STOR 565 Fall 2019 Homework 2

Ted Henson

*Remark.* This homework aims to help you go through the necessary preliminary from linear regression. Credits for **Theoretical Part** and **Computational Part** are in total 100 pts. If you receive more points that 100 (say via attempting extra credit/optional questions) then your score will be rounded to 100. **If you are aiming to get full points, it is your duty to make sure you have attempted enough problems to get 100 pts**. For **Computational Part**, please complete your answer in the **RMarkdown** file and summit your printed PDF (or doc or html) homework created by it.

## Computational Part

1. (*21 pt*) Consider the dataset “Boston” in predicting the crime rate at Boston with associated covariates.

head(Boston)

## crim zn indus chas nox rm age dis rad tax ptratio black  
## 1 0.00632 18 2.31 0 0.538 6.575 65.2 4.0900 1 296 15.3 396.90  
## 2 0.02731 0 7.07 0 0.469 6.421 78.9 4.9671 2 242 17.8 396.90  
## 3 0.02729 0 7.07 0 0.469 7.185 61.1 4.9671 2 242 17.8 392.83  
## 4 0.03237 0 2.18 0 0.458 6.998 45.8 6.0622 3 222 18.7 394.63  
## 5 0.06905 0 2.18 0 0.458 7.147 54.2 6.0622 3 222 18.7 396.90  
## 6 0.02985 0 2.18 0 0.458 6.430 58.7 6.0622 3 222 18.7 394.12  
## lstat medv  
## 1 4.98 24.0  
## 2 9.14 21.6  
## 3 4.03 34.7  
## 4 2.94 33.4  
## 5 5.33 36.2  
## 6 5.21 28.7

Suppose you would like to predict the crime rate with explantory variables

* medv - Median value of owner-occupied homes
* dis - Weighted mean of distances to employement centers
* indus - Proportion of non-retail business acres

Run the linear model using the code below. You can do so either by copying and pasting the code into the R console, or by clicking the green arrow in the code ‘chunk’ (grey box where the code is written).

mod1 <- lm(crim ~ medv + dis + indus, data = Boston)  
summary(mod1)

##   
## Call:  
## lm(formula = crim ~ medv + dis + indus, data = Boston)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.625 -3.345 -1.242 1.608 78.994   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11.67738 2.12190 5.503 5.95e-08 \*\*\*  
## medv -0.26061 0.04204 -6.199 1.19e-09 \*\*\*  
## dis -0.96320 0.22758 -4.232 2.75e-05 \*\*\*  
## indus 0.13145 0.07728 1.701 0.0896 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.519 on 502 degrees of freedom  
## Multiple R-squared: 0.2404, Adjusted R-squared: 0.2358   
## F-statistic: 52.95 on 3 and 502 DF, p-value: < 2.2e-16

Answer the following questions.

1. What do the following quantities that appear in the above output mean in the linear model? Provide a breif description.
   * t value and Pr(>|t|) of medv

* **Answer:**
* The t value represents a parameters departure from its null hypothesis value (0) according to the student t distribution. P(>|t|) represents the cdf of the t distribution, and represents the likelihood that parameter is nonzero. In this case, the t value of -6.199 for medv corresponds to a 1.19e-09 chance that this nonzero coeffecient found occured due to random chance.
  + Multiple R-squared
* **Answer:**
* The Multiple R-squared is the percentage of variation explained by the independent variables (1 - (Model Sum of squares / Total Sum of squres)). In this case, our model found explains 24.04% of the total variation in our response variable.
  + F-statistic, DF and corresponding p-value
* **Answer:** The F-statistic corresponds to the F distribution. It’s p value represents the probablility that at least 1 coefficient is nonzero. In this case there is a 2.2e-16 that all of our coefficients are zero (these nonzero coefficients occured due to random chance).

1. Are the following sentences True of False? Briefly justify your answer. + indus is not a significant predictor of crim at the 0.1 level.

\*\*Answer:\*\*   
  
False. The p value for the indus coefficient is less than .1 so it is a significant predictor at the .1 level.  
  
\*\*\*  
+ `Multiple R-squared` is preferred to `Adjusted R-squared` as it takes into account all the variables.  
  
