Homework 5

Joe Diaz

9/25/2019

Question 8.1

Describe a situation or problem from your job, everyday life, current events, etc., for which a linear regression model would be appropriate. List some (up to 5) predictors that you might use.

A nice use case for a linear regression would be predicting the value of a used car to see if you are getting a good deal or not.

Some features I would use are:

- Make
- Model
- Age
- color
- Location of vendor

Question 8.2

Using crime data from http://www.statsci.org/data/general/uscrime.txt (file uscrime.txt, description at http://www.statsci.org/data/general/uscrime.html), use regression (a useful R function is **lm or glm) to predict the observed crime rate in a city with the following data:**

```
#df <- read.table("uscrime.txt", stringsAsFactors = FALSE, header = TRUE)
df<-read.delim("http://www.statsci.org/data/general/uscrime.txt")


model <- glm(Crime ~ . , data=df, family="gaussian")
summary(model)
```

```
##
## Call:
## glm(formula = Crime ~ ., family = "gaussian", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -395.74   -98.09    -6.69   112.99   512.67
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.984e+03  1.628e+03  -3.675 0.000893 ***
## M            8.783e+01  4.171e+01   2.106 0.043443 *
## So          -3.803e+00  1.488e+02  -0.026 0.979765
## Ed           1.883e+02  6.209e+01   3.033 0.004861 **
## Po1          1.928e+02  1.061e+02   1.817 0.078892 .
## Po2         -1.094e+02  1.175e+02  -0.931 0.358830
## LF          -6.638e+02  1.470e+03  -0.452 0.654654
## M.F          1.741e+01  2.035e+01   0.855 0.398995
## Pop         -7.330e-01  1.290e+00  -0.568 0.573845
## NW           4.204e+00  6.481e+00   0.649 0.521279
```

```
## U1          -5.827e+03  4.210e+03  -1.384 0.176238
## U2           1.678e+02  8.234e+01   2.038 0.050161 .
## Wealth       9.617e-02  1.037e-01   0.928 0.360754
## Ineq         7.067e+01  2.272e+01   3.111 0.003983 **
## Prob        -4.855e+03  2.272e+03  -2.137 0.040627 *
## Time        -3.479e+00  7.165e+00  -0.486 0.630708
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 43707.93)
##
##     Null deviance: 6880928  on 46  degrees of freedom
## Residual deviance: 1354946  on 31  degrees of freedom
## AIC: 650.03
##
## Number of Fisher Scoring iterations: 2
```

Model Validation

I referenced this to perform cross-validation with glm https://stat.ethz.ch/R-manual/R-patched/library/boot/html/cv.glm.html

```
library(boot)
# sum of squared differences
sq_diff <- sum((df$Crime - mean(df$Crime))^2)

cross_val_model <- cv.glm(df,model,K=7)

# R squared
1 - cross_val_model$delta[1]*nrow(df)/sq_diff
```

```
## [1] 0.4536012
```

R-Squared is fairly low when using all the available features!