Stat 245 HW #8

Müfitcan Atalay

3/15/2019

First we download the dataset

```
setwd("/Users/mufitcan/Downloads")
x <- dget("NewHaven.txt")
```

a)

now we can summarize the dataset

```
names(x)
```

```
##  [1] "parcelID"       "Address"        "Lat"            "Long"
##  [5] "owner"          "CurVal"         "size"           "LivingArea"
##  [9] "TotalBedrooms"  "TotalBathrooms" "ACtype"         "Grade"
## [13] "Depreciation"   "Year"           "Garage"         "Garage.area"
## [17] "condo"          "house"
```

```
dim(x)
```

```
## [1] 18104    18
```

The dataset contains information on 18104 parcel deliveries. there are 17 variables describing the each parcel delivery that contais information about the delivery location, recipient and quantitative features of the house/location being delivered to.

Some of these variables are:

ParcelID: unique number for each parcel

Address: the adress the parcel is being delivered

owner: the recipient of or the owner of the buiding being delivered to

Year: year of delivery

TotalBathrooms: number of bathrooms

house: dummy for if the location is a house etc.

b) and c)

I will study all the parcels delivered after 1995 because, to me, this seems like the more relevant period for us to study from a time and technology standpoint. I also pick, Year, house, TotalBedrooms, Garage and size as my 5 variables, because this way I can investigate what type of parcels go to which types of places.

I will investigate how the other 4 variables predict size of a parcel

Year: year of delivery, thinking of it like a time fixed effect

house: is a dummy for whether the location being delivere to is a house

TotalBedrooms: this is the number of bedrooms in the location being deliverd to. This will essentially be a proxy for how nice the house/location is and possibly how rich of a neighboorhood the house/location is in.

Garage: a dummy for whether the location has a garage or not, I imagine if a house/locaiton has a garage, then they might be delivering different items then people without garages.

size: the size of the parcel being delivered, essentially capturing the nature of the parcel being delivered

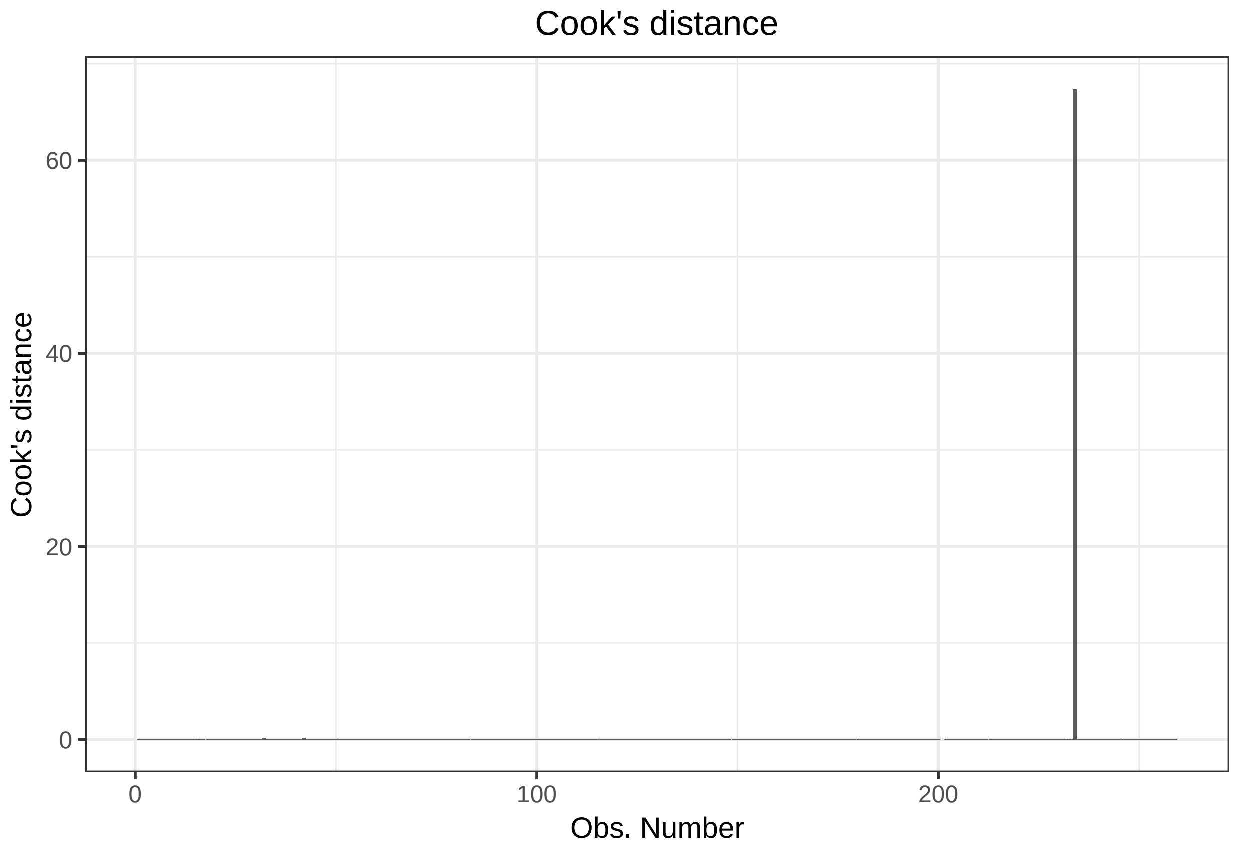
```
#subset
study <- select(filter(x, Year >= 1995), c(Year, house, TotalBedrooms, Garage, size))
#fit
fit1 <- lm(study$size ~ study$Year + study$house + study$TotalBedrooms + study$Garage)
summary(fit1)
```

```
##
## Call:
## lm(formula = study$size ~ study$Year + study$house + study$TotalBedrooms +
##     study$Garage)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62106 -0.09100 -0.01901  0.05995  2.19628
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          24.383589   8.937256   2.728  0.00681 **
## study$Year           -0.012351   0.004472  -2.762  0.00616 **
## study$houseTRUE       0.261547   0.058511   4.470 1.18e-05 ***
## study$TotalBedrooms   0.072108   0.002471  29.186  < 2e-16 ***
## study$Garage          0.071912   0.025196   2.854  0.00467 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.261 on 254 degrees of freedom
## Multiple R-squared:  0.7734, Adjusted R-squared:  0.7698
## F-statistic: 216.7 on 4 and 254 DF,  p-value: < 2.2e-16
```

