DDM Homework 2

Sai Teja Pasula - MS BAIM - PUID: 0032877594

1/29/2021

Objective: Test for the left-digit bias of car buyers and discuss implications for pricing

Data Processing

```
## coerce model year into a factor variable, use 2006 as the reference level
db$modelyear = factor(db$modelyear)
db$modelyear = relevel(db$modelyear,"2006")

## coerce month into a factor variable, use month 9 as the reference level
db$month = factor(db$month)
db$month = relevel(db$month,"9")
summary(db)
```

```
##       sold            price             mile          engine_vol
##  Min.   :0.0000   Min.   : 8.599   Min.   :  3.10   Min.   :1.700
##  1st Qu.:0.0000   1st Qu.:14.998   1st Qu.: 19.55   1st Qu.:2.400
##  Median :0.0000   Median :15.998   Median : 28.01   Median :2.400
##  Mean   :0.1251   Mean   :16.562   Mean   : 32.53   Mean   :2.438
##  3rd Qu.:0.0000   3rd Qu.:17.998   3rd Qu.: 40.53   3rd Qu.:2.500
##  Max.   :1.0000   Max.   :24.998   Max.   :117.25   Max.   :3.500
##
##    wheelbase        model               month      modelyear
##  Min.   :102.0   Length:975          6      :298   2006: 97
##  1st Qu.:107.0   Class :character   7      :130   2007:367
##  Median :109.0   Mode  :character   5      :118   2008:304
##  Mean   :107.7                      2      :109   2009:156
##  3rd Qu.:109.0                      9      : 94   2010: 51
##  Max.   :110.0                      1      : 83
##                                     (Other):143
```

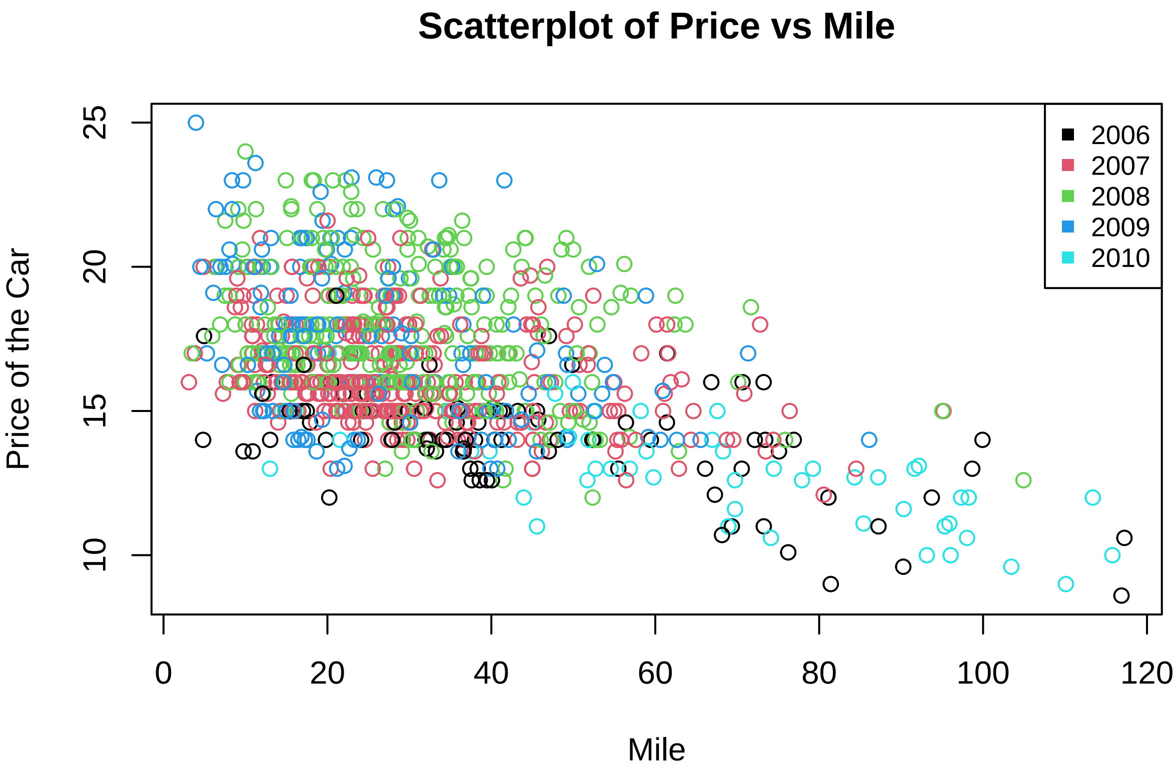
```
## decompose the mile
db$mile10k = floor(db$mile/10)*10
db$mile1k = floor(db$mile - db$mile10k)
db$milermd = db$mile - floor(db$mile)
db$milermd = round(db$milermd,digits = 3)

head(db[,c("mile","mile10k","mile1k","milermd")])
```

```
##     mile mile10k mile1k milermd
## 1 21.057      20      1   0.057
## 2 39.445      30      9   0.445
## 3 45.727      40      5   0.727
## 4 20.251      20      0   0.251
## 5 40.415      40      0   0.415
## 6 50.365      50      0   0.365
```

