ACT02 - Data Mining

Team 7 | Captain: Nadine Rose | Members: Kyle Scott, Lakshya Rathore, Bailey LaRea

Monday, January 31, 2022

Problem 1.1

We are given a 3x5 table with a response as well as two potential models to predict that response. We need to calculate for each model, the SSE, MSE, R^2, MAPE and MAE. More in depth descriptions of calculations will be found within the Problem 2 Section.

```r
# copy data from assignment
response <- c(3, 4, 5, 6, 7)
model_1 <- c(3.2, 4.3, 4.9, 5.7, 6.9)
model_2 <- c(3.3, 4.2, 4.8, 5.9, 7.1)

# save as data frame
q1 <- data.frame(response, model_1, model_2)
```

```r
residuals_1 <- q1$response - q1$model_1
residuals_2 <- q1$response - q1$model_2

sse_1 <- sum((residuals_1)^2)
sse_2 <- sum((residuals_2)^2)
```

```r
n <- length(response)

mse_1 <- sse_1 / n
mse_2 <- sse_2 / n
```

```r
rsq_1 <- cor(q1$response, q1$model_1)^2
rsq_2 <- cor(q1$response, q1$model_2)^2
```

```r
mape_1 <- sum(abs(q1$response - q1$model_1) / q1$response) / n * 100
mape_2 <- sum(abs(q1$response - q1$model_2) / q1$response) / n * 100
```

```r
mae_1 <- sum(abs(q1$response - q1$model_1)) / n
mae_2 <- sum(abs(q1$response - q1$model_2)) / n
```

```r
library(kableExtra)
Measure <- c('SSE', 'MSE', 'R^2', 'MAPE', 'MAE')
Model1 <- c(sse_1, mse_1, rsq_1, mape_1, mae_1)
Model2 <- c(sse_2, mse_2, rsq_2, mape_2, mae_2)

tbl <- data.frame(Measure, Model1, Model2)

kable(tbl, caption = "Statistical Values for Model 1 and Model 2", linesep = "\\addlinespace",
      digits = 3, booktabs = T, format = 'pandoc')
```

Table 1: Statistical Values for Model 1 and Model 2

Measure	Model1	Model2
SSE	0.240	0.190
MSE	0.048	0.038
R^2	0.988	0.986
MAPE	4.519	4.419
MAE	0.200	0.180

Problem 1.2

Question 1

We are tasked with minimizing $\text{Var}(\alpha X + (1-\alpha)Y)$

The roadmap for this proof will be to first expand the equation and then to take the first partial derivative with respect to α. Finally, we will check that the critical point is indeed a minimum by finding the second partial derivative with respect to alpha and show that it is positive.

$$\text{Var}(\alpha X + (1-\alpha)Y) = \alpha^2\sigma_X^2 + (1-\alpha)^2\sigma_Y^2 + 2\alpha(1-\alpha)\sigma_{XY} \tag{1}$$

Begin by taking the first partial,

$$\frac{\partial}{\partial\alpha}\text{Var}(\alpha X + (1-\alpha)Y) = 2\alpha\sigma_X^2 - 2\sigma_Y^2 + 2\alpha\sigma_Y^2 + 2\sigma_{XY} - 4\alpha\sigma_{XY} \tag{2}$$

Set equation equal to 0 and find the critical points,

$$2\alpha\sigma_X^2 - 2\sigma_Y^2 + 2\alpha\sigma_Y^2 + 2\sigma_{XY} - 4\alpha\sigma_{XY} = 0 \tag{3}$$

From this, we get

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}} \tag{4}$$

To check that this is a minimum, take the second partial,

$$\frac{\partial^2}{\partial\alpha^2}\text{Var}(\alpha X + (1-\alpha)Y) = 2\sigma_X^2 + 2\sigma_Y^2 - 4\sigma_{XY} = 2\text{Var}(X-Y) \geq 0 \tag{5}$$

Because Variance is always positive. ■

Question 2

(a) $1 - \frac{1}{n}$

(b) $1 - \frac{1}{n}$

(c) For bootstrapping sampling with replacement, we have the probability that the j^{th} observation is not the in bootstrap sample is the product of probabilities that each bootstrap observation is not the j^{th} observation from the original sample as these probabilities are independent. For a total of n observations, using the product rule gives us $(1 - \frac{1}{n})^n$