We see that all of our values are significant at the $\alpha = 0.05$ level and our adjusted R-swuared is 0.7698, which are all good signs that we have a good fit. Additionally our F-statistic is 216.7 which is also a good sign that our fit is pretty good.

d)

```
cook = cooks.distance(fit1)
p4<-ggplot(fit1, aes(seq_along(.cooksd), .cooksd))+geom_bar(stat="identity", position="identity")
  p4<-p4+xlab("Obs. Number")+ylab("Cook's distance")
  p4<-p4+ggtitle("Cook's distance")+theme_bw()
  p4<-p4+ theme(text = element_text(size = 11))
  p4<-p4+  theme(plot.title = element_text(hjust = 0.5))
p4
```



From the graph above it’s clear that there is an one outlier with a huge cook’s distance. We have to remove all points that have a cook’s distance 3 times the mean of the cook’s distance.

```
#Removing Outliers
m_cook = mean(cook)
n_study = study[cook < 3*m_cook,]
#New Fit
fit2 <- lm(n_study$size ~ n_study$Year + n_study$house + n_study$TotalBedrooms + n_study$Garage)
summary(fit2)
```

```
##
## Call:
## lm(formula = n_study$size ~ n_study$Year + n_study$house + n_study$TotalBedrooms +
##     n_study$Garage)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.24535 -0.09091 -0.04161  0.01316  2.16706
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           15.681619   8.393150   1.868 0.062864 .
## n_study$Year          -0.007891   0.004202  -1.878 0.061543 .
## n_study$houseTRUE      0.214046   0.054737   3.910 0.000118 ***
## n_study$TotalBedrooms  0.028123   0.007127   3.946 0.000103 ***
## n_study$Garage         0.039927   0.023870   1.673 0.095625 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

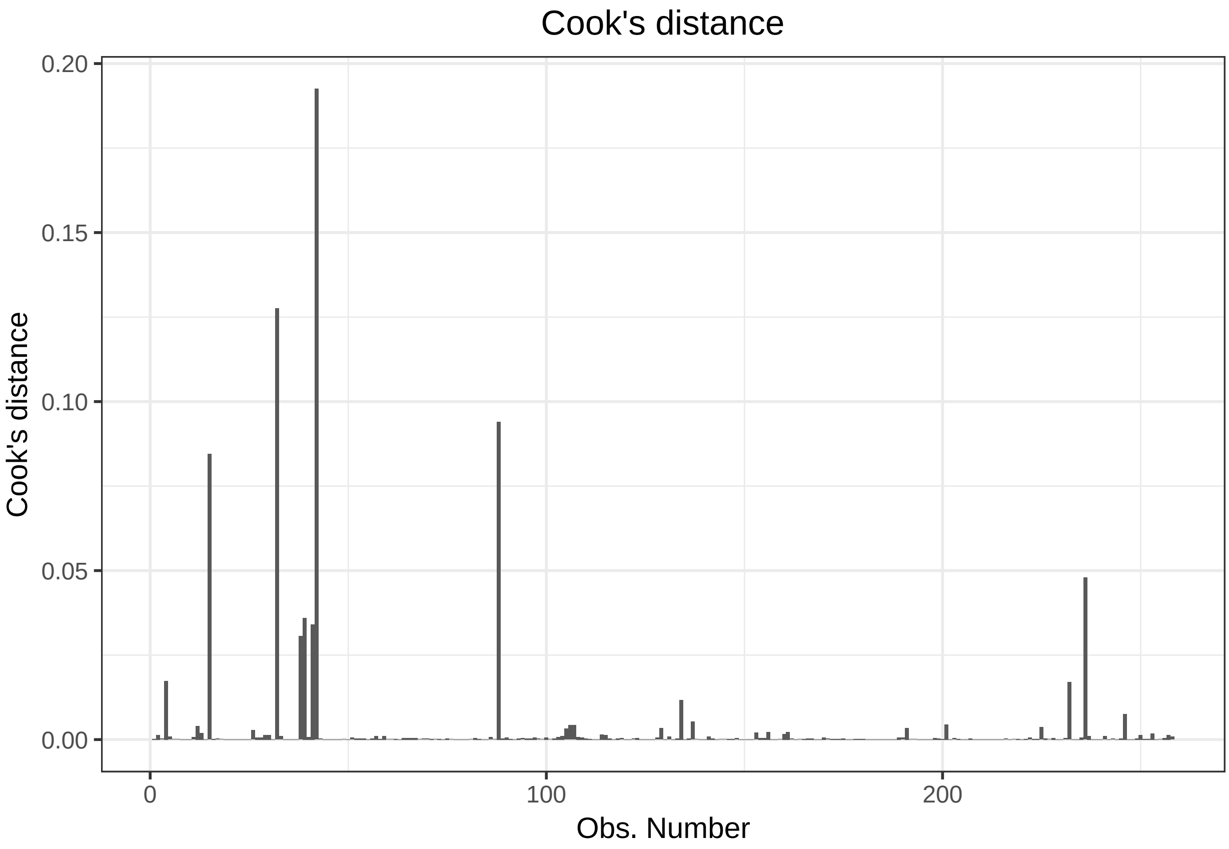
```
##
## Residual standard error: 0.242 on 253 degrees of freedom
## Multiple R-squared:  0.0917, Adjusted R-squared:  0.07734
## F-statistic: 6.386 on 4 and 253 DF,  p-value: 6.55e-05
```

The signficance level of some of our estimates have gone down, but are still significant at the $\alpha = 0.1$. , but our house dummy and our total bedrooms dummy is still significant.

e)

Now we can check whether there are still any outliers and remove them again.