Question 1: Plot a scatterplot of price against mile. Briey explain the major patterns in the price-mile relationship.

```
## plot price against mile - add legend
plot(db$mile,db$price,main="Scatterplot of Price vs Mile",
     xlab="Mile", ylab="Price of the Car",col = db$modelyear)
legend("topright",legend=c(2006:2010),col=1:5,pch=15,cex=0.8)
```



a) The above plot shows that the price of the car declines with the increase in the number of miles traveled.

b) It also shows that the there are a lot of cars with similar selling price i.e. horizontal lines even though the mileage is different and that’s because mileage is not the only factor affecting the price.

Question 2: Regress price on all car attributes (use decomposed mile) and month. How does the price-mile relationship here compare with that shown in the scatterplot?

```
reg1 = glm(price ~ mile + engine_vol + wheelbase + modelyear + model + month, data = db)
summary(reg1)
```

(i) Linear price regression - with mile

```
##
## Call:
## glm(formula = price ~ mile + engine_vol + wheelbase + modelyear +
##     model + month, data = db)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5950  -0.8845  -0.1548   0.8525   5.2200
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -11.380441   7.840718  -1.451  0.14698
## mile           -0.055388   0.002681 -20.660  < 2e-16 ***
## engine_vol      2.280628   0.126116  18.084  < 2e-16 ***
## wheelbase       0.214451   0.072103   2.974  0.00301 **
## modelyear2007   1.100902   0.160575   6.856 1.27e-11 ***
## modelyear2008   2.183889   0.170964  12.774  < 2e-16 ***
## modelyear2009   2.749893   0.188361  14.599  < 2e-16 ***
## modelyear2010  -0.013414   0.240218  -0.056  0.95548
## modelAltima    -0.890920   0.128976  -6.908 8.99e-12 ***
## modelCamry     -1.146315   0.119600  -9.585  < 2e-16 ***
## modelCivic     -0.026002   0.273040  -0.095  0.92415
## modelCorolla   -0.532251   0.525675  -1.013  0.31155
## modelSonata    -3.287185   0.272207 -12.076  < 2e-16 ***
## month1          0.614090   0.199128   3.084  0.00210 **
## month2          1.342006   0.187721   7.149 1.74e-12 ***
## month3          1.087955   0.234062   4.648 3.82e-06 ***
## month4          0.039009   0.222036   0.176  0.86058
## month5         -0.047729   0.190916  -0.250  0.80264
## month6         -0.071252   0.167869  -0.424  0.67133
## month7         -0.412861   0.182054  -2.268  0.02356 *
## month8         -0.056583   0.275996  -0.205  0.83760
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for gaussian family taken to be 1.731135)
##
##     Null deviance: 6343.1  on 974  degrees of freedom
## Residual deviance: 1651.5  on 954  degrees of freedom
## AIC: 3324.8
##
## Number of Fisher Scoring iterations: 2
```

```
reg2 = glm(price ~ mile10k + mile1k + milermd + engine_vol + model + modelyear  + month, data = db)
summary(reg2)
```

