Introduction to Machine Learning
Solution for Unit Exercises

2 Exercises

2.1 Consider a pair of quantitative variables (X, Y) with a joint PDF given by

$$\begin{aligned} \pi(x, y) &= 2 \quad \text{if } x \geq 0, y \geq 0 \text{ and } x + y \leq 1, \text{ and} \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Suppose we observe $X = x = 0.2$ and would like to predict the corresponding Y.

(a) Under the MSE criteria, what is the expression for the best regressor of Y given $X = x$?

(b) Now suppose that $\pi(x, y)$ is unknown to you but we have a data set consisting of iid samples $\{(x_i, y_i),\ i = 1, 2, \cdots, N\}$ from $\pi(x, y)$. Based on the expression derived in (a), determine a class of functions $\mathcal{C}$ that will be optimal in estimating the best regressor. Find the estimate of the best regressor based on $\mathcal{C}$ and the data set $\{(x_i, y_i),\ i = 1, 2, \cdots, N\}$.

(c) Determine a class of functions that is more restrictive compared to $\mathcal{C}$, and another class of functions that is more flexible compared to $\mathcal{C}$.

Assignment Project Exam Help
https://powcoder.com
Add Wechat powcoder

SOLUTION:

(a) Under the MSE criteria, we wish to find the predictor $f(X)$ which is a function of X that minimizes

$$MSE = E(Y - f(X))^2$$

where the expectation is taken with respect to the joint PDF of (X, Y), $\pi(x, y)$. From lectures, the best regressor turns out to be $\hat{f}(x) = E(Y|X = x)$ which is the conditional expectation of Y given $X = x$ with respect to the conditional PDF of Y given $X = x$. The conditional PDF $\pi(y|x)$ is given by the formula $\pi(y|x) = \pi(x, y)/\pi(x)$ where $\pi(x)$ is the marginal of X evaluated at $X = x$ given by

$$\pi(x) = \int_{y=0}^{1-x} \pi(x, y)\, dy = \int_{y=0}^{1-x} 2\, dy = 2(1 - x).$$

It follows that the conditional PDF $\pi(y|x)$ is

$$\begin{aligned} \pi(y|x) = \frac{2}{2(1-x)} = \frac{1}{1-x} \quad &\text{if } 0 \leq y \leq 1 - x, \text{ and} \\ = 0 \quad &\text{otherwise.} \end{aligned}$$

Therefore,

$$E(Y|X=x) = \int_0^{1-x} y\,\pi(y|x)\,dy = \int_0^{1-x} y\,\frac{1}{1-x}\,dy = \frac{1}{1-x}\left[\frac{y^2}{2}\right]_{y=0}^{1-x} = \frac{1-x}{2},$$

a linear function in x.

(b) Based on the form of $E(Y|X=x)$, as a function of x, the optimal class to be considered is the class of all linear functions $\mathcal{C} = \{f(x) : f(x) = \beta_0 + \beta_1 x\}$. The objective function to be minimized based on the training sample is the empirical MSE given by

$$MSE(\beta_0, \beta_1) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_i)^2$$

with respect to β_0 and β_1. The standard least squares estimates of β_0 and β_1 results from the minimization.

(c) A more restrictive class could be the class of all constant functions whereas a more flexible class could be the class of all quadratic polynomials in x.

2.2 The `Advertising` dataset consists of the `sales` of a product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: `TV`, `radio`, and `newspaper`.

Assignment Project Exam Help

(a) Obtain scatterplots of `sales` versus each of the three different media. What do you observe regarding the general trend in these scatterplots?

https://powcoder.com

(b) Find the least squares regression line for each scatterplot and obtain summaries of the fit.

Add WeChat powcoder

(c) Is simple linear regression an adequate class of models for explaining the data in the scatterplots? Provide relevant diagnostics.

(d) Provide advice with suitable quantitative evidence on what could be the best media to allocate funds to for increasing sales.

SOLUTION:

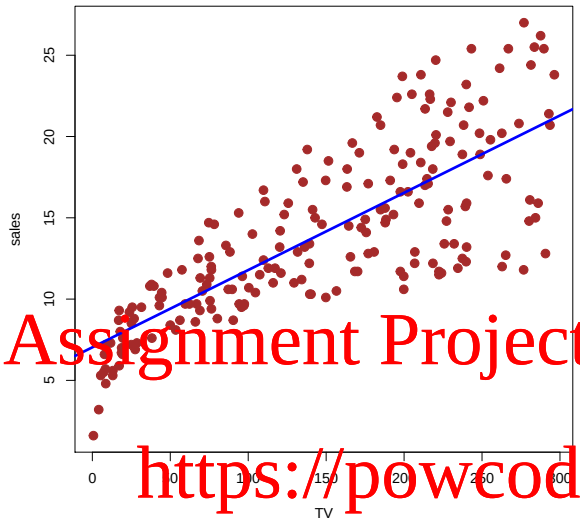
(a) The R codes and resulting scatter plots are given below:

```
#2.2
setwd("C:\\Users\\Dell\\OneDrive\\Documents\\Teaching\\202021\\F70TS"
#Set the correct working directory in R
advert <- read.csv("Advertising.csv")
#Read the csv file
str(advert)
```

```
## 'data.frame': 200 obs. of  5 variables:
##  $ X        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ TV       : num  230.1 44.5 17.2 151.5 180.8 ...
##  $ radio    : num  37.8 39.3 45.9 41.3 10.8 48.9 32.8 19.6 2.1 2
##  $ newspaper: num  69.2 45.1 69.3 58.5 58.4 75 23.5 11.6 1 21.2
##  $ sales    : num  22.1 10.4 9.3 18.5 12.9 7.2 11.8 13.2 4.8 10.
```

```r
#Investigate the dataframe

#Least squares fits
TV_fit <- lm(sales ~ TV, data = advert)
with(advert, {
plot(TV, sales, type = "p", pch= 20, cex=2, col="brown")
abline(TV_fit, lwd=3, col="blue")
}
)
```



