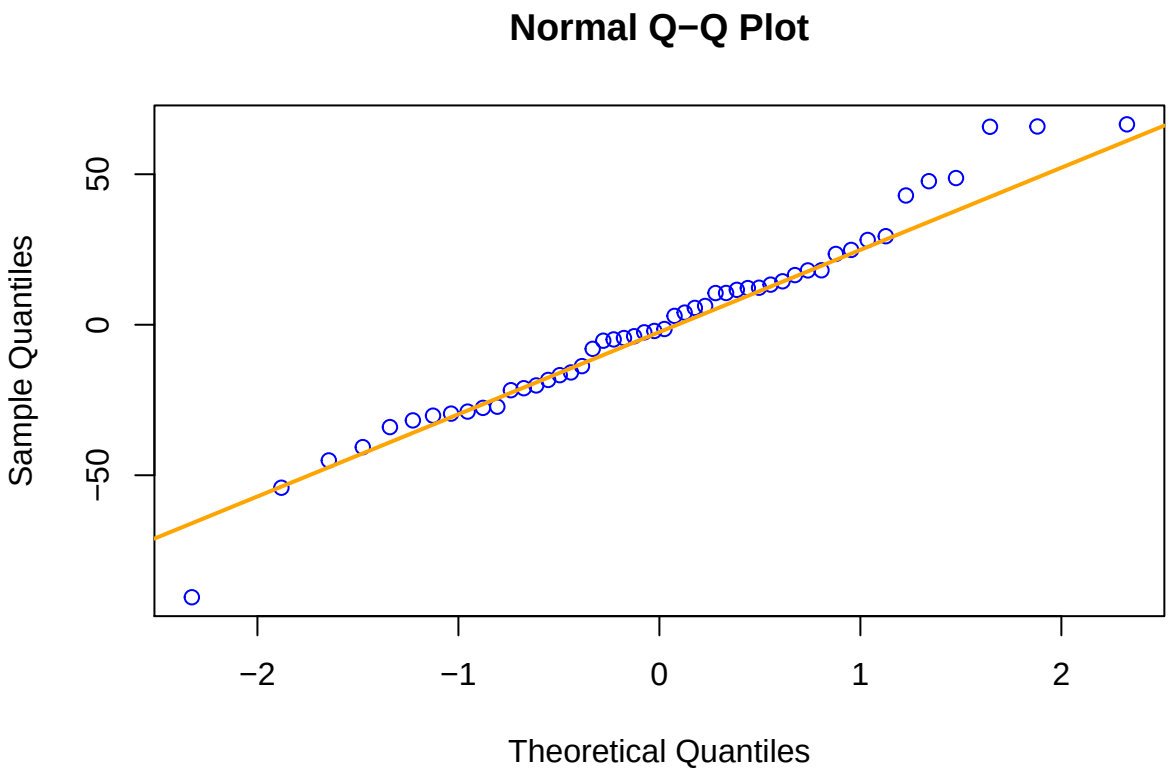
```
plot(fit$fitted.values,fit$residuals)
abline(h=0,lty=2)
```




The residuals versus the fitted values plot suggests that the variance for errors may not be constant since points in the plot seem to have a curved pattern.