```
cook1 = cooks.distance(fit2)
p4<-ggplot(fit2, aes(seq_along(.cooksd), .cooksd))+geom_bar(stat="identity", position="identity")
  p4<-p4+xlab("Obs. Number")+ylab("Cook's distance")
  p4<-p4+ggtitle("Cook's distance")+theme_bw()
  p4<-p4+ theme(text = element_text(size = 11))
  p4<-p4+  theme(plot.title = element_text(hjust = 0.5))
p4
```



```
m_cook1 = 3*mean(cook1)
m_cook1
```

```
## [1] 0.009437017
```

There seems to be some outliers, we will remove some these again.

```r
n_study1 = n_study[cook1 < m_cook1,]
fit3 <- lm(n_study1$size ~ n_study1$Year + n_study1$house + n_study1$TotalBedrooms + n_study1$Garage)
summary(fit3)
```

```
##
## Call:
## lm(formula = n_study1$size ~ n_study1$Year + n_study1$house +
##     n_study1$TotalBedrooms + n_study1$Garage)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.10107 -0.05155 -0.01317  0.02665  0.34400
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.704039   2.695139   1.003    0.317
## n_study1$Year          -0.001377   0.001349  -1.020    0.309
## n_study1$houseTRUE      0.166398   0.017464   9.528  < 2e-16 ***
## n_study1$TotalBedrooms  0.012387   0.002886   4.292 2.56e-05 ***
## n_study1$Garage         0.033908   0.007437   4.559 8.16e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07442 on 242 degrees of freedom
## Multiple R-squared:   0.33,  Adjusted R-squared:  0.3189
## F-statistic:  29.8 on 4 and 242 DF,  p-value: < 2.2e-16
```

Our fit seems to have changed significantly, however our adjusted R-sqaured has improved quite a bit so it seems as thought we did the right thing.

From our last table summarizing the results of our fit we can say the following things. Firstly it seems as though chosing between 1995 to our time was the right decision because the estimate for Year is not statistically significant. This basically shows that peope's expectation of what they can get delivered has not changed much since 1995 and, possibly, that delivery companies ability to deliver parcels of a certain size has stayed the same across time. As we had originally reasoned the nature of your delivery location, so whether you live in a house, how many bedrooms you have and whether you have a garage or not are significatn indicators to the size of the parcel you may order. This makes sense because all of these are essentially proxies for how rich you are or what type of activities you engage in or how many people live in that location. This are all factors that could influence what type of good you order and therefore influence the size of your parcel. For example if you have a lot of children you may order a lot of toys, which may come in large parcels. Similarly if you have a garage, you may order materials for a DIY project and that could necessitate a large parcel. Alternatively you could not live in a house, and you could just be getting mail like bills. The fact that these variables are significant make sense.