Question 5

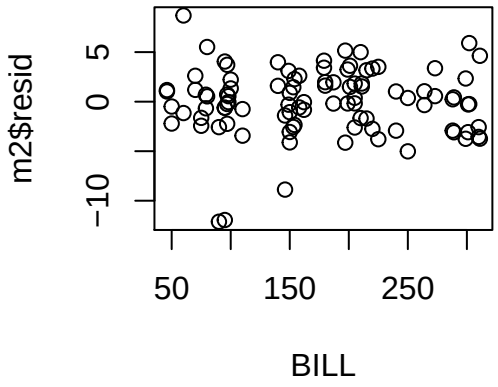
Since we have a very small number of variables we will do 'classical' model selection by starting with the saturated model and seeing if we can remove the interaction term.

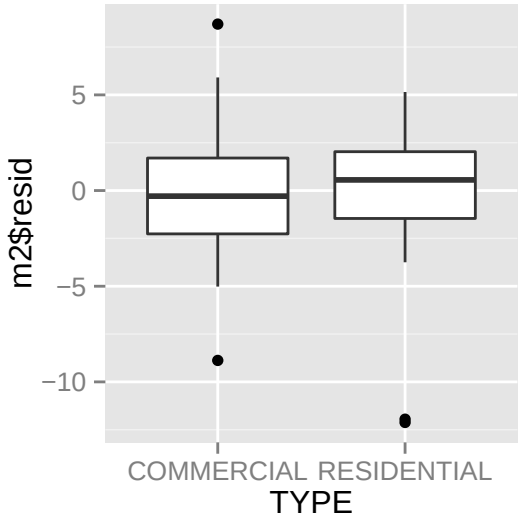
```
m2 = lm(LATE ~ BILL * TYPE)
summary(m2)$coef
```

```
##                       Estimate Std. Error t value  Pr(>|t|)
## (Intercept)           101.7582   1.198504   84.90 3.475e-89
## BILL                   -0.1910   0.006285  -30.38 9.386e-50
## TYPERESIDENTIAL       -99.5486   1.694940  -58.73 9.905e-75
## BILL:TYPERESIDENTIAL    0.3566   0.008888   40.13 4.370e-60
```

We see that we need to include all covariates in our model by the small values of the wald tests. Next we will do model checking to make sure we have a good fit. First we will check for patterns in our residuals.

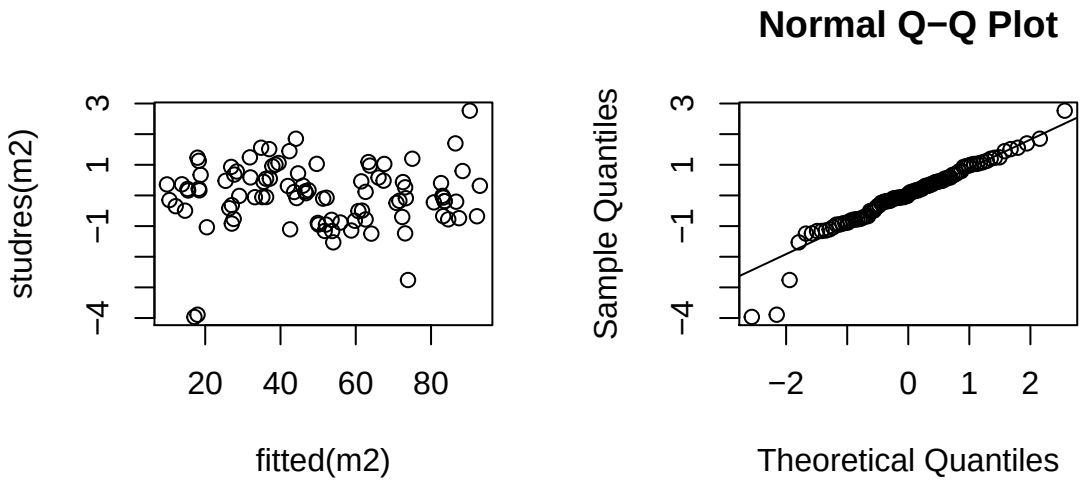
```
plot(BILL, m2$resid)
ggplot(data, aes(x = TYPE, y = m2$resid)) + geom_boxplot()
```





