# Solutions to Assignment 2

Rongfei Jin

January 24, 2025

## Conceptual 1

(a) Flexible methods are better because there are enough samples to reduce variance without overfitting. Additionally, having a small number of predictors reduces the complexity of the flexible model—both in terms of search space and time complexity—compared to the same flexible model with more predictors. As a result, the flexible model can capture more information from the sample than inflexible methods.

(b) An inflexible method is preferable here. The small number of observations increases the variance of the model, making it more prone to overfitting. By contrast, an inflexible model makes fewer assumptions and is therefore less likely to overfit.

(c) A flexible method works better. Due to the strong linearity, an inflexible model cannot further reduce the bias, even if the dataset is large.

(d) An inflexible method is better in this situation. The high variance of the error terms suggests the sample might be somewhat incorrectly collected. A flexible model would try to fit the randomness introduced by the error terms, leading to a higher chance of overfitting.

## Conceptual 2

(a) Regression: Salary is continuous. Inference is the focus here, since we want to understand what causes salary growth and utilize that information about the predictors.

(b) Classification: Success and failure are discrete outcomes. Inference is the focus here as well, since the company needs to identify the factors that lead to more success.

(c) Regression: The rate of change is continuous. We focus on prediction in this case, because we cannot alter any of the predictors but can still utilize insights about the rate of change.

## Conceptual 5

Advantages of Flexible method:

1. is better at capturing the non-linearity
2. Allows greater modification based on specific area knowledge
3. Allows the change of pattern in data

Disadvantages of Flexible method:

1. Has a higher chance of overfitting
2. has less interpretability
3. Requires more data
4. May have high time complexity

Senarios where flexible methods are preferable

1. The pattern is highly non-linear
2. Our proposed model is very different from the assumptions of inflexible models
3. The pattern evolutes over time
4. We don't care about interpretability

Senarios where inflexible methods are preferable

1. The pattern is highly linear
2. We want to interpret the model
3. We have limited data

## Conceptual 6

Parametric models assume functional forms of the models where the Non-Parametric models don't. Examples of parametric models linear, logistic regression, Linear SVM, Naive Bayes
Advantages of Parametric models:

1. has more interpretability
2. works better on small datasets
3. less likely to overfit
4. requires less computations

Disadvantages of Parametric method:

1. may underfit the data when the true model is highly non-linear
2. requires correct prior knowledge on the data to construct the model

## Conceptual 7

we compute the euclidean distance by

$$d_i = \sqrt{X_{i1}^2 + X_{i2}^2 + X_{i3}^2} \text{ where } X_{ij} \text{ means the value jth predictor of ith observation}$$

We can easily compute the values to be

$$\{d_1, d_2, d_3, d_4, d_5, d_6\} = \{3, 2, 3.16, 2.24, 1.41, 1.73\}$$

when $k = 1$ we choose the one nearest neighbor which in this case is $X_5$ thus the prediction is the same as the $Y_5 = \text{Green}$

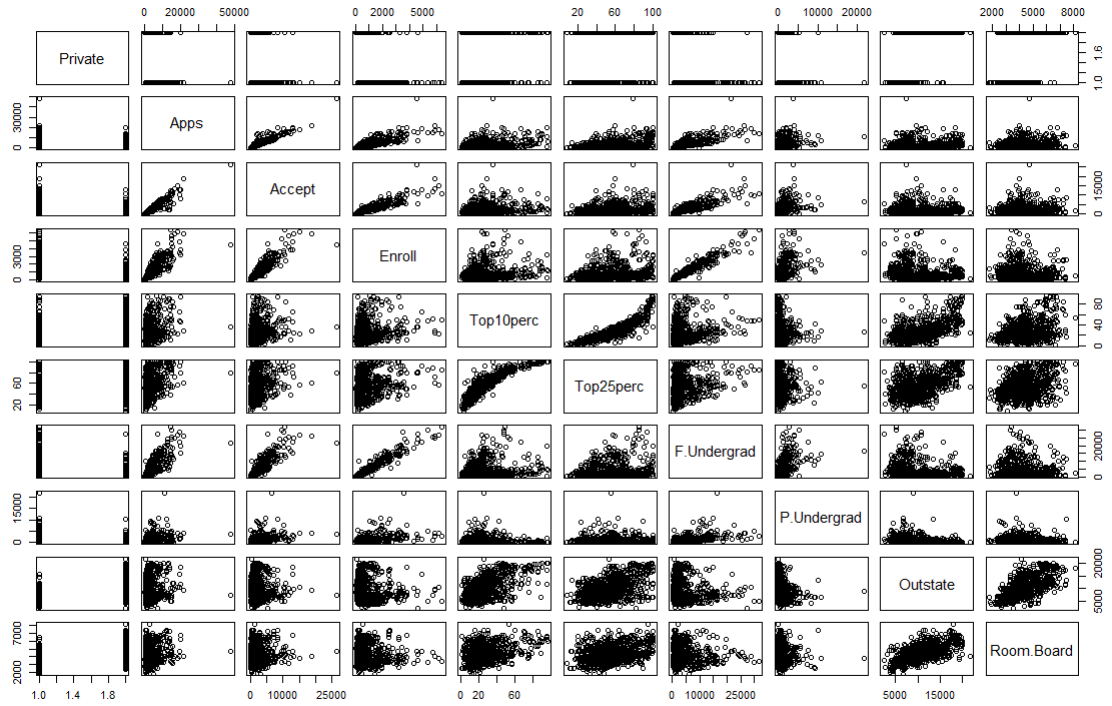
when $k = 3$, we choose 3 nearest neighbor which is $X_2, X_5, X_6$ and by majority vote the result is Red