\*\*Answer:\*\*   
  
False. Multiple R-Squared does not consider how many variables you use at all. The adjusted R-squared takes into account that if you had more variables, the improvement in R squared could have happened due to random chance.   
  
\*\*\*   
+ `medv` has a negative effect on the response.  
  
\*\*Answer:\*\*  
  
True. The coefficient for medv in our model is less than zero. Its p value is also less than the .01 (and lower) significance level so it most likely has a negative effect, and is significant.  
  
\*\*\*  
+ Our model residuals appear to be normally distributed.  
  
\begin{hint}  
 You need to access to the model residuals in justifying the last sentence. The following commands might help.  
\end{hint}  
  
```r  
# Obtain the residuals  
res1 <- residuals(mod1)  
  
# Normal QQ-plot of residuals  
plot(mod1, 2)  
```  
  
![](HW-2\_files/figure-docx/unnamed-chunk-1-1.png)<!-- -->  
  
```r  
# Conduct a Normality test via Shapiro-Wilk and Kolmogorov-Smirnov test  
shapiro.test(res1)  
```  
  
```  
##   
## Shapiro-Wilk normality test  
##   
## data: res1  
## W = 0.59766, p-value < 2.2e-16  
```  
  
```r  
ks.test(res1, "pnorm")  
```  
  
```  
##   
## One-sample Kolmogorov-Smirnov test  
##   
## data: res1  
## D = 0.39475, p-value < 2.2e-16  
## alternative hypothesis: two-sided  
```  
  
\*\*Answer:\*\*   
False. There appear to be some large residuals on the upper quantile according to our standardized residual plot. Our Shapiro-Wilk and Kolmogoroz-Smirnoz tests give us extremely low p values that this residual distribution is normal.  
  
\*\*\*

1. (*25 pt*) For this exercise, we will use a dataset with summary information about American colleges and universities in 2013. The following code chunk retrieves it directly from the website, saving you from having to download it. The data is saved in the object called amcoll.

setwd('~/Machine Learning')  
amcoll <- read.csv('College.csv')

Suppose that we are curious about what factors at a university play an important role in the room and board each semester (column Room.Board). Answer the following questions.

1. Based on some research into the area, you believe that the five most important predictors for the room and board amount are

Plot a pairwise scatterplot of these variables along with the room and board cost, and comment on any trends. If you don't know how to plot such a scatterplot, see for example:

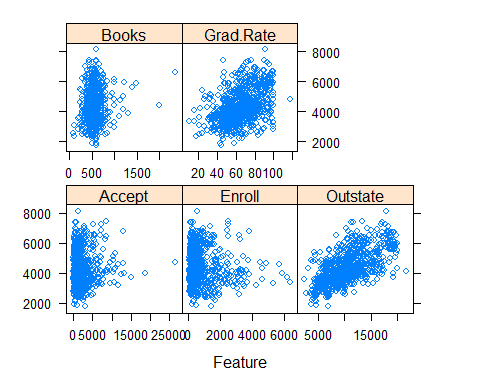
* <http://www.sthda.com/english/wiki/scatter-plot-matrices-r-base-graphs>
* <http://dept.stat.lsa.umich.edu/~jerrick/courses/stat701/notes/ggplot2.html>
* Include your pairwise scatter plot as part of what you turn in.

library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

featurePlot(x=amcoll[,c('Accept', 'Enroll', 'Outstate',  
 'Books', 'Grad.Rate')],  
 y = amcoll[, 'Room.Board'],  
 plot = 'scatter')



Outstate and Grad.Rate appear to have the strongest and most positive relationship with Room.Board. Books, Accept, and Enroll may have a weak relationship.

