Data 621 LMR Ex 6.1

Sin Ying Wong

2/20/2020

LMR Exercise 6.1

Using the `sat` dataset, fit a model with the total SAT score as the response and expend, salary, ratio and takers as predictors. Perform regression diagnostics on this model to answer the following questions. Display any plots that are relevant. Do not provide any plots about which you have nothing to say. Suggest possible improvements or corrections to the model where appropriate.

(a) Check the constant variance assumption for the errors.
(b) Check the normality assumption.
(c) Check for large leverage points.
(d) Check for outliers.
(e) Check for influential points.
(f) Check the structure of the relationship between the predictors and the response.

Let's load up the data.

```
data(sat, package='faraway')
```

A. Check the constant variance assumption for errors

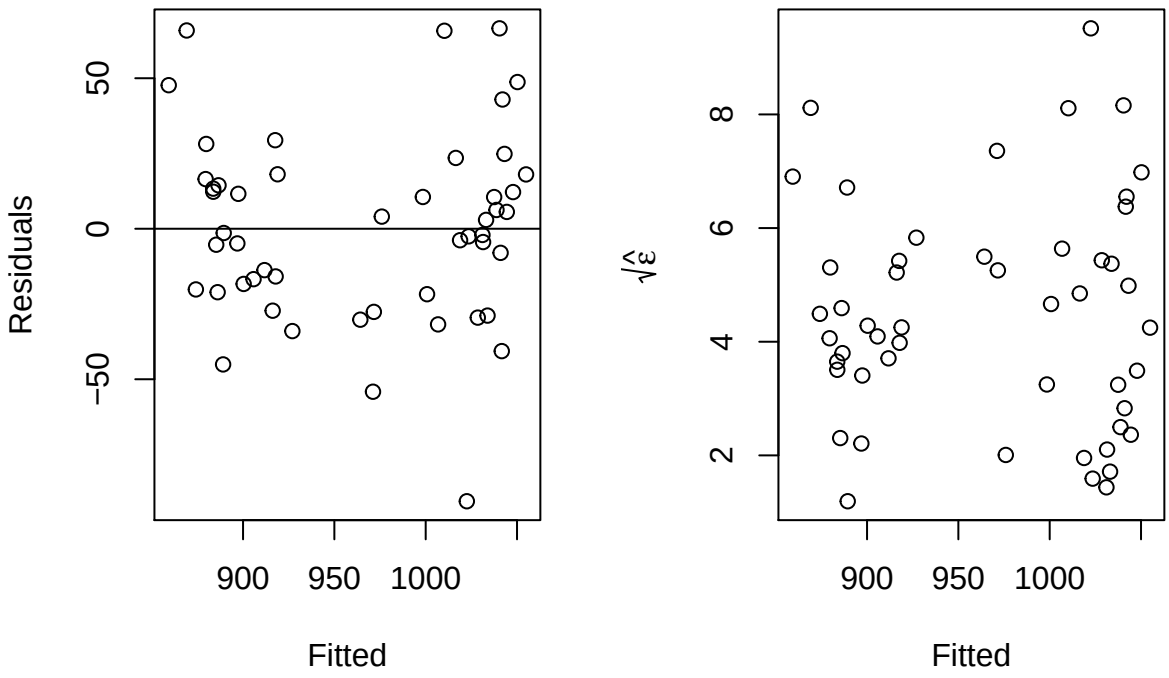
The plots, as seen below, shows approximately constant variation.

```
lmod <- lm(total~expend+salary+ratio+takers, sat)

par(mfrow=c(1,2))

plot(fitted(lmod), residuals(lmod), xlab="Fitted", ylab="Residuals",
     title="Residuals against Fitted values")
abline(h=0)

plot(fitted(lmod), sqrt(abs(residuals(lmod))), xlab="Fitted",
     ylab=expression(sqrt(hat(epsilon))),
     title="Residuals against Fitted values")
abline(h=0)
```