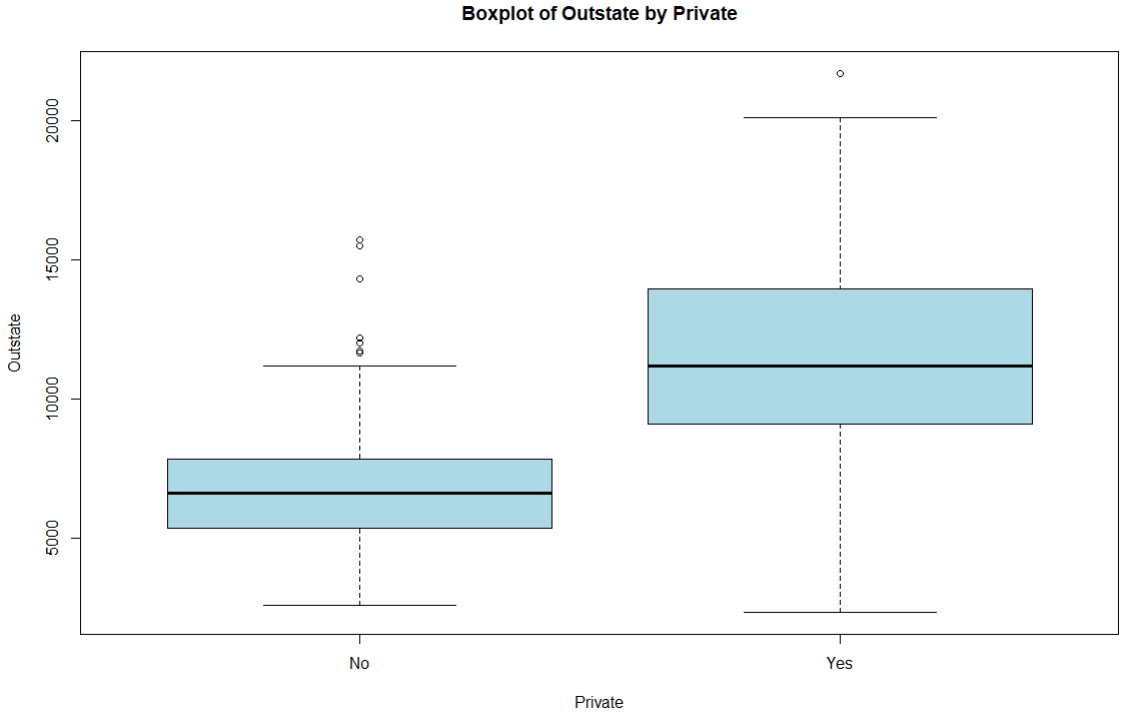
if the decision boundry is highly non-linear, then we prefit a small $k$ since a larger $k$ will underfits the data by including votes that are not in the same group