1. Run a linear model of Room.Board on the 5 features above. Suppose we decide that is our level of significance (so p-values have to be below to count as significant). Discuss the findings of your linear model. In particular you should find that one of the features is **not** significant.

college.mod = lm(Room.Board ~ Books + Grad.Rate + Accept + Enroll + Outstate,  
 data = amcoll)  
summary(college.mod)

##   
## Call:  
## lm(formula = Room.Board ~ Books + Grad.Rate + Accept + Enroll +   
## Outstate, data = amcoll)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2329.8 -544.4 -100.3 496.7 2880.7   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.013e+03 1.532e+02 13.141 < 2e-16 \*\*\*  
## Books 6.458e-01 1.773e-01 3.642 0.000288 \*\*\*  
## Grad.Rate 4.147e+00 2.071e+00 2.003 0.045544 \*   
## Accept 1.409e-01 3.012e-02 4.677 3.43e-06 \*\*\*  
## Enroll -2.905e-01 8.033e-02 -3.616 0.000318 \*\*\*  
## Outstate 1.590e-01 9.135e-03 17.404 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 808.4 on 771 degrees of freedom  
## Multiple R-squared: 0.4601, Adjusted R-squared: 0.4566   
## F-statistic: 131.4 on 5 and 771 DF, p-value: < 2.2e-16

All of our features are below our .01 significance level (the probability that these nonzero coefficients happens due to random chance is close to zero). The Grad.Rate is above our significance level, so we keep our null hypothesis that this coefficient is zero. Our F statistic and corresponding p value below our .01 significance level tell us that at least one of our coefficients is nonzero. Our Multiple R-squared tells us that about 46% of the variation in our response can be explained by this model.

1. Write a function kfold.cv.lm() which performs the following. You can either write this from scracth or use any standard package in R or see the book for example code etc.

* **Input Arguments**:
* - k: integer number of disjoint sets  
  - seed: numeric value to set random number generator seed for reproducability  
  - X: $n \times p$ design matrix  
  - y: $n \times 1$ numeric response  
  - which.betas: $p \times 1$ logical specifying which predictors to be included in a regression

**Output**:

*Avg.MSPE* (average training error over your folds = ),

*Avg.MSE* )

**Description**: Function performs k-fold cross-validation on the linear regression model of on for predictors *which.betas*. Returns both the average MSE of the training data and the average MSPE of the test data.

kfold.cv.lm = function(k = 10, seed = 545, x = amcoll,  
 y = 'Room.Board',  
 which.betas = c('Accept', 'Enroll', 'Outstate',  
 'Books', 'Grad.Rate')){  
   
 cols = c(y, which.betas)  
sub.data = x[,cols]  
  
  
my.formula = as.formula(paste(colnames(cols)[which(colnames(cols) == y)], ' ~ . ', sep = ''))  
  
  
  
sub.data = as.data.frame(sub.data)  
  
  
model = train(  
 as.formula(paste(y, ' ~ .', sep = '')),  
 data = sub.data,  
 method = 'lm',  
 trControl = trainControl(  
 method = "cv",   
 number = k,  
 )  
)  
  
return(model)  
  
}

1. Use your function kfold.cv.lm() to perform 10 folder cross validation on the college data for the following two models:

* the full model on the 5 features above;
* the model where you leave out the feature you found to be insgnificant in (b).

full.model = kfold.cv.lm(k = 10, seed = 545, x = amcoll,  
 y = 'Room.Board',  
 which.betas = c('Accept', 'Enroll', 'Outstate',  
 'Books', 'Grad.Rate'))  
full.model$resample

## RMSE Rsquared MAE Resample  
## 1 831.5576 0.4599087 681.4121 Fold01  
## 2 857.3174 0.4486178 706.4598 Fold02  
## 3 723.3442 0.5345928 586.3384 Fold03  
## 4 900.7207 0.3178661 658.1933 Fold04  
## 5 878.3944 0.4880511 682.6706 Fold05  
## 6 870.7775 0.3879978 698.5737 Fold06  
## 7 845.9284 0.3868709 631.4699 Fold07  
## 8 753.6508 0.5112147 565.3580 Fold08  
## 9 738.9243 0.5125466 575.2803 Fold09  
## 10 712.4757 0.5058883 601.4581 Fold10

reduced.model = kfold.cv.lm(k = 10, seed = 545, x = amcoll,  
 y = 'Room.Board',  
 which.betas = c('Accept', 'Enroll', 'Outstate',  
 'Books'))  
  
