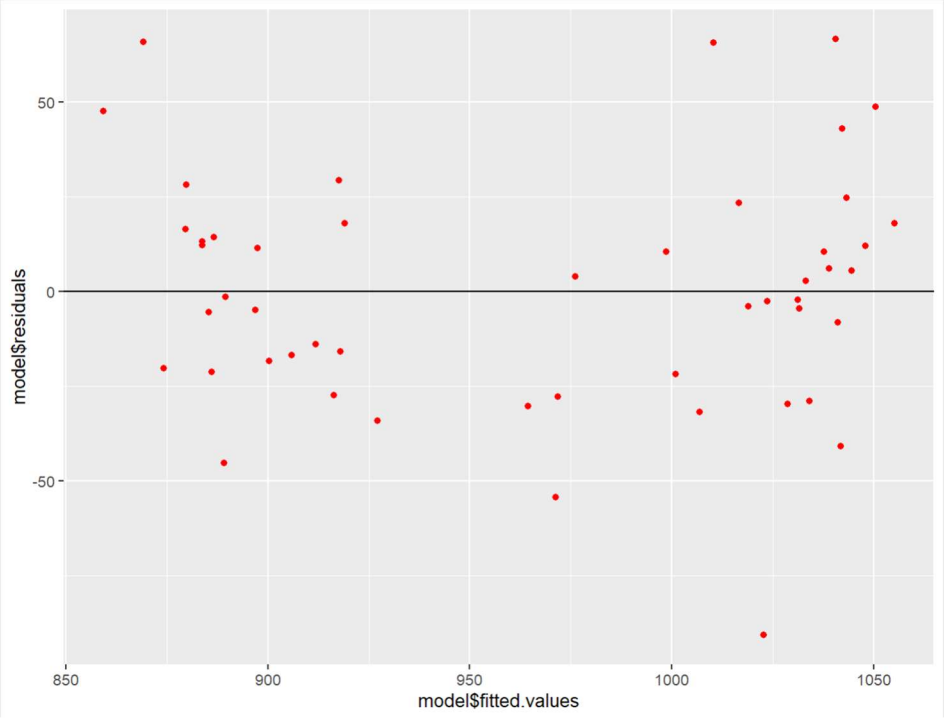
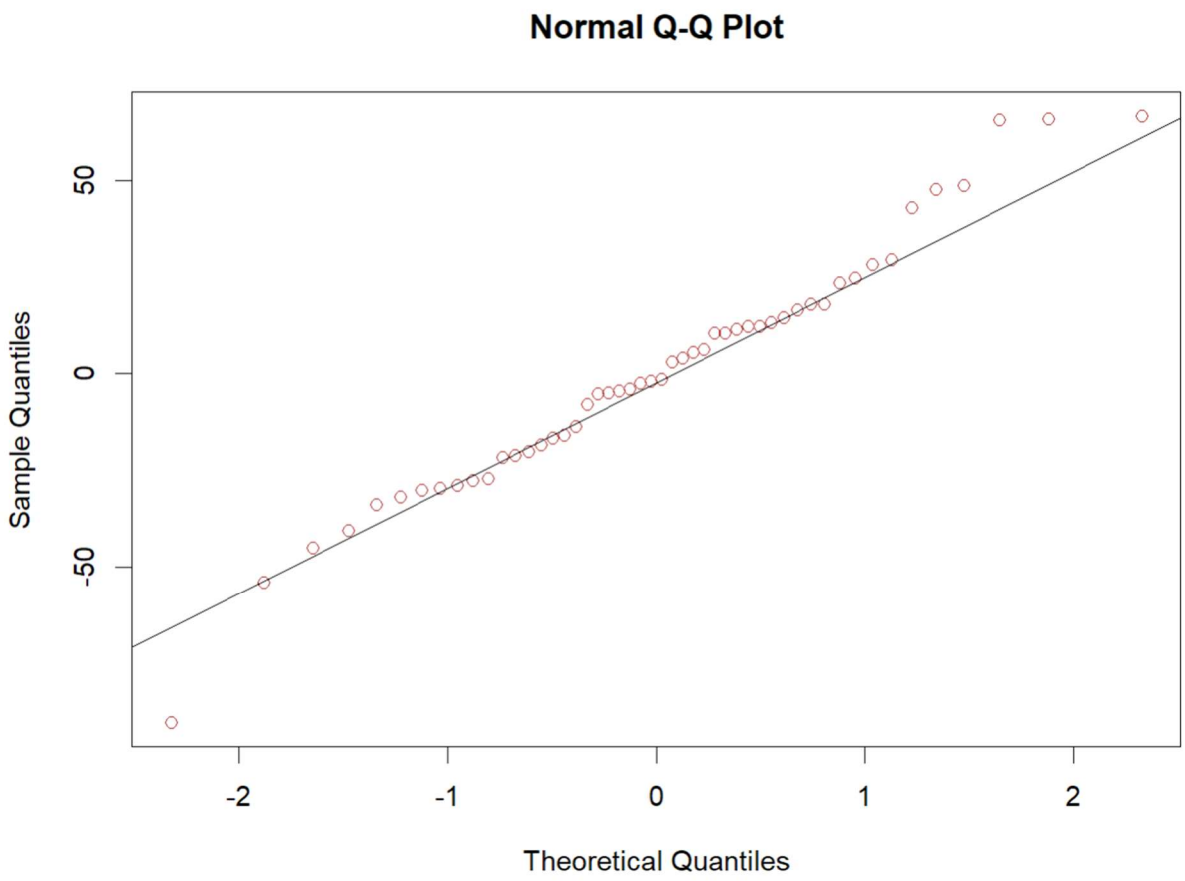
Saad Ahmad

