```
---
title: "HW6"
author: "STAT 131A"
date: 'Due Date: December 2, 2019'
output:
 pdf_document
header-includes:
- \usepackage{framed}
- \usepackage{xcolor}
- \let\oldquote=\quote
- \let\endoldquote=\endquote
- \colorlet{shadecolor}{orange!15}
- \renewenvironment{quote}{\begin{shaded*}\begin{oldquote}}
{\end{oldquote}\end{shaded*}}
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE,
tidy.opts=list(width.cutoff=60), tidy=TRUE)
```

### Question 1

a (2 points) I have a dataset containing average hourly earnings in dollars (wage) and years of education (educ) for 526 individuals. I fit a simple linear regression equation with wage as response and educ as the explanatory variable. This gave me the following equation:

wage = -0.90485 + 0.54136 * (educ).

Which among the following is the correct interpretation for this equation? Give reasons for your answer.

 (i) For every additional four years of education, the average hourly wage increases by 4*0.54 = 2.16 dollars.

 (ii) For every additional year of education, the average hourly wage increases by 54%.

 (iii) For every 1% increase in the number of years of education, the average hourly wage increases by 0.54%.

> This is a lin-lin model. i. For every additional four years of education, the average hourly wage increases by 4*0.54 = 2.16
```

dollars. The coefficient on education is 0.54136 meaning that for every unit increase in education, wage will go up by 0.54136 units. The variables of wage and education are not in percentage units thus they are not the last two options, therefore they will increase by 4 times of 0.54 for every four years of education.

b (2 points) For the same dataset as in the previous part, I fit a simple linear regression equation with log(wage) as response and educ as the explanatory variable. This gave me the following equation:

log(wage) = 0.583773 + 0.082744 * (educ).

Which among the following is the correct interpretation for this equation? Give reasons for your answer.

(i) For every additional year of education, the average hourly wage increases by 0.0827 dollars.

(ii) For every additional year of education, the average hourly wage increases by 8.27 percent.

(iii) For every additional year of education, the average hourly wage increases by 0.0827 percent.

> This is a log-lin model. ii. For every additional year of education, the average hourly wage increases by 8.27 percent. Now that we have converted wage to log, the units are measured in percentages. The coefficient on education is 0.082744 so for every 1 unit increase in education, wage goes up by 0.0827*100%=8.27%.

c (2 points) I have a dataset on the salaries of the CEOs of 209 firms (variable name is salary) along with the sales of the firm (variable is sales). The dataset is from the year 1990. Salary is in thousands of dollars and Sales is in millions of dollars. I fit a simple linear regression with log(salary) as the response variable and log(sales) as the explanatory variable and this gave me the equation:

log(salary) = 4.822 + 0.25667 * log(sales).

Which among the following is the correct interpretation for this equation? Give reasons for your answer.

(i) For a 1 percent increase in sales, the CEO salary

increases by 0.257 percent on average.

  (ii) For a 1 million dollar increase in firm sales, the CEO salary increases by 25.667 thousand dollars on average.

  (iii) For a 1 million dollar increase in firm sales, the CEO salary increases by 2.57 percent.

> This is a log-log model. i. For a 1 percent increase in sales, the CEO salary increases by 0.257 percent on average. The coefficient on sales is 0.25667 so for every 1% increase in sales, there is a 0.257% increase in salary.

\pagebreak

### Question 2

(15 points) The following is a the output of running `lm` on a subset of the imdb dataset you will work with below (Question 4). The below output above has five missing values which are indicated by XXAXX-XXFXX
Using only the available information in the above summary, fill in the missing values. I give you space below for R code, but this is just for using R as a calculator  -- you can't recreate this lm summary with the data given, because this is done on a random subset of the full dataset.

```{r}
XXAXX=sum((1.079e-05-0)/1.095e-05) #t-value
director_facebook_likes
XXBXX=2*pt(0.985, 558, lower=FALSE) #p-value
director_facebook_likes
XXFXX=pf(0.3251, 13, 558, lower.tail = FALSE) #p-value overall



```

> XXCXX=13 , XXDXX=558, XXAXX=0.985, XXBXX=0.325, XXFXX=0.988


```
summary(lmMoviesSmall2)

Call:
```

```
lm(formula = imdb_score ~ ., data = moviesSmall2)

Residuals:
   Min      1Q  Median      3Q     Max
-3.6138 -0.4630  0.0876  0.5490  1.9408