(d) P(j^{th} observation in the bootstrap sample) $= 1 - (1 - \frac{1}{5})^5 \approx 0.672$

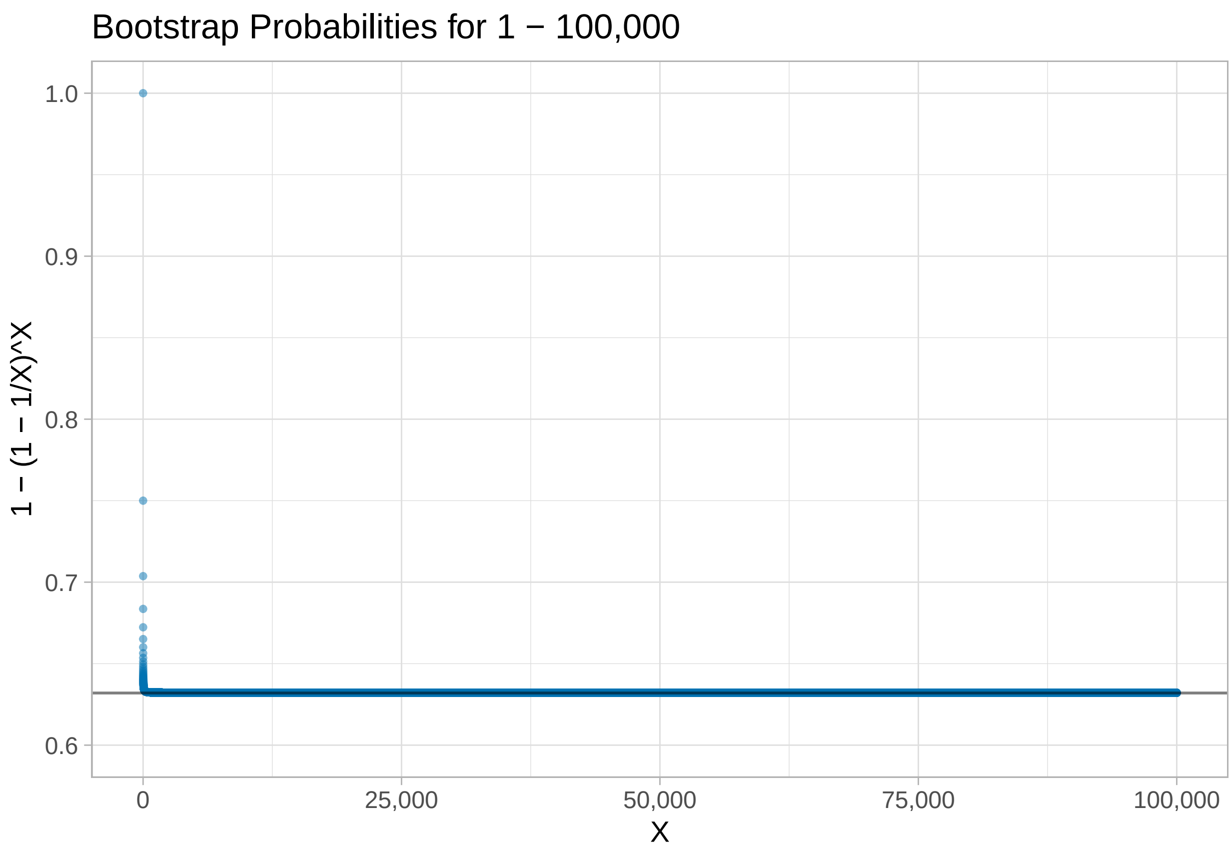
(e) P(j^{th} observation in the bootstrap sample) $= 1 - (1 - \frac{1}{100})^{100} \approx 0.634$

(f) P(j^{th} observation in the bootstrap sample) $= 1 - (1 - \frac{1}{10000})^{10000} \approx 0.632$

(g)

```r
library(ggplot2)
library(scales)
x <- 1:100000
y <- 1 - (1 - 1/x)^x
bsplt <- ggplot(data = NULL, aes(x = x, y = y)) +
  geom_point(alpha = 0.5, size = 0.8, color = "#0072B2") +
  geom_hline(yintercept = 0.632, size = 0.5, alpha = 0.5) +
  ggtitle(label = "Bootstrap Probabilities for 1 - 100,000") +
  xlab("X") +
  ylab("1 - (1 - 1/X)^X") +
  scale_x_continuous(labels = comma) +
  scale_y_continuous(labels = comma) +
  ylim(c(0.6, 1)) +
  theme_light()

bsplt
```



As shown by the h-line, we can see that the graph very quickly begins to converge upon the value 0.632 as we discovered in the 3 parts previous.