Assignment Project Exam Help

https://powcoder.com

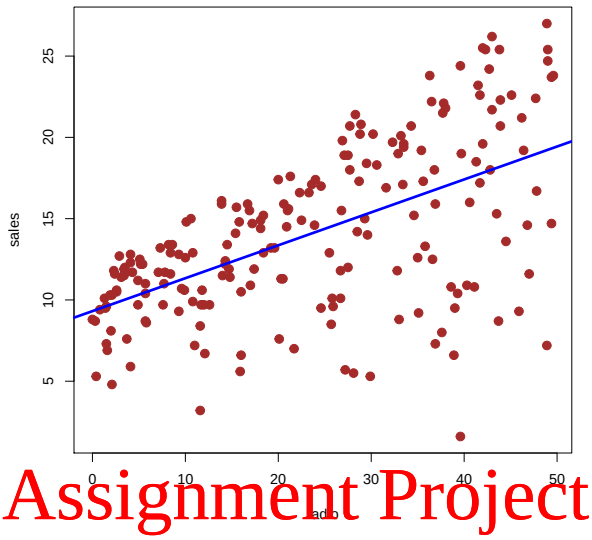
Add WeChat powcoder

```r
summary(TV_fit)
```

```
##
## Call:
## lm(formula = sales ~ TV, data = advert)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.032594   0.457843   15.36   <2e-16 ***
## TV          0.047537   0.002691   17.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 198 degrees of freedom
## Multiple R-squared:  0.6119,Adjusted R-squared:  0.6099
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

```r
#Least squares fits
radio_fit <- lm(sales ~ radio, data = advert)
```

```r
with(advert, {
plot(radio, sales, type = "p", pch= 20, cex=2, col="brown")
abline(radio_fit, lwd=3, col="blue")
}
)
```



Assignment Project Exam Help

```r
summary(radio_fit)
```

https://powcoder.com

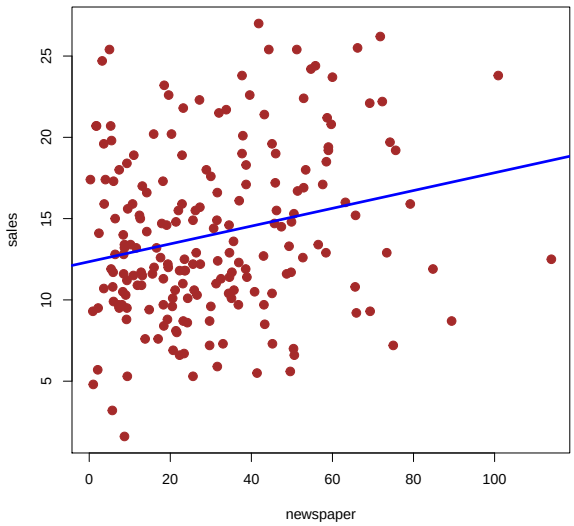
```
##
## Call:
## lm(formula = sales ~ radio, data = advert)
##
```

Add WeChat powcoder

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.7305  -2.1324   0.7707   2.7775   8.1810
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.31164    0.56290  16.542   <2e-16 ***
## radio        0.20250    0.02041   9.921   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.275 on 198 degrees of freedom
## Multiple R-squared:  0.332,Adjusted R-squared:  0.3287
## F-statistic: 98.42 on 1 and 198 DF,  p-value: < 2.2e-16
```

```r
#Least squares fits
newspaper_fit <- lm(sales ~ newspaper, data = advert)
with(advert, {
plot(newspaper, sales, type = "p", pch= 20, cex=2, col="brown")
abline(newspaper_fit, lwd=3, col="blue")
}
```

```
)
```




```
summary(newspaper_fit)
```

Assignment Project Exam Help

```
##
## Call:
## lm(formula = sales ~ newspaper, data = advert)
##
```

https://powcoder.com

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2272  -3.3873  -0.8392   3.5059  12.7751
##
```

Add WeChat powcoder

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.35141    0.62142   19.88  < 2e-16 ***
## newspaper    0.05469    0.01658    3.30  0.00115 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.092 on 198 degrees of freedom
## Multiple R-squared:  0.05212,Adjusted R-squared:  0.04733
## F-statistic: 10.89 on 1 and 198 DF,  p-value: 0.001148
```

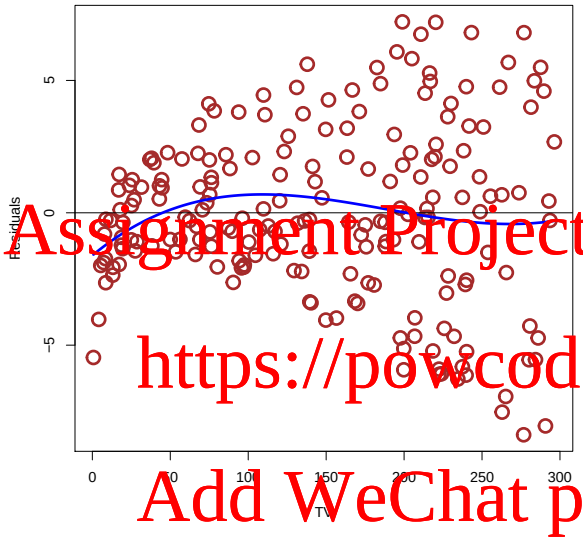
(b) Obtained as above.