reduced.model$resample

## RMSE Rsquared MAE Resample  
## 1 707.9054 0.4926161 575.0894 Fold01  
## 2 932.8296 0.3871647 748.9230 Fold02  
## 3 740.9816 0.5769168 615.4123 Fold03  
## 4 789.5644 0.5230004 606.9923 Fold04  
## 5 805.9746 0.4446741 588.9194 Fold05  
## 6 798.2025 0.5154196 629.6349 Fold06  
## 7 839.3059 0.3860251 667.5502 Fold07  
## 8 865.1133 0.4092752 674.7490 Fold08  
## 9 800.7215 0.4590422 617.8405 Fold09  
## 10 858.2173 0.3703154 665.2368 Fold10

full.model$results

## intercept RMSE Rsquared MAE RMSESD RsquaredSD MAESD  
## 1 TRUE 811.3091 0.4553555 638.7214 71.35711 0.07030792 53.64082

reduced.model$results

## intercept RMSE Rsquared MAE RMSESD RsquaredSD MAESD  
## 1 TRUE 813.8816 0.4564449 639.0348 64.12106 0.06934356 51.17883

Which of the two is a ``better’’ model and why?

**Answer:**

The average RMSE in the folds was lower for our full model than the reduced model. The full model also had a higher average R squared; however, our reduced model had a lower Rsquared standard deviation (between each fold), lower RMAE SD, and lower RMSE SD. We conclude that our reduced model is better due to less variation between folds.

1. (*25 pt*, Textbook Exercises 3.10) This question should be answered using the Carseats data set.

head(Carseats)

## Sales CompPrice Income Advertising Population Price ShelveLoc Age  
## 1 9.50 138 73 11 276 120 Bad 42  
## 2 11.22 111 48 16 260 83 Good 65  
## 3 10.06 113 35 10 269 80 Medium 59  
## 4 7.40 117 100 4 466 97 Medium 55  
## 5 4.15 141 64 3 340 128 Bad 38  
## 6 10.81 124 113 13 501 72 Bad 78  
## Education Urban US  
## 1 17 Yes Yes  
## 2 10 Yes Yes  
## 3 12 Yes Yes  
## 4 14 Yes Yes  
## 5 13 Yes No  
## 6 16 No Yes

1. Fit a multiple regression model to predict Sales using Price, Urban, and US. Then, display a summary of the linear model using the summary function.

mult.fit = lm(Sales ~ Price + Urban + US, data = Carseats)  
summary(mult.fit)

##   
## Call:  
## lm(formula = Sales ~ Price + Urban + US, data = Carseats)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.9206 -1.6220 -0.0564 1.5786 7.0581   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.043469 0.651012 20.036 < 2e-16 \*\*\*  
## Price -0.054459 0.005242 -10.389 < 2e-16 \*\*\*  
## UrbanYes -0.021916 0.271650 -0.081 0.936   
## USYes 1.200573 0.259042 4.635 4.86e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.472 on 396 degrees of freedom  
## Multiple R-squared: 0.2393, Adjusted R-squared: 0.2335   
## F-statistic: 41.52 on 3 and 396 DF, p-value: < 2.2e-16

1. Write a one- or two-sentence interpretation of each coefficient in the model. Be careful: some of the variables in the model are qualitative!

**Answer:**

For every one unit increase in Price, our model decreases our prediction on sales by -0.05446. If Urban is Yes, than our model decreases our prediction on sales by -0.02192. If US is Yes, than our model increases our prediction on sales by 1.20057.

1. Based on the output in part (a): For which of the predictors can you reject the null hypothesis ?

**Answer:** UrbanYes

1. On the basis of your response to the previous question, a model with fewer predictors, using only the predictors for which there is evidence of association with the outcome. Display a summary of the linear model using the summary function.

mult.fit = lm(Sales ~ Price + US, data = Carseats)  
summary(mult.fit)

##   
## Call:  
## lm(formula = Sales ~ Price + US, data = Carseats)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.9269 -1.6286 -0.0574 1.5766 7.0515   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.03079 0.63098 20.652 < 2e-16 \*\*\*  
## Price -0.05448 0.00523 -10.416 < 2e-16 \*\*\*  
## USYes 1.19964 0.25846 4.641 4.71e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.469 on 397 degrees of freedom  
## Multiple R-squared: 0.2393, Adjusted R-squared: 0.2354   
## F-statistic: 62.43 on 2 and 397 DF, p-value: < 2.2e-16

1. In a few sentences: How well do the models in (a) and (d) fit the data? Justify your response with information from the outputs of part (a) and (d).