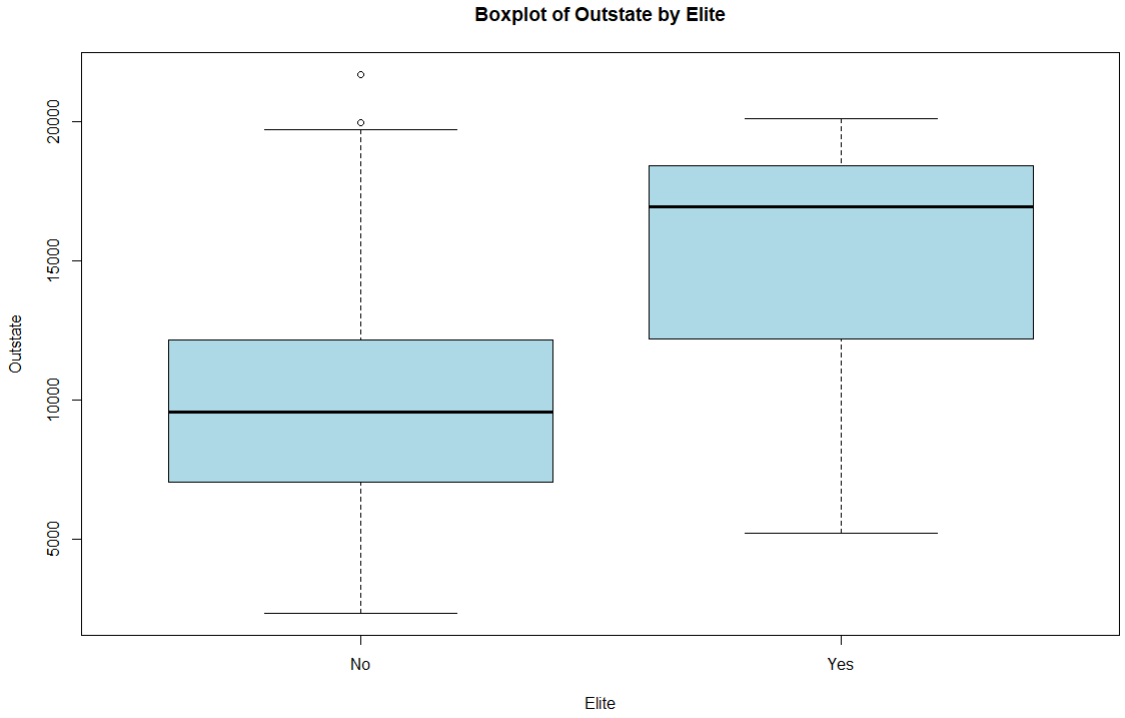
# Applied 8

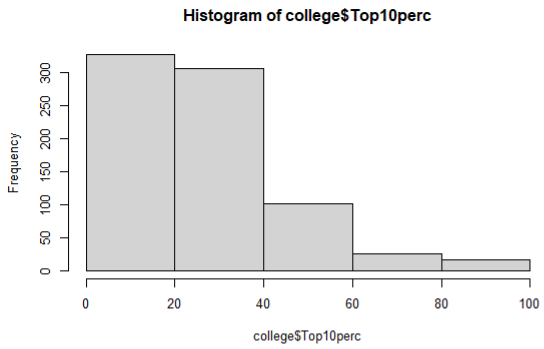
```r
# (a)
college <- read.csv("College.csv")


# (b)
rownames(college) <- college[,1]
college <- college[,-1]

# (c) i
summary(college)

# (c) ii
college[,1] <- as.factor(college[,1])
pairs(college[,1:10])

# (c) iii
plot(college$Private, college$Outstate,
     main = "Boxplot of Outstate by Private",
     xlab = "Private",
     ylab = "Outstate",
     col  = "lightblue")

# (c) iv
Elite <- rep("No", nrow(college))
Elite[college$Top10perc > 50] <- "Yes"
Elite <- as.factor(Elite)
college <- data.frame(college , Elite)

plot(college$Elite, college$Outstate,
     main = "Boxplot of Outstate by Elite",
     xlab = "Elite",
     ylab = "Outstate",
     col  = "lightblue")

par(mfrow = c(2, 2))
hist(college$Top10perc, breaks = 5)
hist(college$Top10perc, breaks = 10)
hist(college$Top10perc, breaks = 15)
hist(college$Top10perc, breaks = 20)
# Elite schools have more top students but they cost more
# Most non-elite schools only receive 20% of the top 10% from the high school class

par(mfrow = c(1, 1))
plot(college$Elite, college$S.F.Ratio,
     main = "Boxplot of S.F ratio by Elite",
     xlab = "Elite",
     ylab = "S.F. Ratio",
     col  = "lightblue")

# An elite school may not have a higher S.F. Ratio than a non-elite school
# And the overall difference in S.F Ratio is not large
```

