

Lab 9: Diagnostics

1. Using the cars2010 data set, run the regression with the following explanatory variables:

EngDispl
Transmission
AirAspirationMethod
TransLockup
TransCreeperGear
DriveDesc
IntakeValvePerCyl
CarlineClassDesc
VarValveLift

```
lm.model=lm(FE~EngDispl+Transmission+AirAspirationMethod+TransLockup+TransCreeperGear+DriveDesc+IntakeValvePerCyl+CarlineClassDesc+VarValveLift,data=cars2010)
```

- a. Let's assume that these observations are ordered throughout time (observation 1 was the first to be observed in time, observation 2 was the 2nd and so forth), check for 1st order autocorrelation using the Durbin-Watson test.

```
> dwtest(lm.model,alternative="greater")
```

Durbin-Watson test

data: lm.model

DW = 1.3354, p-value < 2.2e-16

alternative hypothesis: true autocorrelation is greater than 0

H₀: No Autocorrelation H_A: Autocorrelation, with a test statistic of 1.3354 and a p-value less than 2.2×10^{-16} , we will reject the null hypothesis. There does appear to be significant 1st order autocorrelation present in the data.

- b. Use plots to identify potential influential observations based on the suggested cutoff values.

```
> a =
```

```
ggplot(lm.model,aes(x=n.index,y=rstandard(lm.model)))+geom_point(color="orange")+geom_line(y=-3)+geom_line(y=3)+labs(title = "Internal Studentized Residuals",x="Observation",y="Residuals")
```

```
> b =
```

```
ggplot(lm.model,aes(x=n.index,y=rstudent(lm.model)))+geom_point(color="orange")+geom_line(y=-3)+geom_line(y=3)+labs(title = "External Studentized Residuals",x="Observation",y="Residuals")
```

```
> ##Influential points
```

```

> c =
ggplot(lm.model,aes(x=n.index,y=rstandard(lm.model)))+geom_point(color="orange")+geom_line(y=-
3)+geom_line(y=3)+labs(title = "Internal Studentized Residuals",x="Observation",y="Residuals")

> ##Cook's D

> D.cut=4/(nrow(cars2010)-lm.model$rank)

> d
=ggplot(lm.model,aes(x=n.index,y=cooks.distance(lm.model)))+geom_point(color="orange")+geom_line
(y=D.cut)+labs(title = "Cook's D",x="Observation",y="Cook's Distance")

> ##Dffit

> df.cut=2*(sqrt(lm.model$rank/nrow(cars2010)))

> e
=ggplot(lm.model,aes(x=n.index,y=dffits(lm.model)))+geom_point(color="orange")+geom_line(y=df.cut)
+geom_line(y=-df.cut)+labs(title = "DFFITS",x="Observation",y="DFFITS")

> db.cut=2/sqrt(nrow(cars2010))

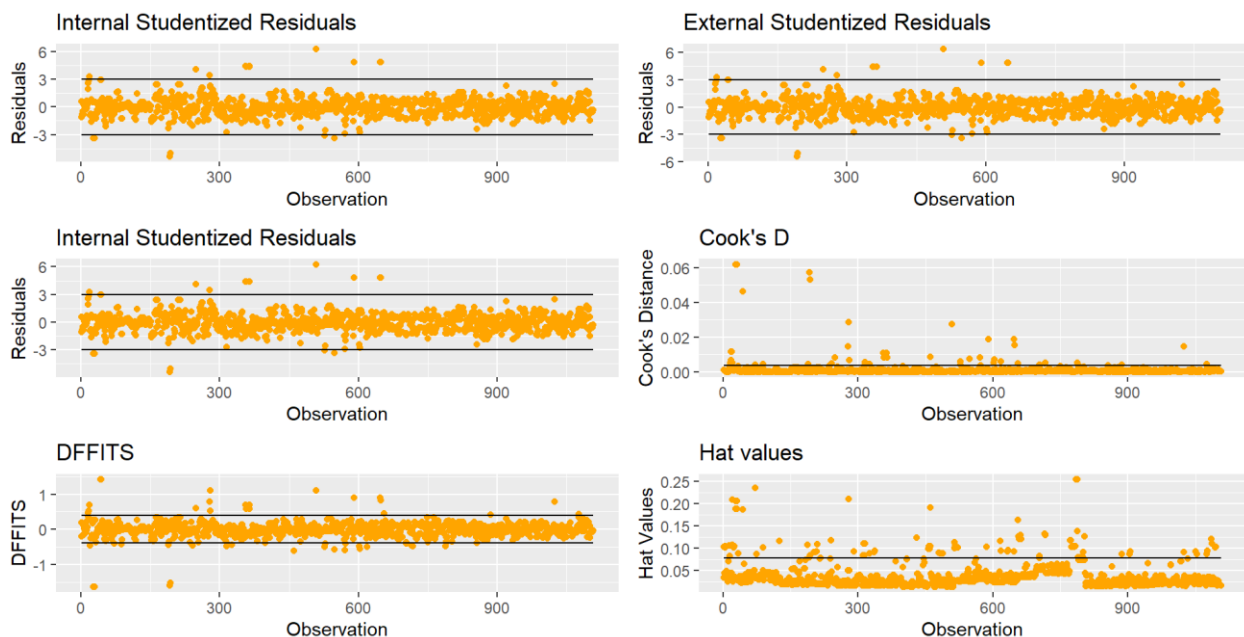
> ##Hat

> hat.cut=2*(lm.model$rank)/nrow(cars2010)

> h =
ggplot(lm.model,aes(x=n.index,y=hatvalues(lm.model)))+geom_point(color="orange")+geom_line(y=hat
.cut)+labs(title = "Hat values",x="Observation",y="Hat Values")

> grid.arrange(a,b,c,d,e,h,ncol=2)

```



- c. Are there any observations with external studentized residuals larger than 3 in magnitude? If so, list the observations.

There are 18 observations:

```
> newcar2[abs(newcar2$Resid)>3,1]
```

```
[1] 18 19 27 30 192 193 248 278 356 357 364 365 508 527  
[15] 548 590 647 648
```