(c) Simple linear regression is not adequate because the largest multiple R^2 value is only around 61% for TV; for radio and newspaper, the multiple R^2 values are only 33% and 5%, respectively, indicating poor fit. The lack of fit is also very clear from the scatterplots with the fitted least squares regression lines.

Residual diagnostics are provided for TV and radio based on the R codes and figures:

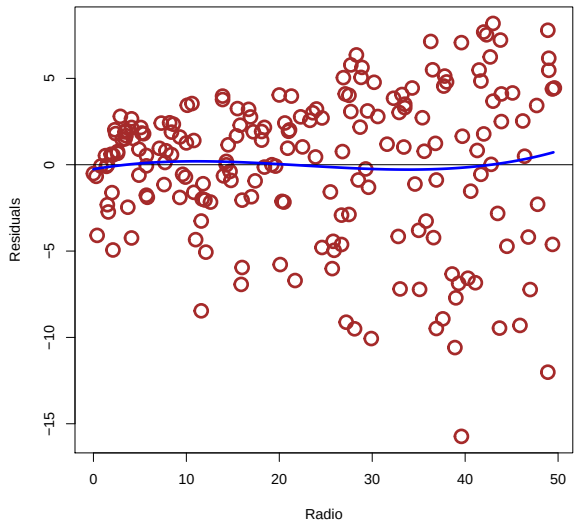
```
#Residual diagnostics for TV fit
resid_TV <- residuals(TV_fit)
plot(advert$TV, resid_TV,
```

```
type = "p", col = "brown",
xlab = "TV", ylab="Residuals", cex = 2, lwd=3)
abline(0,0)
library(splines)
#Obtain data driven trend of resids
resid_df <- data.frame(TV = advert$TV, resids = resid_TV)
#Same lm function works for fitting
resid_fit <- lm(resids ~ bs(TV, df=3), data = resid_df)
#Get predictions
xpoints <- with(resid_df, seq(min(TV), max(TV),0.5))
ypoints = predict(resid_fit, data.frame(TV=xpoints))
lines(xpoints, ypoints, col = "blue", lwd=3)
```



Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

```
#Residual diagnostics for radio fit
resid_radio <- residuals(radio_fit)
plot(advert$radio, resid_radio,
type = "p", col = "brown",
xlab = "Radio", ylab="Residuals", cex = 2, lwd=3)
abline(0,0)
library(splines)
#Obtain data driven trend of resids
resid_df <- data.frame(radio = advert$radio, resids = resid_radio)
#Same lm function works for fitting
resid_fit <- lm(resids ~ bs(radio, df=3), data = resid_df)
#Get predictions
xpoints <- with(resid_df, seq(min(radio), max(radio),0.5))
ypoints = predict(resid_fit, data.frame(radio=xpoints))
lines(xpoints, ypoints, col = "blue", lwd=3)
```



Although the non-parametric mean trends using splines are close to the $y = 0$ line, the variability bands of both residual plots are not uniform. There is significantly larger variability in the residuals for larger funding amounts for both radio and TV. Hence, the model assumption of constant error variance may not be valid.

Assignment Project Exam Help

(d) The best media to invest funds seems to the radio since the estimate of the slope (increase in sales per unit increase in funding) is the largest among all three media. But one must be cautioned that the model fit is poor and the variability is high around the fitted straight line. In other words, if funding is increased, it may turn out that the actual observed sales is less than the anticipated amount (predicted by the regression line) since there is high variability around the predicted value.

https://powcoder.com

Add WeChat powcoder

2.3 On the flexibility of models and the bias-variance decomposition.

(a) Sketch of typical (squared) bias, variance, training error and test error on a single plot, as we increase the flexibility of the class of models used to fit the data. The x-axis should represent increasing degree of flexibility of the model class. There should be four curves. Make sure to label each one.