**Answer:**

Our reduced model has the exact same R-Squared (the percentage of total variation explained by the model is the same). Therefore, there is no reason to add the extra variable ‘Urban’ as it will mostly likely perform worse on future data.

1. (*14 pt Optional*) Note: this question is optinal and if you do want to do it, you will need to do the heavy lifting in terms of finding the data, cleaning the data etc. We will not be able to help you too much with respect to the above data “carpentry” issues.

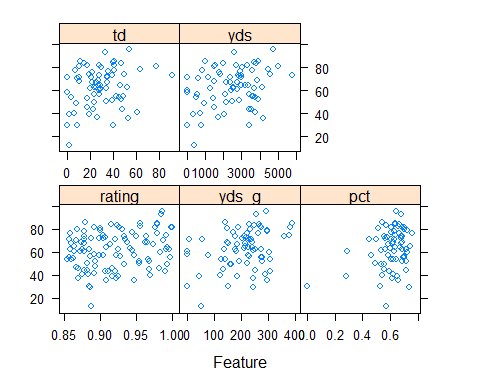
Search online for a dataset that **you are interested in** where you think you can apply linear regression (i.e. your data has a continuous response and a bunch of real valued features). Data sets from the book (ISLR) website are not allowed and more importantly try to find something that makes you curious to find the answers.

1. Include a link and brief description of the data and the kinds of questions you are interested in exploring. ESPN Data: <http://www.espn.com/college-football/qbr> Maxpreps Data: <https://www.maxpreps.com/leaders/football/offense,passing/stat-leaders.htm> 247 Recruiting Data: <https://247sports.com/Season/2020-Football/CompositeRecruitRankings/?InstitutionGroup=HighSchool&PositionGroup=QB>
2. Plot a pairwise scatter plot between the response and some (at least 2) of the features.

library(readr)  
football <- read\_csv("~/Data Journalism/Story Pitch/football.csv")

## Parsed with column specification:  
## cols(  
## playername = col\_character(),  
## total\_qbr = col\_double(),  
## rating = col\_double(),  
## yds\_g = col\_double(),  
## td = col\_double(),  
## interceptions = col\_double(),  
## pct = col\_double(),  
## rate = col\_double(),  
## comp = col\_double(),  
## yds = col\_double(),  
## gp = col\_double()  
## )

featurePlot(x = football[, c('rating', 'yds\_g',  
 'pct',  
 'td', 'yds')],  
 y = football$total\_qbr)



1. Run a linear model to learn the relationship between the features and the response and extract information from the lm function (what variables seem significant and what do not)?

lm.mod = lm(total\_qbr ~ rating + yds\_g + pct + td + yds, data = football)  
summary(lm.mod)

##   
## Call:  
## lm(formula = total\_qbr ~ rating + yds\_g + pct + td + yds, data = football)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -39.514 -10.401 2.494 10.720 30.359   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -9.986378 41.005992 -0.244 0.808  
## rating 52.612208 44.121052 1.192 0.237  
## yds\_g 0.048510 0.053436 0.908 0.367  
## pct 21.941128 22.983159 0.955 0.343  
## td -0.205650 0.297941 -0.690 0.493  
## yds 0.002470 0.005041 0.490 0.626  
##   
## Residual standard error: 16.26 on 64 degrees of freedom  
## (38 observations deleted due to missingness)  
## Multiple R-squared: 0.1556, Adjusted R-squared: 0.08968   
## F-statistic: 2.359 on 5 and 64 DF, p-value: 0.04991

Explain in words (e.g. to someone who has no math or stat background) your findings.

Some of our input variables may have a relationship to Total Quarterback Rating, but the correlations the model found could have occured due to random chance. We cannot definitively conclude that any of these variables are a predictor, but the ESPN rating (rating) is the most likely to have a true relationship to QBR.