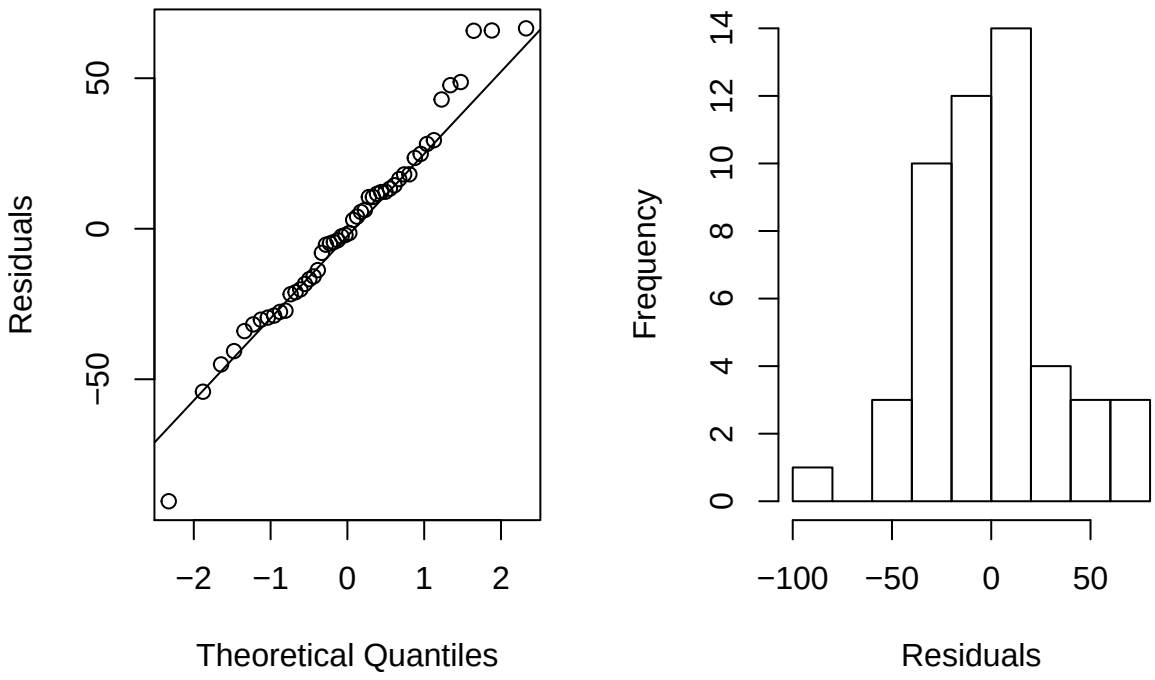

```
var.test(residuals(lmod), sqrt(abs(residuals(lmod))))
```

```
##
##  F test to compare two variances
##
## data:  residuals(lmod) and sqrt(abs(residuals(lmod)))
## F = 251.35, num df = 49, denom df = 49, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  142.6339 442.9224
## sample estimates:
## ratio of variances
##           251.3479
```

B. Check the normality assumption

```
par(mfrow=c(1,2))
qqnorm(residuals(lmod), ylab="Residuals", main="")
qqline(residuals(lmod))

hist(residuals(lmod), xlab="Residuals", main="")
```



C. Check for large leverage points

```
hatv <- hatvalues(lmod)
head(hatv)
```

```
##    Alabama     Alaska    Arizona   Arkansas California   Colorado
## 0.09537668 0.18030612 0.04931612 0.05382878 0.28211791 0.03014533
```