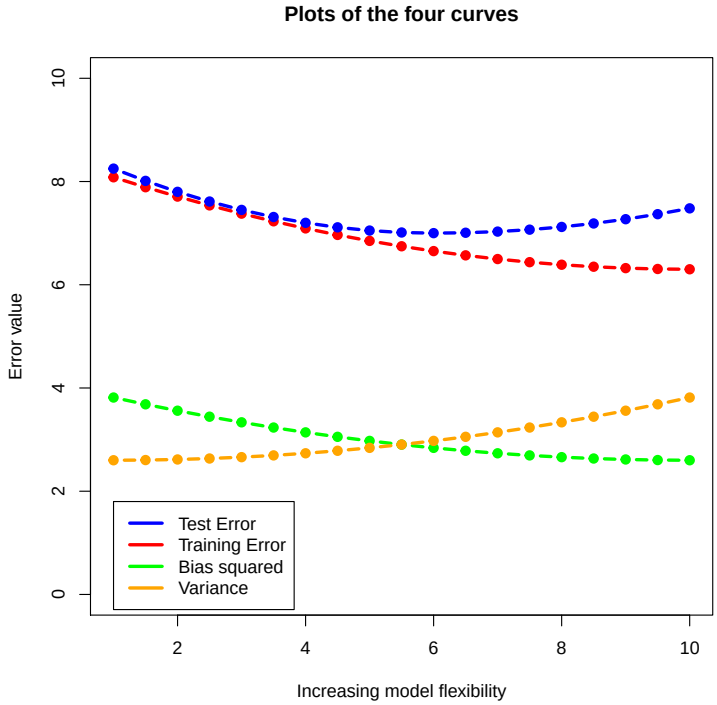
(b) Explain why each curve has the shape that you have drawn.

(c) What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

SOLUTION:

(a) The plot is given as in Figure 1

(b) The training error rate will decrease since the y data in the training data set can be matched exactly with increasing flexibility of the class of models (overfitting).The test error rate will first decrease since the y data is being accurately predicted by the models initially. However, after a certain point, the models will start overfitting and will not provide good predictions on unseen ys as in the test data set. Hence, the test error rate will start to increase. The squared bias will decrease because greater flexibility means that $E(Y|X)$ can be better approximated, on the average, using a more flexible class of



Assignment Project Exam Help

Figure 1: Four different error curves as a function of model flexibility.

models. Variance will increase because more flexible models will result in overfitting on the training data set which will yield high variability for unseen ys.

https://powcoder.com

(c) When the goal is the predict the response variable, one can use a more flexible model. The predictor will be treated as a black box in this case. A more flexible model will provide a better estimate of the best regressor $E(Y|X)$, and this is its advantage. When the goal is to do inference, e.g. to study the relationships between the variables, then a simpler model will be better for interpretating the parameters with regard to the aims of the inference.

Add WeChat powcoder

2.4 Consider the `Auto` data set from the `ISRL` package.

(a) How many rows are in this data set? How many columns? What do the rows and columns represent?

(b) Obtain a scatter plot of `horsepower` versus `mpg` and comment on the trend.

(c) Obtain the least squares regression line of `horsepower` on `mpg` and plot it on the scatterplot. Comment on the fit visually and based on residual diagnostics.

(d) Use a cross validation procedure to find the best regressor of `horsepower` on `mpg` based on a class of models $\mathcal{C}_p$ where

$$\mathcal{C}_p = \left\{ f(x) \ : \ f(x) = \beta_0 + \beta_1\, x + \beta_2\, x^2 + \cdots \beta_p\, x^p \right\}$$

with range of p in $p_1 \leq p \leq p_2$ where p_1 and p_2 are chosen appropriately by you (HINT: Test/Validation error rate must be a U shaped curve for you to choose the minimum).

SOLUTION:

(a) The number of rows is 392 and the number of columns is 9. The R codes are as below. Each row represents an observation on a vehicle and the columns represent the measurements (observations) made on that vehicle. The R codes are as below:

```
#2.4
library(ISLR)
attach(Auto)
str(Auto)
```

```
## 'data.frame': 392 obs. of  9 variables:
##  $ mpg          : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ cylinders    : num  8 8 8 8 8 8 8 8 8 8 ...
##  $ displacement: num  307 350 318 304 302 429 454 440 455 390
##  $ horsepower   : num  130 165 150 150 140 198 220 215 225 190
##  $ weight       : num  3504 3693 3436 3433 3449 ...
##  $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ year         : num  70 70 70 70 70 70 70 70 70 70 ...
##  $ origin       : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ name         : Factor w/ 304 levels "amc ambassador brougham
```

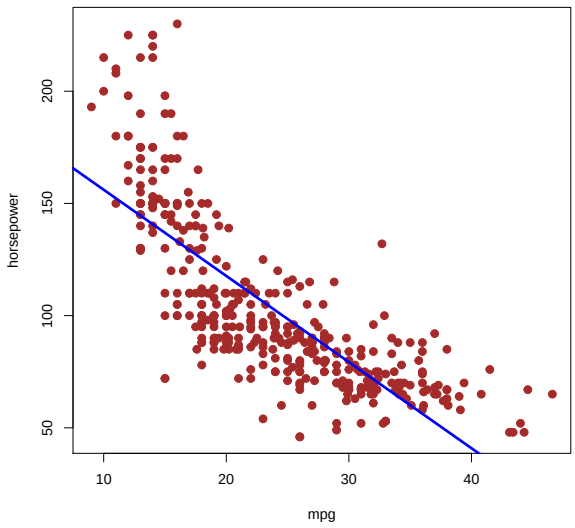
Assignment Project Exam Help

(b) The R codes and figures are given below. The scatter plot shows a decreasing nonlinear trend.