**b)**

```
qqnorm(resid(fit), col = "blue")
qqline(resid(fit), col = "orange", lwd = 2)
```



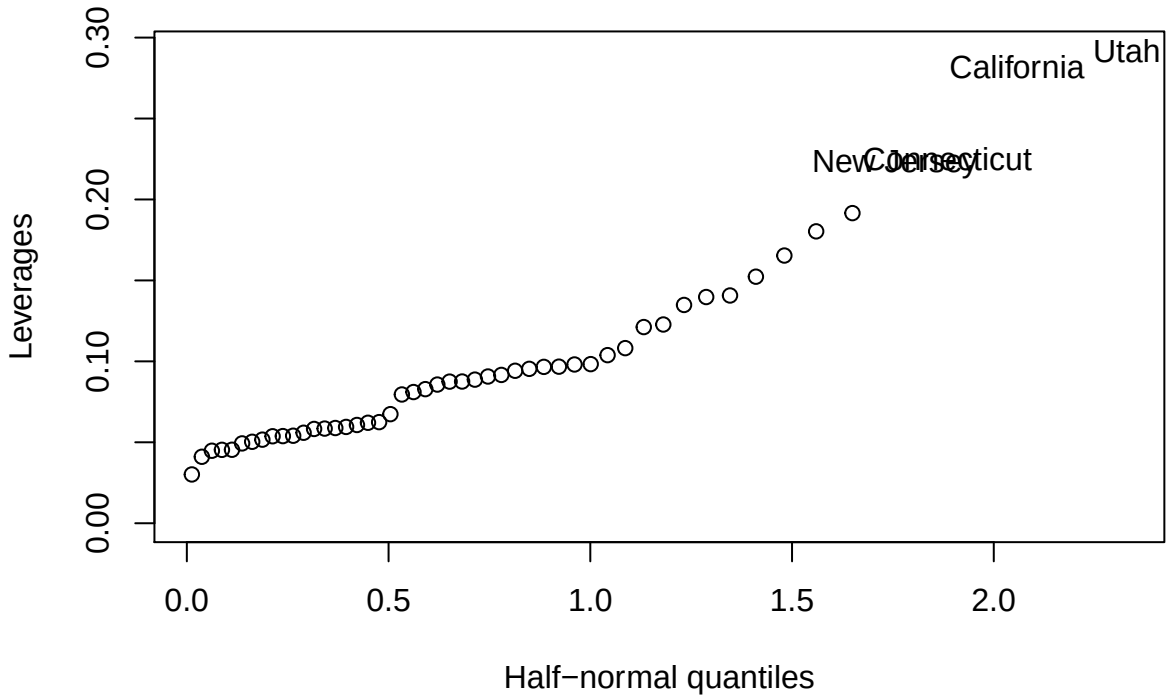

Since the points of the plot do not closely follow a straight line, this suggests that the data do not come from a normal distribution.

c)

```
n = 50
p = 5
lev = influence(fit)$hat
lev[lev>2*p/n]
```

```
##  California Connecticut  New Jersey        Utah
##   0.2821179   0.2254519   0.2220978   0.2921128
```

```
halfnorm(lev, 4, labs = row.names(sat), ylab = "Leverages")
```




From the plot we can see 4 large leverage points corresponding to California, Connecticut, New Jersey and Utah.

**d)**

```
cv=qt(0.05/(2*n),df=df.residual(fit))
(hlobs <- which(influence(fit)$hat > 2 * p / n))
```

```
##  California Connecticut  New Jersey        Utah
##           5           7          30          44
```

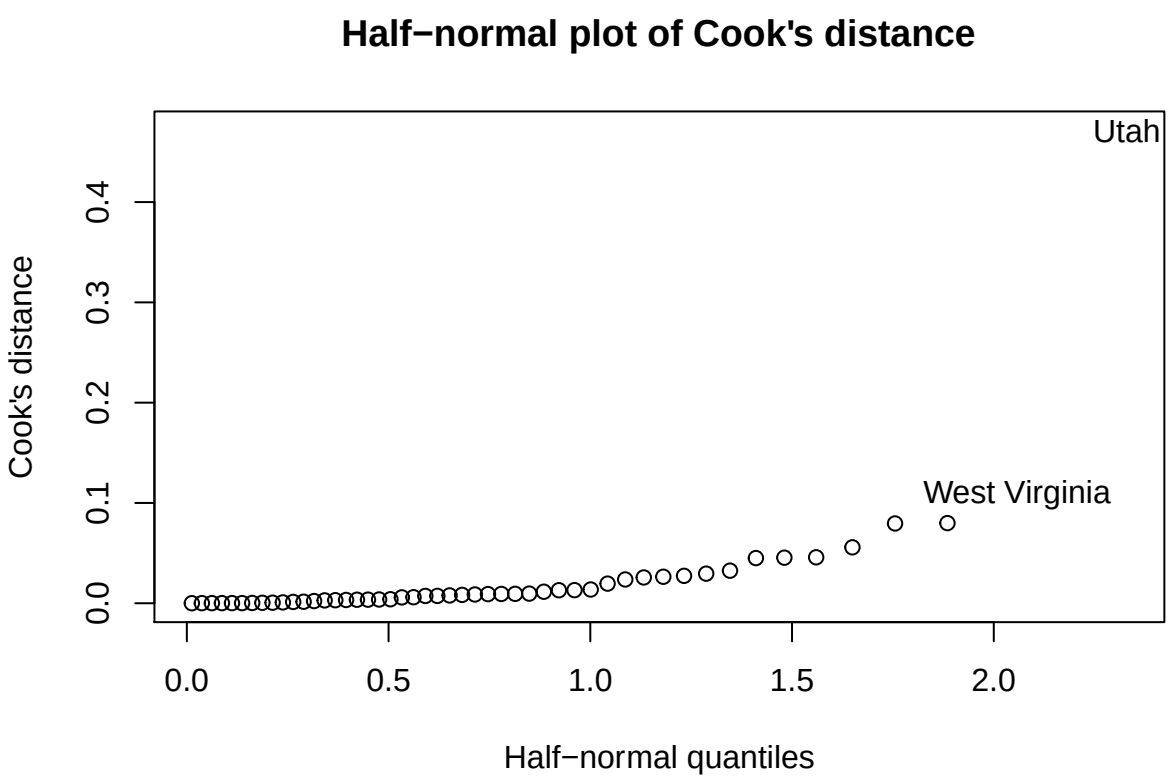
```
which(abs(rstudent(fit)[hlobs]) > abs(cv))
```

```
## named integer(0)
```