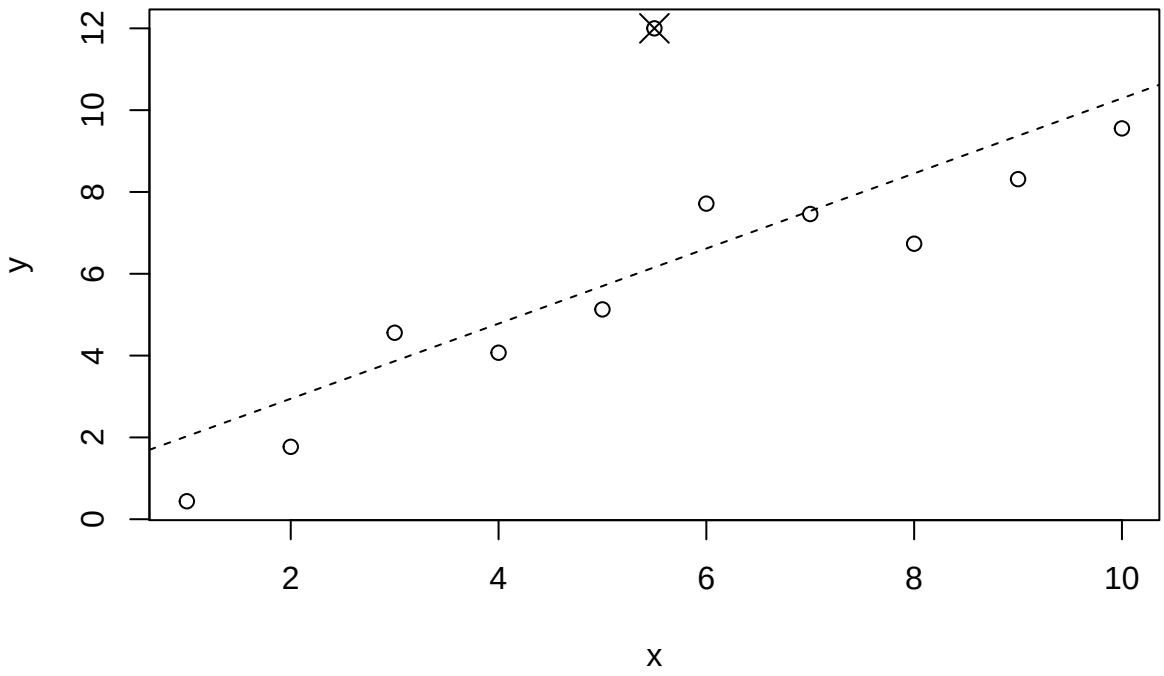
```
sum(hatv)
```

```
## [1] 5
```

D. Check for outliers

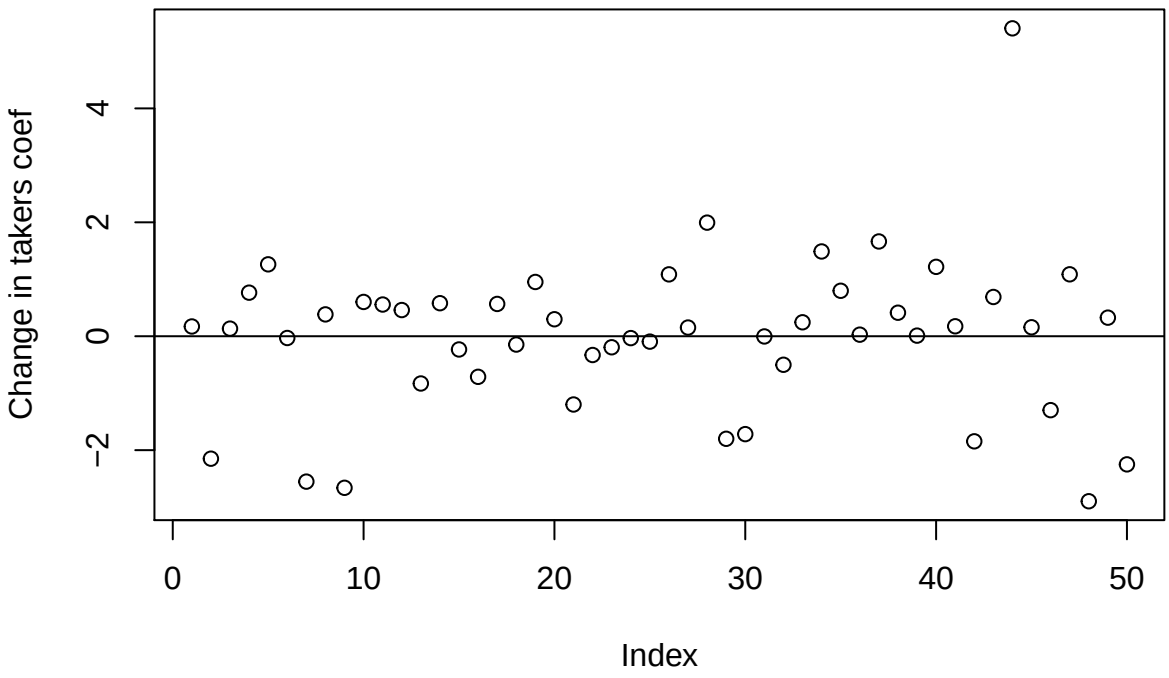
```
set.seed(123)
testdata <- data.frame(x=1:10, y=1:10+rnorm(10))
lmod1 <- lm(y~x, testdata)

p1 <- c(5.5,12)
lmod2 <- lm(y~x, rbind(testdata, p1))
plot(y~x, rbind(testdata, p1))
points(5.5, 12, pch=4, cex=2)
abline(lmod)
abline(lmod2, lty=2)
```



E. Check for influential points

```
plot(dfbeta(lmod)[,2], ylab="Change in takers coef")
abline(h=0)
```



F. Check the structure of the relationship between the predictors and the response

```
summary(lmod)
```

```
##
## Call:
## lm(formula = total ~ expend + salary + ratio + 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
## salary         1.6379     2.3872   0.686    0.496
## ratio         -3.6242     3.2154  -1.127    0.266
## 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
```

```
d <- residuals(lm(total~expend+salary+ratio+takers, sat))
m <- residuals(lm(takers~expend+salary+ratio, sat))
plot(m, d, xlab="takers residuals", ylab="Sat Totals residuals")
abline(0, coef(lmod)['takers'])
```