https://powcoder.com

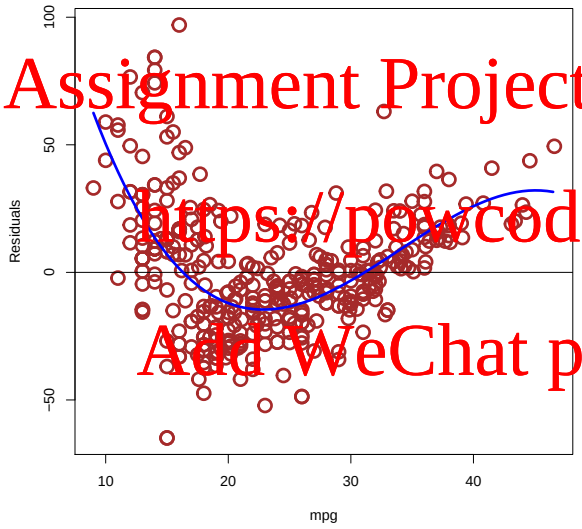
```
with(Auto)
plot(mpg, horsepower, type="p", pch=20, cex=1.8, col="brown")
)
lm_fit <- lm(horsepower ~ mpg, data = Auto)
abline(lm_fit, col="blue", lwd=3)
```

Add WeChat powcoder



(c) Based on the scatterplot above and the fitted line, we can conclude that simple linear regression is not a good fit to the scatterplot points. Residual diagnostics can be carried out to further confirm this fact. The R codes and figure are as follows:

```r
#Residual diagnostics for AUto fit
resid_auto <- residuals(lm_fit)
plot(Auto$mpg, resid_auto,
type = "p", col = "brown",
xlab = "mpg", ylab="Residuals", cex = 2, lwd=3)
abline(0,0)
library(splines)
#Obtain data driven trend of resids
resid_df <- data.frame(mpg = Auto$mpg, resids = resid_auto)
#Same lm function works for fitting
resid_fit <- lm(resids ~ bs(mpg, df=3), data = resid_df)
#Get predictions
xpoints <- with(resid_df, seq(min(mpg), max(mpg),0.5))
ypoints = predict(resid_fit, data.frame(mpg=xpoints))
lines(xpoints, ypoints, col = "blue", lwd=3)
```



Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

The non-parametric fit based on splines indicate that the mean trend of the residuals signficantly deviates from 0 which further gives evidence to the fact that simple linear regression does not provide a good fit to the points in the scatterplot.

(d) The range of p in $p_1 \leq p \leq p_2$ should be chosen so that the test/validation error rate shows a U-shape. The relevant R codes and figure are as follows:

```r
#Now let's do full CV
library(dplyr) #For the pipe operator
```

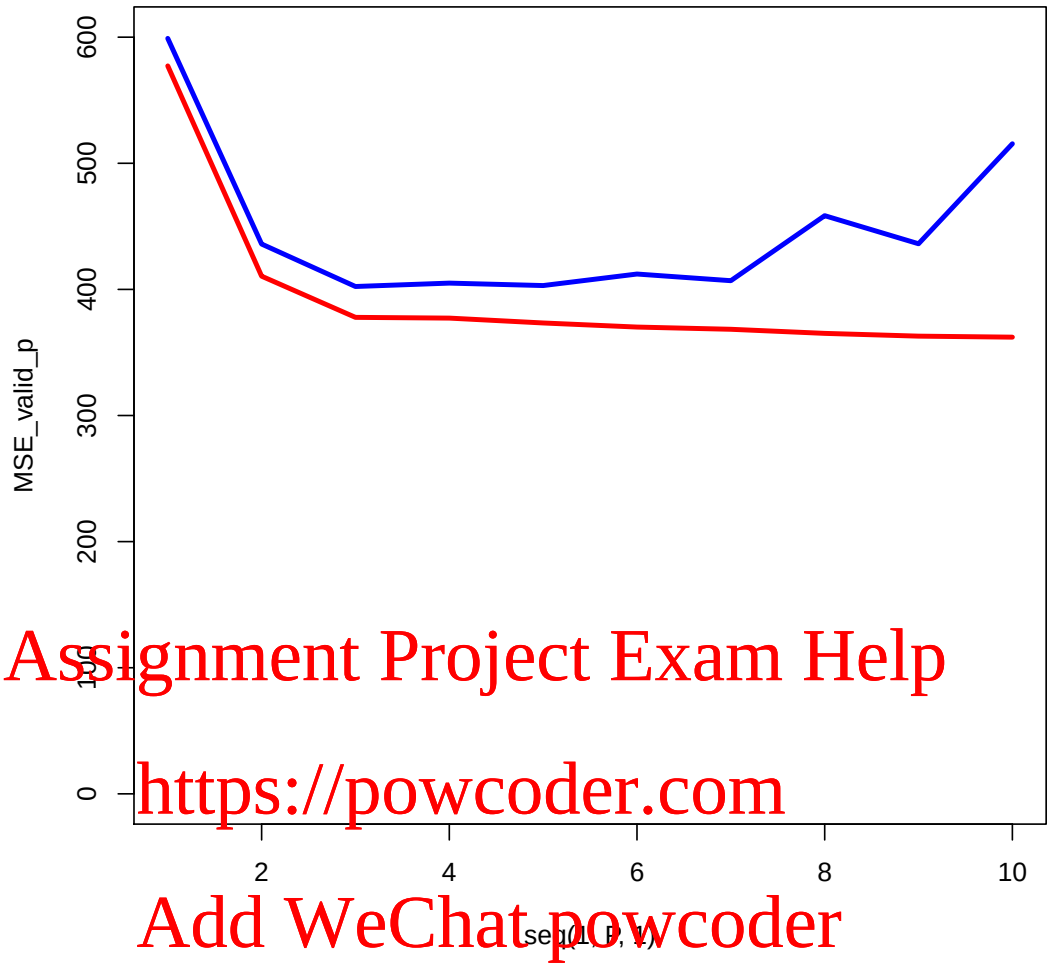
```
##
## Attaching package:  'dplyr'
## The following objects are masked from 'package:stats':
##
##    filter, lag
## The following objects are masked from 'package:base':
```

```
##
##    intersect, setdiff, setequal, union

  K = 50;
  P = 10;
  MSE_train_mat <- rep(list(vector("numeric",K)),P)
  MSE_valid_mat <- rep(list(vector("numeric",K)),P)

  for (k in 1:K){
  #Training dataset data.frame
  train <- Auto %>% sample_frac(0.7)
  #Validation dataset data.frame
  valid <- Auto %>% setdiff(train)
  #Determine class of learners which are polynomails
  #from degree 1 to 6
  for (p in 1:P){
  poly_train_fit <- lm(horsepower ~ poly(mpg,p), data = train)
  poly_train_predict <- predict(poly_train_fit, train)
  poly_valid_predict <- predict(poly_train_fit, valid)
  MSE_train_mat[[p]][k] <- with(train,
   mean((horsepower - poly_train_predict)^2))
  MSE_valid_mat[[p]][k] <- with(valid,
   mean((horsepower - poly_valid_predict)^2))
  }
  }
  MSE_train_p <- sapply(MSE_train_mat, mean)
  MSE_valid_p <- sapply(MSE_valid_mat, mean)
  plot(seq(1,P,1), MSE_valid_p, type="l",col="blue",
  ylim = c(0,600), lwd=3)
  lines(seq(1,P,1), MSE_train_p, type="l",col="red", lwd=3)
```