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                4.425e+01  9.482e+00   4.667 3.84e-06
***
num_critic_for_reviews     2.540e-03  5.055e-04   5.025 6.78e-07
***
duration                   1.111e-02  1.710e-03   6.497 1.81e-10
***
director_facebook_likes    1.079e-05  1.095e-05   XXAXX    XXBXX
actor_3_facebook_likes     9.128e-05  5.226e-05   1.747
0.08126 .
actor_1_facebook_likes     8.848e-05  3.153e-05   2.807  0.00518
**
gross                      1.662e-10  6.693e-10   0.248  0.80399
num_voted_users            3.746e-06  4.309e-07   8.694  < 2e-16
***
cast_total_facebook_likes -7.583e-05  3.127e-05  -2.425  0.01564
*
num_user_for_reviews      -7.565e-04  1.575e-04  -4.804 2.01e-06
***
budget                    -4.223e-09  1.025e-09  -4.122 4.33e-05
***
title_year                -1.973e-02  4.727e-03  -4.175 3.46e-05
***
actor_2_facebook_likes     6.026e-05  3.242e-05   1.859
0.06362 .
movie_facebook_likes      -1.150e-06  2.366e-06  -0.486  0.62723
---
Signif. codes:  0  0.001  0.01  0.05  0.1  1

Residual standard error: 0.8023 on 558 degrees of freedom
Multiple R-squared:  0.4432,  Adjusted R-squared:  0.4302
F-statistic: 34.16 on XXCXX and XXDXX DF,  p-value: XXFXX
```

\pagebreak

### Question 3

Consider the data in `ceodata_num.csv` which consists of 209 firms and has data on the salary of the CEO, sales of the firm, and the firm type. The data is from the year 1990.

```{r}
ceodata<-read.csv("ceodata_num.csv")
ceodata
```

`salary`: Salary of the CEO in thousands of dollars
`sales`: Sales of company in millions of dollars
`FirmType`: the type of company as numeric values 1-4 which correspond to:

* 1=consumer product
* 2=finance
* 3=industry
* 4=utility

a (10 points) Fit a regression in R with `sales` and `FirmType` as a predictor of `salary`. Interpret each of the coefficient estimates given by R, except for the intercept.

```{r}
FirmType.f=factor(ceodata$FirmType)
summary(lm(salary~sales+FirmType.f, data=ceodata))

```

> For every one million dollar increase on sales, there is a 12.49 dollar increase on average salary holding all other variables constant. For finance type firms, for every 1 unit increase, there is a 349,571.42 dollar decrease in average sales relative to the consumer product firm holding all other variables constant. For industry type firms, for every 1 unit increase, there is a 586,686.42 dollar decrease in average sales relative to the consumer product firm holding all other variables constant. For utility type firms, for every 1 unit increase, there is a 939,552.51 dollar decrease in average sales relative to the consumer product firm holding all other variables constant. Only the industry and utility firm type coefficient is statistically significant since its p-value is less than 0.5.

b (10 points) Fit a regression that allows for a different slope

for the different types of firms. Give a summary of the results and interpret each coefficient, except for the intercept.

```{r}
summary(lm(salary~sales*FirmType.f, data=ceodata))

```

> For every one million dollar increase on sales, there is a 1.715 dollar increase on average salary holding all other variables constant. For finance type firms, for every 1 unit increase, there is a 616,400 dollar decrease in average sales relative to the consumer product firm holding all other variables constant. For industry type firms, for every 1 unit increase, there is a 806,000 dollar decrease in average sales relative to the consumer product firm holding all other variables constant. For utility type firms, for every 1 unit increase, there is a 1,375,000 dollar decrease in average sales relative to the consumer product firm holding all other variables constant. The product terms means that three separate slopes per each level of the factor are being fit. Only the industry and utility firm type coefficient is statistically significant since its p-value is less than 0.5.

c. (10 points) Evaluate whether this model in part b is an improvement over the model in part a.

```{r}
rega=lm(salary~sales+FirmType.f, data=ceodata)
regb =lm(salary~sales*FirmType.f, data=ceodata)
anova(rega, regb)
aic1=AIC(rega)
aic2=AIC(regb)
```

> The model in part b is not an improvement over the model in part a because the AIC is 3611.155 which is larger than the AIC of the part a model at 3608.476.

d. (20 points) Run diagnostics on the model you found in part a, and determine whether there are any problems with this model that should be addressed. If so, explain next steps you might take to try to improve this model.

```{r}
```

```
# Code for diagnostics
par(mfrow = c(2, 3))
plot(lm(salary~sales+FirmType.f, data=ceodata), which=1:6)
```

> There appears to be problems with the model. In the residual vs fitted plot, the points are clustering, however this may not be an issue since there are different firm types having to be accounted for. However, The QQ plot appears to be not normal since the right side of residuals are not aligned. Looking at the scale-location plot there appears to be heteroskedasticity as the residual points are in a cone shape. With Cook's distance there are a few outliers which can be effecting the fit of the data. Some of the model assumptions are violated and therefore there are issues with the data. Going forward we should remove the outliers to perhaps get a better overall view.

\pagebreak

### Question 4

Consider data from the website www.imdb.com giving scores of movies. Read in the data with the following code:

```{r}
movies<-read.csv("imdb_simplified.csv",header=TRUE)
movies<-movies[,-grep("name",names(movies))]
movies<-movies[,!names(movies) %in% "movie_title"] ## LINE3
```

a. (5 points) In the code above, explain what LINE 2 and LINE3 are doing (see `?grep`).

> Line 2 the grep function takes all the character vectors from the names of the dataset movies and removes any columns that have "name" in their name.
Line 3 will subset movies and if column names of the movie dataset are not "movie_title" they will be FALSE if it is TRUE it will be removed. Thus, the column "movie_title" is removed.

b. (10 points) Find best submodel based on AIC,*without including the categorical variables (`genre` or `content_rating`)*. (Hint: If you use `regsubsets`, that function requires you to set `nvmax` as the maximum size submodel you want to consider, and thus have to set it larger than the largest possible model if you

want to compare all possible sizes)

```{r}
# The best submodel based on AIC
newmovies <- movies[ -c(10,16) ]
require(leaps)
regsub_out = regsubsets(imdb_score~.,newmovies, nvmax=14)
LOOCV<-function(lm){
   vals<-residuals(lm)/(1-lm.influence(lm)$hat)
   sum(vals^2)/length(vals)
}

calculateCriterion<-function(x=NULL,y,dataset,lmObj=NULL){
   #dataset contains only explanatory variables
   #x is a vector of logicals, length equal to number of
explanatory variables in dataset, telling us which variables to
keep
   #sigma2 is estimate of model on full dataset
   # either x or lmObj must be given to specify the smaller lm
model
   sigma2=summary(lm(y~.,data=dataset))$sigma^2
   if(is.null(lmObj)) lmObj<-lm(y ~ .,
data=dataset[,x,drop=FALSE]) #don't include intercept
   sumlmObj<-summary(lmObj)
   n<-nrow(dataset)
   p<-sum(x)
   RSS<-sumlmObj$sigma^2*(n-p-1)
   c(R2=sumlmObj$r.squared,
       R2adj=sumlmObj$adj.r.squared,
       "RSS/n"=RSS/n,
       LOOCV=LOOCV(lmObj),
       Cp=RSS/n+2*sigma2*(p+1)/n,
       CpAlt=RSS/sigma2-n+2*(p+1),
       AIC=AIC(lmObj), # n*log(RSS/n)+2*p +constant,
       BIC=BIC(lmObj) # n*log(RSS/n)+p*log(n) + constant
   )
}

critSeat<-apply(summary(regsub_out)$which[,-
1],1,calculateCriterion,
   y=movies$imdb_score,
   dataset=newmovies[-13])
critSeat<-t(critSeat)
critSeat[,7]
```

```

> The best model is size 11 since it has the lowest AIC at 6503.695.

c. (10 points) Use CV to compare the best models of each possible size K, as found by comparing RSS.

```{r}
# The best submodel based on CV
critSeat[,4]
```

> The best model is model size 11 which has the lowest LOOCV value at 0.6498450.

d. (5 points) Plot the AIC and CV as a function of model size, and comment on whether AIC and CV lead to the same answer.

```{r}
# plots of aic and cv vs. model size
par(mfrow = c(1, 2))
AIC=critSeat[,7]
CV=critSeat[,4]
size=c(1:13)
plot(size,CV)
plot(size,AIC)

```

> Looking at the plots of CV as a function of size and AIC as a function of size they end up looking very similar to one another so they should end up giving the same results which looking at the part 4a and 4b they both signify that model 11 is the best model.

e. (10 points) If you had far more variables, you would not be able to find the best among all submodels, and could use stepwise regression. Use the `step` function to find a good submodel (based on AIC). Does it find best model based on AIC? If not, report the AIC of the model `step` does find.

```{r}

```
# applying step
model=lm(imdb_score~.,newmovies)
stepwise<-step(model)
stepwise
```

> The step function gives the model lm(formula = imdb_score ~ num_critic_for_reviews + duration + actor_3_facebook_likes + actor_1_facebook_likes + num_voted_users + cast_total_facebook_likes + num_user_for_reviews + budget + title_year + actor_2_facebook_likes + movie_facebook_likes,data = newmovies) with the AIC of -1191.79.The step function does find the best model. The step function iteratively removes the least necessary variables to get the best submodel with the lowest AIC. It is an 11 variable model which is the same as with the AIC from part 4a and LOOCV part 4b.

f. (5 points) In the help of `regsubsets` it states that it will not run if there are more than 50 variables (because it will take too long to try all of the submodels).

If I try to run `regsubsets` on the 15 explanatory variables in the dataset `movie`, i.e. include the categorical variables `genre` and `content_rating`,  it says that it has reached that limit:

```
> regsubsets(imdb_score~., movies,nvmax=30)
Error in leaps.exhaustive(a, really.big) :
 Exhaustive search will be S L O W, must specify really.big=T
```

Given an explanation for how `regsubsets` determined there are more than 50 variables.

> Regsubsets cannot run on more than 50 variables unless really.big is set to TRUE. However, in this case it will determine that there are more than 50 variables although there are actually less variables because the categorical variables are included. It would have worked with categorical variables however, under genre and content_rating there are many different types of each which make over 50 variables including the original 15 when creating the regression subsets.