We can see that none of the observations is rejected as an outlier after Bonferroni adjustment for the sample size.

**e)**

```
cook = cooks.distance(fit)
halfnorm(cook, labs = row.names(sat), ylab = "Cook's distance",
main = "Half-normal plot of Cook's distance")
```




```
max(cook)
```

```
## [1] 0.4715287
```

According to the rule-of-thumb (CD >= 1), there are not influential observations. However, there is one observation that is too far from the rest which correspond to "Utah"

**f)**

```
summary(fit)
```

```
##
## Call:
## lm(formula = total ~ expend + ratio + salary + takers, data = sat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -90.531 -20.855  -1.746  15.979  66.571
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1045.9715    52.8698  19.784  < 2e-16 ***
## expend         4.4626    10.5465   0.423    0.674
## ratio         -3.6242     3.2154  -1.127    0.266
```

```
## salary         1.6379     2.3872   0.686    0.496
## takers        -2.9045     0.2313 -12.559 2.61e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.7 on 45 degrees of freedom
## Multiple R-squared:  0.8246, Adjusted R-squared:  0.809
## F-statistic: 52.88 on 4 and 45 DF,  p-value: < 2.2e-16
```

expend is the least significant variable, p-value = 0.8439.

```
fit1 = update(fit, .~. - expend)
summary(fit1)
```

```
##
## Call:
## lm(formula = total ~ ratio + salary + takers, data = sat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -89.244 -21.485  -0.798  17.685  68.262
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1057.8982    44.3287  23.865   <2e-16 ***
## ratio         -4.6394     2.1215  -2.187   0.0339 *
## salary         2.5525     1.0045   2.541   0.0145 *
## takers        -2.9134     0.2282 -12.764   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.41 on 46 degrees of freedom
## Multiple R-squared:  0.8239, Adjusted R-squared:  0.8124
## F-statistic: 71.72 on 3 and 46 DF,  p-value: < 2.2e-16
```

All variables are significant at 0.10. "Best" model is lm(formula = total ~ ratio + salary + takers, data = sat)

g)

```
step(fit, direction = "backward")
```

```
## Start:  AIC=353.48
## total ~ expend + ratio + salary + takers
##
##          Df Sum of Sq    RSS    AIC
## - expend  1       191  48315 351.67
## - salary  1       503  48627 352.00
```

```
## - ratio   1      1359  49483 352.87
## <none>                 48124 353.48
## - takers  1    168688 216812 426.74
##
## Step:  AIC=351.67
## total ~ ratio + salary + takers
##
##          Df Sum of Sq    RSS    AIC
## <none>                 48315 351.67
## - ratio   1      5023  53338 354.62
## - salary  1      6782  55097 356.24
## - takers  1    171126 219441 425.34
```

```
##
## Call:
## lm(formula = total ~ ratio + salary + takers, data = sat)
##
## Coefficients:
## (Intercept)        ratio       salary       takers
##    1057.898       -4.639        2.552       -2.913
```