(ii) Linear price regression - with mile replaced by decomposed mile digits

```
##
## Call:
## glm(formula = price ~ mile10k + mile1k + milermd + engine_vol +
##     model + modelyear + month, data = db)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7010  -0.8611  -0.1553   0.8245   5.1999
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   11.862227   0.419176  28.299  < 2e-16 ***
## mile10k       -0.055302   0.002692 -20.547  < 2e-16 ***
## mile1k        -0.065550   0.015090  -4.344 1.55e-05 ***
## milermd        0.154222   0.145726   1.058 0.290184
## engine_vol     2.285679   0.126593  18.055  < 2e-16 ***
## modelAltima   -0.818766   0.127091  -6.442 1.86e-10 ***
## modelCamry    -1.143767   0.120153  -9.519  < 2e-16 ***
## modelCivic    -0.656356   0.171206  -3.834 0.000135 ***
## modelCorolla  -1.997748   0.194948 -10.248  < 2e-16 ***
## modelSonata   -3.720641   0.232408 -16.009  < 2e-16 ***
## modelyear2007  1.052141   0.160938   6.538 1.02e-10 ***
## modelyear2008  2.268664   0.168932  13.429  < 2e-16 ***
## modelyear2009  2.831005   0.187043  15.136  < 2e-16 ***
## modelyear2010 -0.140476   0.238737  -0.588 0.556393
## month1         0.651672   0.200217   3.255 0.001175 **
## month2         1.347402   0.188469   7.149 1.74e-12 ***
## month3         1.049963   0.235786   4.453 9.47e-06 ***
## month4         0.030028   0.223240   0.135 0.893028
## month5        -0.067199   0.191676  -0.351 0.725977
## month6        -0.051488   0.168729  -0.305 0.760316
## month7        -0.378920   0.183120  -2.069 0.038793 *
## month8         0.014435   0.276416   0.052 0.958362
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for gaussian family taken to be 1.744392)
##
##     Null deviance: 6343.1  on 974  degrees of freedom
## Residual deviance: 1662.4  on 953  degrees of freedom
## AIC: 3333.2
##
## Number of Fisher Scoring iterations: 2
```

a) We can infer that the left digits of the decomposed price (mile10k,mile1k) behave in a similar way that of the mile in the first scatter plot. But the last mile digits are positively affecting the price and not significant enough to predict the price while the left ones are significant with almost 100% confidence

Question 3: Fit a logistic regression for whether a car was sold on the first day to investigate the LDB of car buyers. Does car buyers show LDB in their attention to the digits of price? Briefly explain your answer.

```
## decompose the price
db$pricedol10 = floor(db$price/10)*10
db$pricedol1 = floor(db$price - db$pricedol10)
db$pricemd = db$price - floor(db$price)
db$pricemd = round(db$pricemd,digits = 3)

#fit the regression
reg3 = glm(sold ~ pricedol10 + pricedol1 + pricemd + mile + modelyear + model + month, data = db,family=
summary(reg3)
```

```
##
## Call:
## glm(formula = sold ~ pricedol10 + pricedol1 + pricemd + mile +
##     modelyear + model + month, family = binomial, data = db)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2621  -0.5478  -0.4570  -0.3338   2.6101
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     4.275219   1.500008   2.850 0.004370 **
## pricedol10     -0.325389   0.078167  -4.163 3.14e-05 ***
## pricedol1      -0.341494   0.088195  -3.872 0.000108 ***
## pricemd         0.151137   0.432142   0.350 0.726535
## mile           -0.019061   0.007348  -2.594 0.009488 **
## modelyear2007 -0.236699   0.344392  -0.687 0.491897
## modelyear2008 -0.165017   0.391704  -0.421 0.673551
## modelyear2009  0.111373   0.434294   0.256 0.797606
## modelyear2010 -0.158345   0.483600  -0.327 0.743343
## modelAltima   -0.479452   0.299860  -1.599 0.109839
## modelCamry    -0.732183   0.298625  -2.452 0.014213 *
## modelCivic    -1.308283   0.384095  -3.406 0.000659 ***
## modelCorolla  -1.345434   0.452230  -2.975 0.002929 **
## modelSonata   -0.940100   0.544224  -1.727 0.084093 .
## month1         0.636279   0.466129   1.365 0.172244
## month2         0.495265   0.473916   1.045 0.296000
## month3         1.307674   0.499469   2.618 0.008841 **
## month4        -0.241944   0.563715  -0.429 0.667781
## month5        -0.197364   0.462717  -0.427 0.669721
## month6        -0.150968   0.412952  -0.366 0.714677
## month7         0.047761   0.436325   0.109 0.912836
## month8        -1.468331   1.080359  -1.359 0.174110
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 735.19  on 974  degrees of freedom
## Residual deviance: 691.10  on 953  degrees of freedom
## AIC: 735.1
##
## Number of Fisher Scoring iterations: 6
```

a) After observing the results of the above model, we can say that the left digit bias is evident for sold vs price, because the left digits(pricedol10, pricedol1) are significant with almost 100% confidence whereas the right most ones aren't significant enough to predict the selling probability.

Question 4: Briefly discuss the implications of your fndings above for the pricing of used cars.

a) We have concluded that the left digit bias exists when a consumer is trying to purchase a used car. But that left digit bias can be across multiple variables i.e. in the above example, it is price and miles. But both these variables are interdependent with each other. So the store managers can take the left digit bias into account, but focus on the most important variable (here it is price) to apply this left digit bias.
b) LDB is present
c) first two digits are important
d) The store managers can be more proactive while setting a price for the car i.e. they can increase the last digits part of the price for more profit margin because the