(h)

```
store <- rep(NA, 10000)
for (i in 1:10000) {
  store[i] <- sum(sample(1:100, rep = TRUE) == 4) > 0
}
mean(store)
```

```
## [1] 0.6363
```

We can see that the results produced by this code comes close to the limit we found in the previous problem. This is showing how within a sample, we will find roughly the same results as the theoretical probabilities.

Question 3

(a) k-fold cross validation is implemented by having a single parameter that refers to the number of groups that a data set will be split into. When a certain value for k is chosen, it will replace k and k=15 means it is 15-fold. It is used in ML to learn modeling on data not seen. It is less biased than a test split.

   (i) You will shuffle the data

   (ii) Split into k groups

   (iii) Fit a model on the training set to investigate the test set

   (iv) Keep the evaluation score and get rid of the model

(b) Advantages and disadvantages to k-fold cross validation:

   (i) The advantages of k-fold cross validation relative to validation set approach is that the computation time is reduced as the repetition of the process is limited to k times, it has reduced bias, the variance of the resulting estimate is reduced as k increases, and every data point get to be tested exactly once and is used for training (k-1) times. The limitation of k-fold cross validation is that the training algorithm is computationally intensive as the algorithm has to rerun from scratch for k times. Validation set approach is better in some ways for example using this method the MSE can be calculated using any modeling algorithm and without even sorting the data.

   (ii) The advantage of using k-fold cross validation vs LOOCV is that it takes into consideration more than one observation. It is also very intensive computationally vs k-fold. k-fold is also better since it has less bias. LOOCV is better as it addresses the drawback of using smaller datasets.

Problem 2.1

```r
# Read in the csv file
library(readr)
library(kableExtra)
housePrices <- read_csv("House_Prices_PRED.csv")
kable(head(housePrices), caption = "Head of HousePrices Data", linesep = "\\addlinespace",
      digits = 3, booktabs = T, format = 'pandoc')
```

Table 2: Head of HousePrices Data

Id	P_SalePrice	SalePrice
1	206307.7	208500
2	179044.5	181500
3	217258.4	223500
4	161547.6	140000
5	272594.2	250000
6	154557.5	143000

Problem 2.2

$$\text{SSE} = \sum_{i=1}^{n}(Y_i - \hat{f}(x_i))^2 \tag{6}$$

```r
# Calculate the residuals of the model
residuals <- housePrices$SalePrice - housePrices$P_SalePrice


# Square the residual of the model
sse <- (residuals)^2
# Find the sum to get SSE
sse <- sum(sse)
```

Sum of Squares Error (SSE) = 740,014,639,177.164

$$\text{MSE} = \frac{\text{SSE}}{n} \tag{7}$$

```r
# Find the length of our dataset
n <- length(residuals)
# Divide the SSE by n to find MSE
mse <- sse/n
```

Average or Mean Squared Error (MSE) = 506,859,341.902

Problem 2.3

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{f}(x_i))^2}{\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2} \tag{8}$$

Luckily, there is a more simplistic way to calculate R^2, which is as follows:

$$R^2 = cor(Y_i, \hat{f}(x_i))^2 \tag{9}$$

```r
# Find R^2 using our 2nd formula
rsq <- cor(housePrices$SalePrice, housePrices$P_SalePrice)^2
```

$R^2 = 0.923$

Problem 2.4

$$\text{MAPE} = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{Y_i - \hat{f}(x_i)}{Y_i} \right| \tag{10}$$

```
# Find the summation portion of the equation
mape <- sum(abs((housePrices$SalePrice - housePrices$P_SalePrice) /
      housePrices$SalePrice))
# Divide by n and turn the number into a percentage by multiplying by 100
mape <- mape / n * 100
```

MAPE = 7.026%

Problem 2.5

$$\text{MAE} = \frac{\sum_{i=1}^{n} \left| \hat{f}(x_i) - Y_i \right|}{n} \tag{11}$$

```
# Find MAE using the above equation
mae <- sum(abs(housePrices$SalePrice - housePrices$P_SalePrice)) / n
```