"Best" model is lm(formula = total ~ ratio + salary + takers, data = sat)

**h)**

```
attach(sat)
step(lm(total ~ 1), total ~ expend + ratio + salary + takers,
direction = "forward")
```

```
## Start:  AIC=432.5
## total ~ 1
##
##          Df Sum of Sq    RSS    AIC
## + takers  1    215875  58433 357.18
## + salary  1     53078 221230 423.75
## + expend  1     39722 234586 426.68
## <none>                274308 432.50
## + ratio   1      1811 272497 434.17
##
## Step:  AIC=357.18
## total ~ takers
##
##          Df Sum of Sq   RSS    AIC
## + expend  1    8913.1 49520 350.91
## + salary  1    5094.8 53338 354.62
## + ratio   1    3336.2 55097 356.24
## <none>                58433 357.18
```

```
##
## Step:  AIC=350.91
## total ~ takers + expend
##
##          Df Sum of Sq   RSS    AIC
## <none>                49520 350.91
## + ratio   1    892.74 48627 352.00
## + salary  1     37.52 49483 352.87

##
## Call:
## lm(formula = total ~ takers + expend)
##
## Coefficients:
## (Intercept)       takers       expend
##     993.832       -2.851       12.287
```

"Best" model is lm(formula = total ~ takers + expend)

**i)** (h) is preferred with lower AIC compared to (g)

**j)** (g):

```
summary(lm(formula = total ~ ratio + salary + takers, data = sat))
```

```
##
## Call:
## lm(formula = total ~ ratio + salary + takers, data = sat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -89.244 -21.485  -0.798  17.685  68.262
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1057.8982    44.3287  23.865   <2e-16 ***
## ratio         -4.6394     2.1215  -2.187   0.0339 *
## salary         2.5525     1.0045   2.541   0.0145 *
## takers        -2.9134     0.2282 -12.764   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.41 on 46 degrees of freedom
## Multiple R-squared:  0.8239, Adjusted R-squared:  0.8124
## F-statistic: 71.72 on 3 and 46 DF,  p-value: < 2.2e-16
```

(h):

```r
summary(lm(formula = total ~ expend + takers, data = sat))
```

```
##
## Call:
## lm(formula = total ~ expend + takers, data = sat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -88.400 -22.884   1.968  19.142  68.755
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 993.8317    21.8332  45.519  < 2e-16 ***
## expend       12.2865     4.2243   2.909  0.00553 **
## takers       -2.8509     0.2151 -13.253  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.46 on 47 degrees of freedom
## Multiple R-squared:  0.8195, Adjusted R-squared:  0.8118
## F-statistic: 106.7 on 2 and 47 DF,  p-value: < 2.2e-16
```

(g) is preferred with larger Adjusted R-squared.

## Problem 2

### c)

```r
n=34
p.null=3
p.full=5
RSS.null=528
RSS.full=448
```

```r
aic.null = n*log(RSS.null/n) + 2*p.null
aic.full = n*log(RSS.full/n) + 2*p.full
c(aic.null, aic.full)
```

```
## [1] 99.25302 97.66671
```

Since aic.full<aic.null, full model is preferred.

### d)

```r
bic.null = n*log(RSS.null/n) + log(n)*p.null
bic.full = n*log(RSS.full/n) + log(n)*p.full
c(bic.null, bic.full)
```

```
## [1] 103.8321 105.2985
```

Since bic.null<bic.full, null model is preferred.

e)

```
syy=748
rsq.null=1-RSS.null/syy
rsq.full=0.40107
Adjusted.Rsq.null = 1-(1-rsq.null)*(n-1)/(n-p.null-1)
Adjusted.Rsq.full = 1-(1-rsq.full)*(n-1)/(n-p.full-1)
c(Adjusted.Rsq.null,Adjusted.Rsq.full)
```

```
## [1] 0.2235294 0.2941182
```

Full model is preferred since full model has larger adjusted R-square.