With no obvious patterns in the resudials, we will now check for outliers

```
plot(fitted(m2), studres(m2))
qqnorm(studres(m2))
qqline(studres(m2))
```



Although a couple of points may arguably be outliers, we will keep them in the model as they are not too bad and the outliers are unlikely to change the estimates of the model. Next we will check for poitns of influence

```
sort(lm.influence(m2)$hat)[1:10]

##      46      54      90      23      33      87      28      86      20
## 0.02091 0.02091 0.02095 0.02095 0.02136 0.02136 0.02140 0.02140 0.02175
##      59
## 0.02175
```

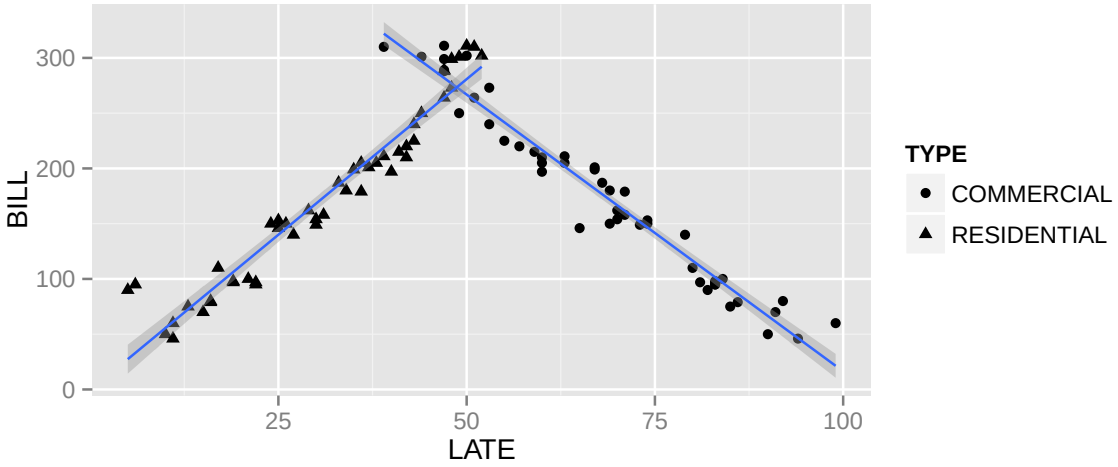
We see that there are no influencial points. Now that we are satisfied with our model we can interpret it. Recall the parameters

```
coef(m2)

##          (Intercept)                 BILL      TYPERESIDENTIAL
##              101.7582              -0.1910              -99.5486
## BILL:TYPERESIDENTIAL
##               0.3566
```

Which we can interpret as follows. For a commercial building with a small bill, we will have to wait about 100 days. For a residental building with a small bill we will have to wait about 0 days. As the bill increases by one, we wil expect to wait .2 less days for a commercial building, and .15 more days for a residental building. We can plot these regression lines as follows.

```
ggplot(data, aes(x = LATE, y = BILL, shape = TYPE)) + geom_point() + stat_smooth(method = lm)
```



Using this plot we can say that the marketing department should make different claims for commercial and residential clients. It seems they can indeed they can collect most of their residential claims in 60 days, however they are unlikely to collect their commercial claims in 60 days. In order to balance simplicity and accuracy I would make the following four claims.

- Residential claims under $1500 can be collected in 30 days
- Residential claims over $150 can be collected in 60 days
- Commercial claims over $150 can be collected in 75 days
- Commercial claims under $150 can be collected in 100 days