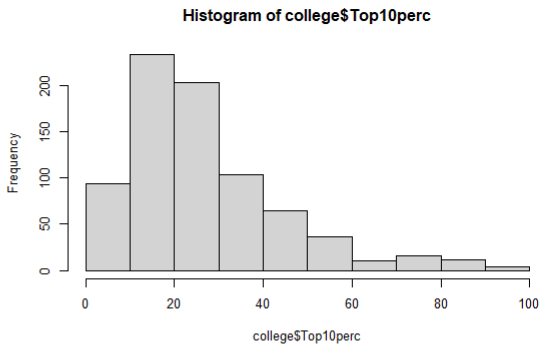
```
> summary(Auto)
      mpg          cylinders      displacement     horsepower        weight      acceleration        year           origin                      name
 Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613   Min.   : 8.00   Min.   :70.00   Min.   :1.000   amc matador       :  5
 1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225   1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000   ford pinto        :  5
 Median :22.75   Median :4.000   Median :151.0   Median : 93.5   Median :2804   Median :15.50   Median :76.00   Median :1.000   toyota corolla    :  5
 Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5   Mean   :2978   Mean   :15.54   Mean   :75.98   Mean   :1.577   amc gremlin       :  4
 3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615   3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000   amc hornet        :  4
 Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140   Max.   :24.80   Max.   :82.00   Max.   :3.000   chevrolet chevette:  4
                                                                                                                                 (other)           :365
```

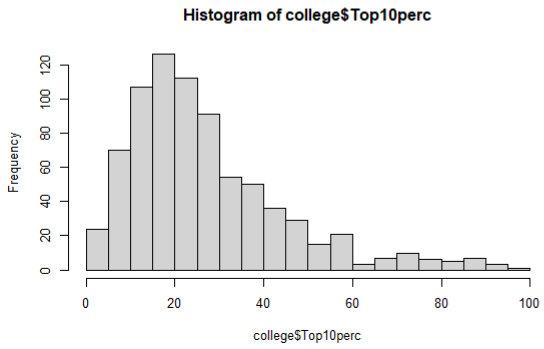

Private
Apps
Accept
Enroll
Top10perc
Top25perc
F.Undergrad
P.Undergrad
Outstate
Room.Board