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

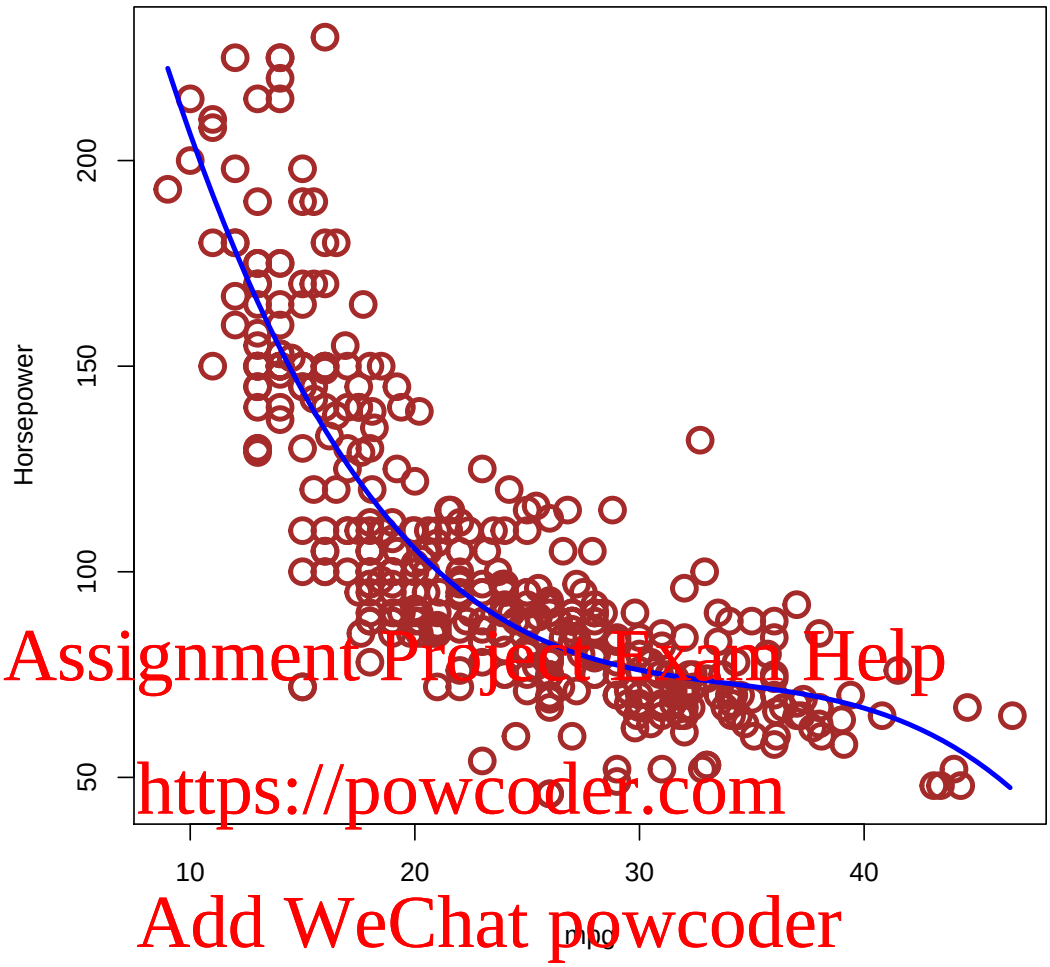


Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

```
# #Best fit is p*
pstar <- which.min(MSE_valid_p)
poly_best_fit <- lm(horsepower ~ poly(mpg,pstar), data = Auto)
with(Auto,
plot(mpg, horsepower, type = "p", col = "brown",
xlab = "mpg", ylab="Horsepower", cex = 2, lwd=3)
)
#This part is to fit fhat to the scatter plot
xpoints = with(Auto, seq(min(mpg), max(mpg),0.5))
#prediction using fhat at xpoints
ypoints <- predict(poly_best_fit, data.frame(mpg=xpoints))
#Plot the points on scatter plot
lines(xpoints, ypoints, col = "blue", lwd=3)
```



Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

2.5 This question is related to the `Weekly` data set, which is part of the ISLR package. The `Weekly` data set contains $1,089$ weekly returns for the S&P 500 stock index for 21 years, from the beginning of 1990 to the end of 2010.

(a) Use the full data set to perform a logistic regression with `Direction` as the response and the five lag variables plus `Volume` as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?

(b) Compute the misclassification error rate on the entire data set.

(c) Now fit the logistic regression model using a training data period from 1990 to 2008, with `Lag2` as the only predictor. (HINT: Use the function `filter` from the `dplyr` package to form the training and test sets).

(d) Compute the misclassification error rate on the test data set.

SOLUTION:

(a) The R codes are as follows:

```
#2.5
attach(Weekly)
library(dplyr) #for the filter function later
glm_fit <- glm(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
Volume, data = Weekly, family=binomial)
summary(glm_fit)
```

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = binomial, data = Weekly)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6949  -1.2565   0.9913   1.0849   1.4579
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.26686    0.08593   3.106   0.0019 **
## Lag1        -0.04127    0.02641  -1.563   0.1181
## Lag2         0.05844    0.02686   2.175   0.0296 *
## Lag3        -0.01606    0.02666  -0.602   0.5469
## Lag4        -0.02779    0.02646  -1.050   0.2937
## Lag5        -0.01447    0.02638  -0.549   0.5833
## Volume      -0.02274    0.03690  -0.616   0.5377
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1496.2  on 1088  degrees of freedom
## Residual deviance: 1486.4  on 1082  degrees of freedom
## AIC: 1500.4
##
## Number of Fisher Scoring iterations: 4
```

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

The p-value corresponding to Lag2 is singifcant at 5% level whereas all other independent variables have p-values that are insignificant. We conclude that Lag2 is the only variable that is significant to the classification.

(b) R codes to compute the misclassification error rate on the entire data set:

```
glm_probs <- predict(glm_fit, newdata = Weekly,
type = "response")
glm_pred <- ifelse(glm_probs > 0.5,"Up","Down" )
with( Weekly,
mean(glm_pred != Direction)
)
```

```
## [1] 0.4389348
```

(c) The R codes and output are as follows:

```
WeeklyTrain <- filter(Weekly, Year <= 2008)
WeeklyTest <- filter(Weekly, Year > 2008)
glm_fit_Train <- glm(Direction ~ Lag2, data = WeeklyTrain, family=b
summary(glm_fit_Train)
```

```
##
## Call:
## glm(formula = Direction ~ Lag2, family = binomial, data = Weekly
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.536  -1.264   1.021   1.091   1.368
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.20326    0.06428   3.162  0.00157 **
## Lag2         0.05810    0.02870   2.024  0.04298 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1354.7  on 984  degrees of freedom
## Residual deviance: 1350.5  on 983  degrees of freedom
## AIC: 1354.5
##
## Number of Fisher Scoring iterations: 4
```

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

(d) R codes to compute the misclassification error rate on the test data set and output.

```
glm_probs_Test <- predict(glm_fit_Train,
 newdata = WeeklyTest, type = "response")
glm_pred_Test <- ifelse(glm_probs_Test > 0.5,"Up","Down" )
with(WeeklyTest,
mean(glm_pred_Test != Direction)
)
```

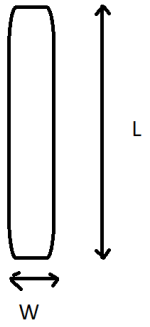
```
## [1] 0.375
```

2.6 This exercise illustrates the use of the `kmeans` clustering algorithm on a dataset in the `cluster.datasets` package. This dataset consists of milk composition of various mammals and the aim is to group the mammals according to similarities in the composition of their milk.

(a) Install the `cluster.datasets` package from CRAN and make the datasets available to you in R. Consider the `all.mammals.milk.1956` data set for `kmeans` clustering.

(b) Run the `kmeans` clustering for different number of clusters, K. Choose the optimal K, K^*, based on the elbow criteria.

(c) `kmeans` clustering is sensitive to scaling of the variables.



Consider this example: Let L and W represent the length and width of a skiboard, respectively, both measured in meters. It is clear that $L >> W$. If we want to cluster a collection of skiboards with data on (L, W) using `kmeans`, the clustering will be dominated by L. To avoid this, we scale both variables so that they are comparable. This is done using the R function `scale`. Typing `help(scale)` will give you the details on how this function scales all *appropriate* columns in a data frame so that they are comparable to each other.

(d) Investigate the variables in the `all.mammals.milk.1956` data set. Should the variables be scaled prior to running `kmeans`? Why?

(e) Run the `kmeans` algorithm on the scaled data set. Find the optimal K^* as before. Has your findings changed?

(f) Repeat (b) using GMMs. Use the option `modelNames = "EII"` in `mclustBIC`.

Assignment Project Exam Help

SOLUTION:

https://powcoder.com

(a) The relevant R codes and output are as follows:

Add WeChat powcoder