Question 1

```
library(faraway)
library(ggplot2)

data(sat)
?sat
head(sat)

model = lm(sat$total ~ sat$expend + sat$salary + sat$ratio + sat$takers)
model
summary(model)

ggplot(data = sat,
       aes(x = model$fitted.values, y = model$residuals)) +
       geom_abline(slope = 0) +
       geom_point(col = "red")
```



```
qqnorm(model$residuals, col = "brown")
qqline(model$residuals, col = "black")
```



Takeaways:

- The plot of Fitted.Values vs Residuals shows scattered data points which implies it is homoscedasticity and it lacks heteroscedasticity.
- The residuals are distributed normally.

Question 2

```
data(longley)
?longley
head(longley)

model2 = lm(longley$Employed ~ longley$GNP.deflator + longley$GNP +
longley$Unemployed + longley$Armed.Forces + longley$Population + longley$Year)
summary(model2)
```

```
Call:
lm(formula = longley$Employed ~ longley$GNP.deflator + longley$GNP +
    longley$Unemployed + longley$Armed.Forces + longley$Population +
    longley$Year)

Residuals:
     Min       1Q   Median       3Q      Max
-0.41011 -0.15767 -0.02816  0.10155  0.45539

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          -3.482e+03  8.904e+02  -3.911 0.003560 **
longley$GNP.deflator  1.506e-02  8.492e-02   0.177 0.863141
longley$GNP          -3.582e-02  3.349e-02  -1.070 0.312681
longley$Unemployed   -2.020e-02  4.884e-03  -4.136 0.002535 **
longley$Armed.Forces -1.033e-02  2.143e-03  -4.822 0.000944 ***
longley$Population   -5.110e-02  2.261e-01  -0.226 0.826212
longley$Year          1.829e+00  4.555e-01   4.016 0.003037 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3049 on 9 degrees of freedom
Multiple R-squared:  0.9955,    Adjusted R-squared:  0.9925
F-statistic: 330.3 on 6 and 9 DF,  p-value: 4.984e-10
```

Compute and comment on the condition numbers.

kappa(longley[,-1])

The number is really high.

Compute and comment on the correlations between the predictors

?cor

cor(longley)

```
> cor(longley)
             GNP.deflator       GNP Unemployed Armed.Forces Population      Year  Employed
GNP.deflator    1.0000000 0.9915892  0.6206334    0.4647442  0.9791634 0.9911492 0.9708985
GNP             0.9915892 1.0000000  0.6042609    0.4464368  0.9910901 0.9952735 0.9835516
Unemployed      0.6206334 0.6042609  1.0000000   -0.1774206  0.6865515 0.6682566 0.5024981
Armed.Forces    0.4647442 0.4464368 -0.1774206    1.0000000  0.3644163 0.4172451 0.4573074
Population      0.9791634 0.9910901  0.6865515    0.3644163  1.0000000 0.9939528 0.9603906
Year            0.9911492 0.9952735  0.6682566    0.4172451  0.9939528 1.0000000 0.9713295
Employed        0.9708985 0.9835516  0.5024981    0.4573074  0.9603906 0.9713295 1.0000000
> ##### There is a high correlation (p) between the variables
```

There is a high correlation (p) between the variables

Compute and comment on the variance inflation factors.

library(faraway)

model3 = lm(Employed ~ ., data = longley)

vif(model3)

```
> library(faraway)
> model3 = lm(Employed ~ ., data = longley)
> vif(model3)
GNP.deflator          GNP   Unemployed Armed.Forces   Population         Year
   135.53244   1788.51348     33.61889      3.58893    399.15102    758.98060
> ##### On comparing VIF among other variables, Armed.Forces has 3.59.
```

On comparing VIF among other variables, Armed.Forces has 3.59.

Choose a reduced set of predictors that does not exhibit as much collinearity as the full set, fit a new linear model with this reduced set, and comment on the differences between the reduced model and the full model.

```
model4 = lm(longley$Employed ~ longley$GNP.deflator + longley$GNP
                              + longley$Population + longley$Year)

summary(model4)
```

```
Call:
lm(formula = longley$Employed ~ longley$GNP.deflator + longley$GNP +
    longley$Population + longley$Year)

Residuals:
     Min       1Q   Median       3Q      Max
-0.70332 -0.40823 -0.04122  0.30160  0.89940

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -296.73890  861.69595  -0.344 0.737064
longley$GNP.deflator   -0.18159    0.12492  -1.454 0.173973
longley$GNP             0.08090    0.01716   4.715 0.000635 ***
longley$Population     -0.52802    0.21825  -2.419 0.034047 *
longley$Year            0.21037    0.45449   0.463 0.652490
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5392 on 11 degrees of freedom
Multiple R-squared:  0.9827,    Adjusted R-squared:  0.9764
F-statistic: 156.4 on 4 and 11 DF,  p-value: 1.299e-09
```

These variables which I used are highly collinear.