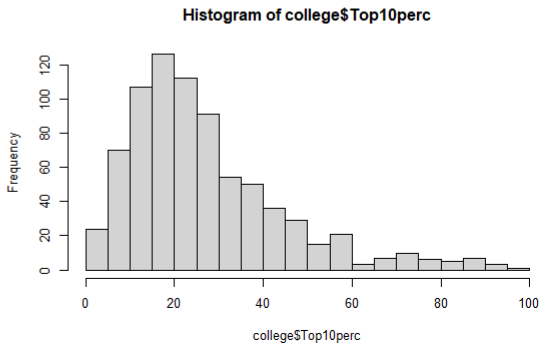

Boxplot of Outstate by Private
Outstate
Private
No
Yes
5000
10000
15000
20000

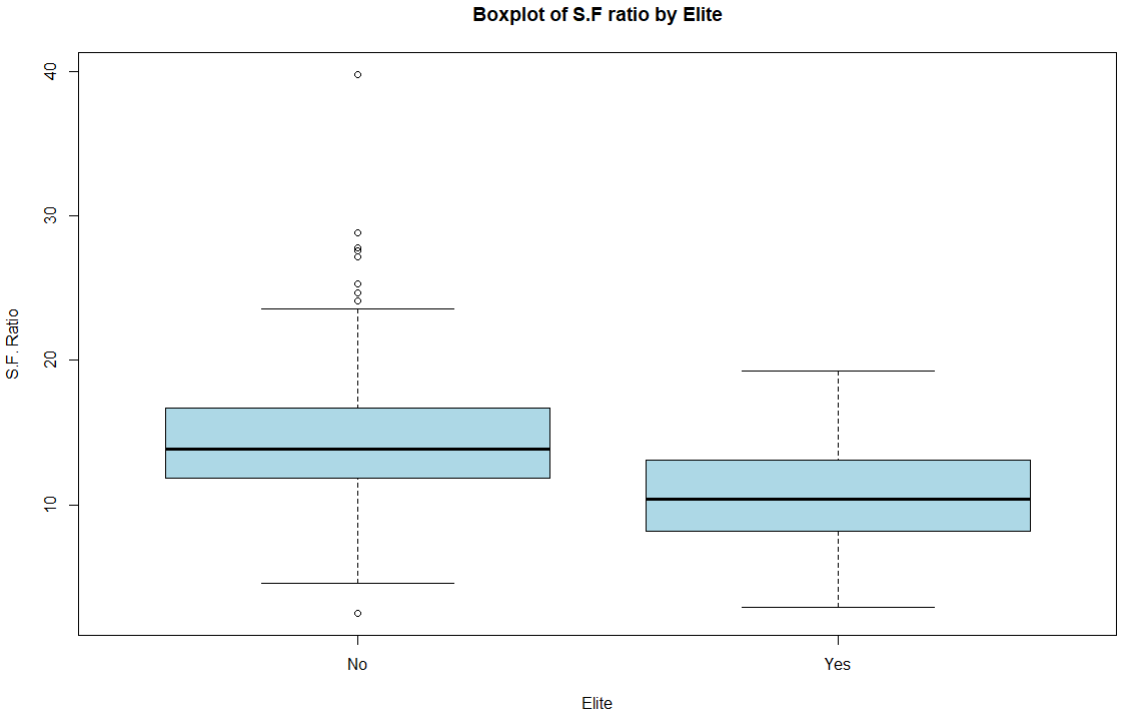

Boxplot of Outstate by Elite
Outstate
5000
10000
15000
20000
No
Yes
Elite


Histogram of college$Top10perc
Frequency
0
50
100
150
200
250
300
0
20
40
60
80
100
college$Top10perc


Histogram of college$Top10perc
Frequency
0
50
100
150
200
0
20
40
60
80
100
college$Top10perc


Histogram of college$Top10perc
Frequency
0
20
40
60
80
100
120
0
20
40
60
80
100
college$Top10perc

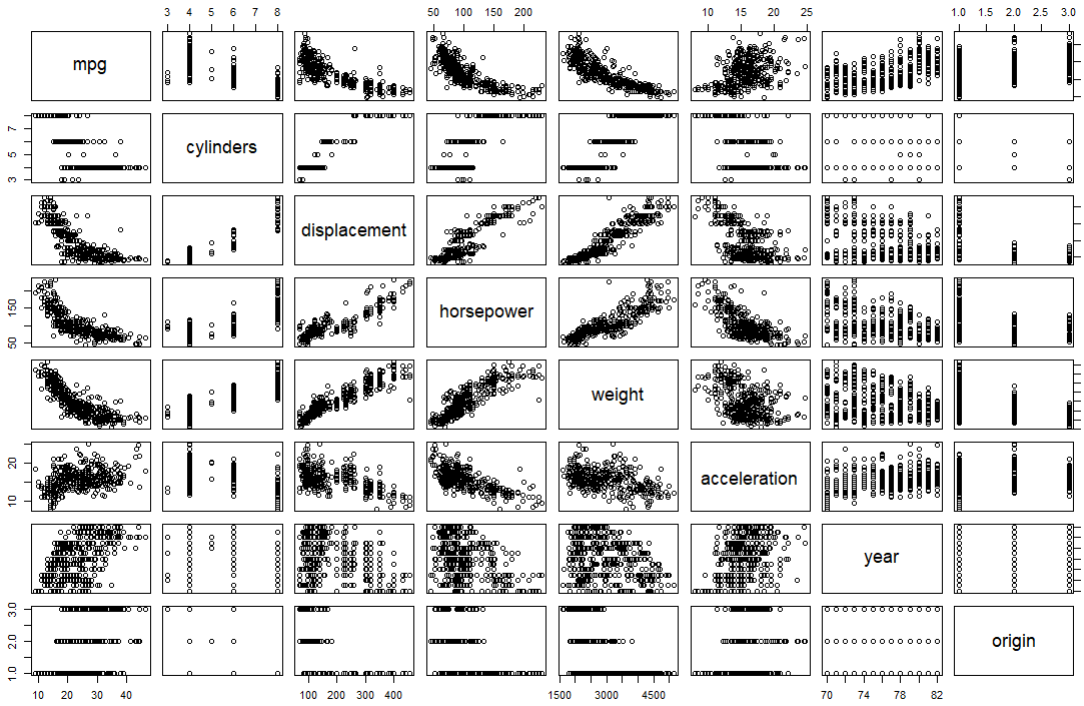

Histogram of college$Top10perc
Frequency
0
20
40
60
80
100
120
0
20
40
60
80
100
college$Top10perc