```
#2.6
#install.packages("cluster.datasets")
library(cluster.datasets)
df <- all.mammals.milk.1956
df1 <- df[,2:6]
#taking only the quantitative variables
#for clustering
K = 10;
W <- vector("numeric", K)
for (k in 1:K){
km_out <- kmeans(df1, k, nstart=20)
W[k] <- km_out$tot.withinss
}
plot(c(1:K), W, type="b", pch=20, cex=2, main="W versus k",
xlab="Number of clusters", ylab="Within cluster SS")
```

(b) R codes are as in part (a). Based on Figure 2, the optimal K^* is 4.

(c) Yes, the variables should be scaled since they are of different magnitudes.

(d) The R codes and figure are as follows:

```
df1scaled <- scale(df1)
K = 10;
W <- vector("numeric", K)
for (k in 1:K){
km_out <- kmeans(df1scaled, k, nstart=20)
```

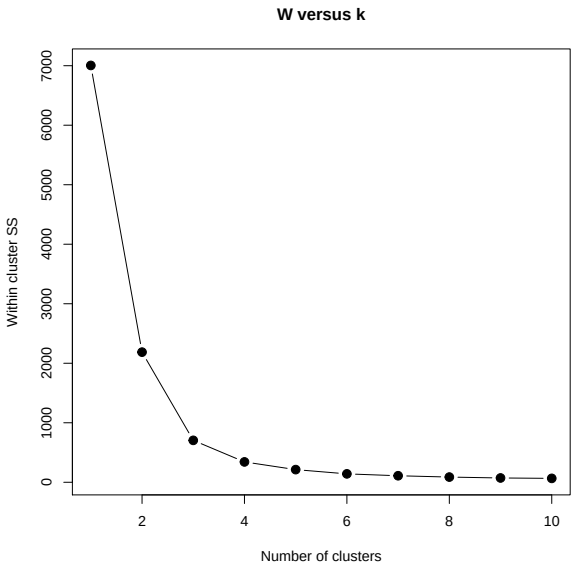


Figure 2: Plot of within SS versus K for original quantitative variables in the dataset.

```
W[k] <- km_out$tot.withinss
}
plot(c(1:K), W, type="b", pch=20, cex=2, main="W versus k",
xlab="Number of clusters", ylab="Within cluster SS")
```

Assignment Project Exam Help
https://powcoder.com

Based on Figure 3, the optimal K^* is still 4. However, note the change in the range of values of the y-axis. There is no change in K^* but this is just a coincidence for this particular dataset. Always remember to scale your variables if they are of different magnitudes.

Add WeChat powcoder

(e)
```
library(mclust)

BIC = mclustBIC(df1, modelNames = "EEI")
plot(BIC)
```

Based on Figure 4, the optimal K^* is 7.

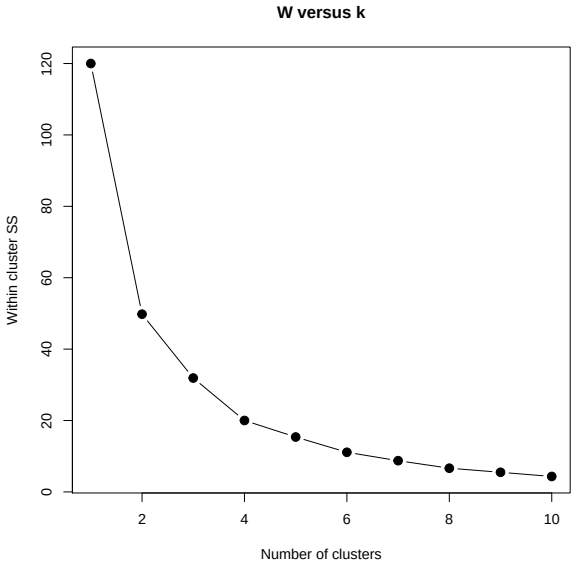


Figure 3: Plot of within SS versus K for the scaled quantitative variables in the dataset.

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

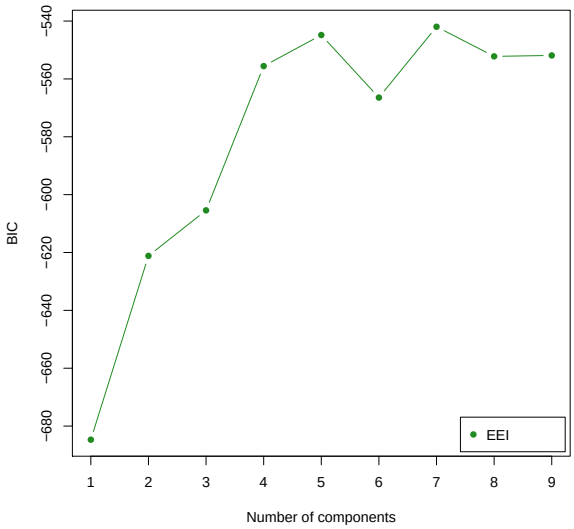


Figure 4: Plot of BIC values versus K.