Stat 500 – Homework 4 (Solutions)

Read in data and fit the model.

```
> library(faraway)
> data(sat)
> g<-lm(total˜expend+salary+ratio+takers,data=sat)
```

1.
   ```
   > plot(g$fit, g$res, xlab="Fitted", ylab="Residuals")
   > abline(h=0)
   > plot(sat$takers, g$res, xlab="Takers", ylab="Residuals")
   ```

   In the residual versus fitted plot in Figure 1(a) we notice a non-linear relationship. After examining plots of each of the variables versus the residuals, we notice takers shows a non-linear relationship, which appears to be quadratic (see Figure 1(b)). So to fix this we will add $takers^2$ to the model.

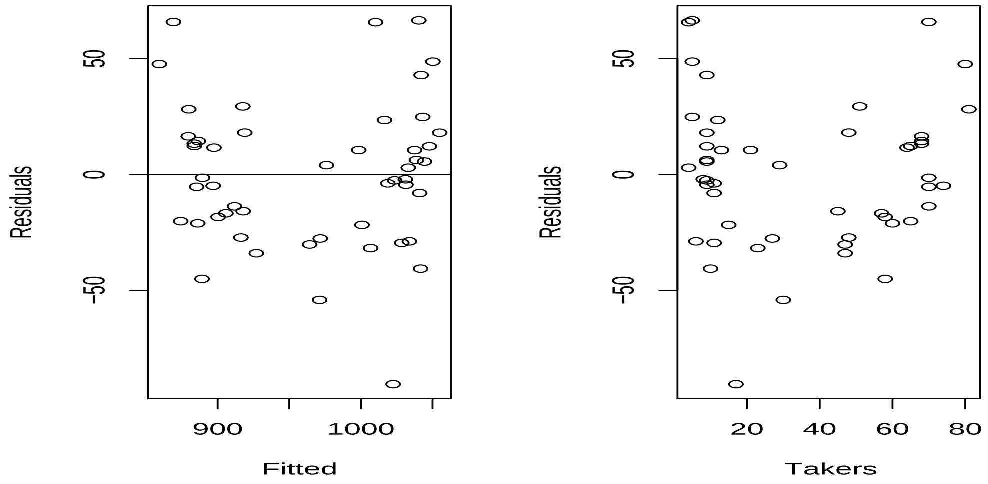

Figure 1: (a) residuals vs. fitted shows some non-linearity worthy of investigation. (b) residuals vs. takers shows quadratic relationship between takers and the residuals

```
> gadj<-lm(total˜expend+salary+ratio+takers+I(takers^2),data=sat)
> plot(gadj$fit, gadj$res, xlab="Fitted", ylab="Residuals")
> abline(h=0)
> qqnorm(gadj$res)
```

Now the residual versus fitted plot in Figure 2(a) looks much better and there does not appear to be any problems with non-constant variance.

2. The qq-plot in Figure 2(b) shows no real issue with normality.
3. A half-normal plot will help identify leverage points.

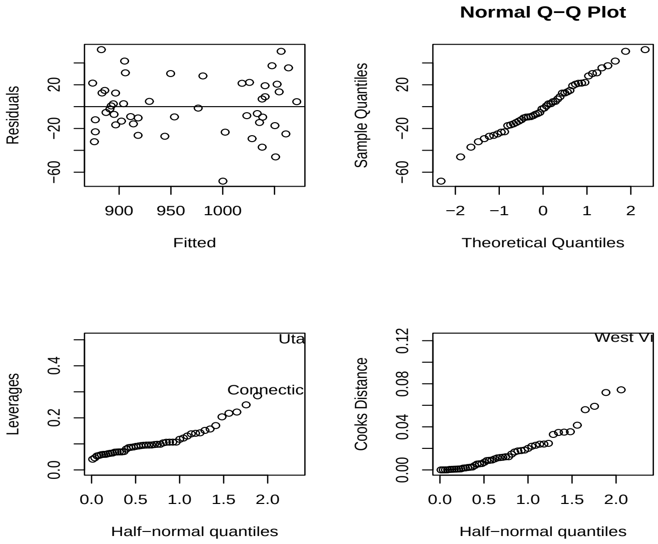

Figure 2: For the adjusted model:(a) fitted vs. residuals (b) normal q-q plot (c) Half-normal plot of leverages (d) Half-normal plot of Cook's distances

```
>  halfnorm(lm.influence(gadj)$hat,labs=row.names(sat),ylab="Leverages")
```

From Figure 2(c), we see that **Utah** and possibly **Connecticut** are high leverage points.

4. In this exercise we are going to use jackknife residuals and the Bonferroni correction to see if any observations stand out.

```
> jack<-rstudent(gadj)
> jack[order(abs(jack),decreasing=TRUE)][1:5]
West Virginia New Hampshire  North Dakota      Arkansas
    -2.940391      2.165300      2.097323     -1.861728
       Oregon
     1.711113
> qt(0.05/(50*2),44)
[1] -3.525801
```

From this somewhat conservative test, we see that none of these residuals are significant.

2. (a) Now

$$\begin{aligned}\hat{\epsilon} &= y - \hat{y} \\ &= y - Hy \\ &= (I-H)y \\ &= (I-H)X\beta + (I-H)\epsilon\end{aligned}$$

Since first term $(I-H)X\beta$ is deterministic,

$$\begin{aligned}\mathrm{Var}(\hat{\epsilon}) &= \mathrm{Var}((I-H)\epsilon) \\ &= (I-H)^\top \mathrm{Var}(\epsilon)(I-H) \\ &= (I-H)^\top \mathrm{Var}(\epsilon)(I-H) \\ &= (I-H)^\top\ (\sigma^2 I)\ (I-H) \\ &= \sigma^2(I-H)(I-H) \qquad\qquad \text{since } H \text{ is projection matrix} \\ &= \sigma^2(I-2H+H^2) \\ &= \sigma^2(I-H) \qquad\qquad \text{since } H \text{ is projection matrix}\end{aligned}$$

invoking that H is a projection matrix.

(b) As shown in the part **(a)**, we showed that residuals do not have a constant variance, but have a standard deviation depending on the diagonal element of H. So we know that the variance is not constant, and so to have the best diagnostic plot on the constancy of the variance, it would be be best to plot the normalized studentized residuals whose variance is a constant, provided that the original linear model assumptions hold.