MAE = 12,470.834

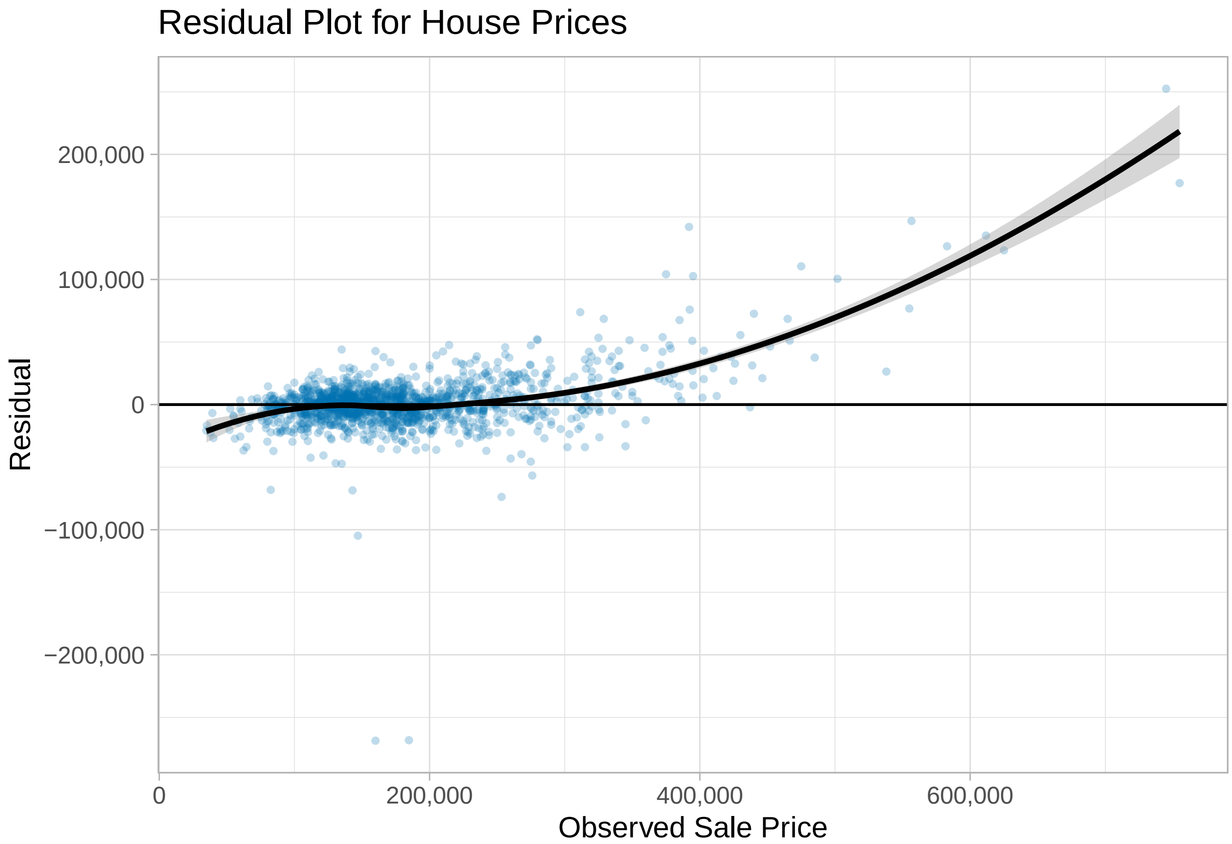
Problem 2.6

```r
library(ggplot2)
library(ggthemes)
library(scales)

# add a column to our dataframe for residuals
housePrices <- cbind(housePrices,residuals)

# build residual plot with observed sale price on the x and residuals on the y
resid_plt <- ggplot(data = housePrices,
                    aes(x = SalePrice, y = residuals)) +
            geom_point(alpha = 0.25, size = 0.8, color = '#0072B2') +
            geom_smooth(method = 'loess', color = 'black') +
            geom_hline(yintercept = 0) +
            ggtitle(label = "Residual Plot for House Prices") +
            xlab("Observed Sale Price") +
            ylab("Residual") +
            scale_x_continuous(labels = comma) +
            scale_y_continuous(labels = comma) +
            theme_light()

resid_plt
```


Residual Plot for House Prices
200,000
100,000
0
−100,000
−200,000
Residual
0
200,000
400,000
600,000
Observed Sale Price