Boxplot of S.F ratio by Elite
S.F. Ratio
10
20
30
40
No
Yes
Elite

## Applied 9

```r
library(ISLR2)

# (a)
summary(Auto)
sapply(Auto, class)
# Quantitative: mpg, displacement, horsepower, weight, acceleration, year
# Qualitative: cylinders, origin

# (b) (c)
continuous = c("mpg", "displacement", "horsepower", "weight", "acceleration", "year")
for (v in continuous) {
  cat(v,"\n")
  cat("range:", range(Auto[[v]]), "\t")
  cat("mean:", mean(Auto[[v]]), "\t")
  cat("sd:", sd(Auto[[v]]), "\n\n")
}

# (d)
Auto_subset <- Auto[-c(10:85),]

for (v in continuous) {
  cat(v,"\n")
  cat("range:", range(Auto_subset[[v]]), "\t")
  cat("mean:", mean(Auto_subset[[v]]), "\t")
  cat("sd:", sd(Auto_subset[[v]]), "\n\n")
}

pairs(Auto[,-ncol(Auto)])
# More Cylinder means less mpg
# Engine displacement and mpg are negatively related
# Acceleration and mpg are not strongly correlated

# (e)

# A few variables can be used to predict mpg,
# such as cylinders, displacement, horsepower, weight
# because they shows strong linear pattern with mpg
```

```
mpg
range: 11 46.6  mean: 24.40443  sd: 7.867283

displacement
range: 68 455   mean: 187.2405  sd: 99.67837

horsepower
range: 46 230   mean: 100.7215  sd: 35.70885

weight
range: 1649 4997        mean: 2935.972  sd: 811.3002

acceleration
range: 8.5 24.8         mean: 15.7269   sd: 2.693721

year
range: 70 82    mean: 77.14557  sd: 3.106217
```


mpg
cylinders
displacement
horsepower
weight
acceleration
year
origin

# Additional

(a) For any points $(y_i, x_i)$, the residual can be expressed as $\epsilon_i = y_i - \beta x_i$. The total residual sum of square is

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \beta x_i)^2 = \sum_{i=1}^{n}(y_i^2 - 2\beta x_i + \beta^2 x_i^2)$$

We then minimize the RSS with respect to $\beta$ by taking derivative of RSS

$$\frac{d}{d\beta}(\text{RSS}) = -2\beta\sum_{i=1}^{n}(y_i x_i) + 2\beta\sum_{i=1}^{n} x_i^2$$

We set the derivative to 0 and obtain the solution to minimization

$$\frac{d}{d\beta}(\text{RSS}) = 0$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}$$

(b) We first check if the $\hat{\beta}$ is not biased given $\epsilon \sim N(0, \sigma^2)$

$$\begin{aligned}
\mathbf{E}(\hat{\beta}) &= \mathbf{E}(\frac{\sum_{i=1}^{n}(\beta x_i + \epsilon)x_i}{\sum_{i=1}^{n} x_i^2}) \\
&= \frac{\mathbf{E}[\sum_{i=1}^{n}(\beta x_i^2 + x_i\epsilon)]}{\sum_{i=1}^{n} x_i^2} \\
&= \frac{[\beta\sum_{i=1}^{n} x_i^2 + \mathbf{E}(\epsilon)\sum_{i=1}^{n} x_i]}{\sum_{i=1}^{n} x_i^2} \\
&= \frac{[\beta\sum_{i=1}^{n} x_i^2 + 0 \cdot \sum_{i=1}^{n} x_i]}{\sum_{i=1}^{n} x_i^2} \\
&= \beta
\end{aligned}$$

$\hat{\beta}$ is indeed unbiased, thus

$$\begin{aligned}
\text{MSE}(\hat{\beta}) &= \text{Var}(\hat{\beta}) + 0 \\
&= \text{Var}(\frac{\sum_{i=1}^{n}(\beta x_i + \epsilon)x_i}{\sum_{i=1}^{n} x_i^2}) \\
&= \frac{1}{(\sum_{i=1}^{n} x_i^2)^2}\,\text{Var}(\epsilon)\sum_{i=1}^{n} x_i^2 \\
&= \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}
\end{aligned}$$

(c) For $\beta$ to be unbiased, the residuals must be centered on 0, which means there's no intercept in the true model that causes a systemtatic shift to residuals, x and y must have a relationship that passes through the origin.

(d)

$$\begin{aligned}
\hat{\beta}_1 - \hat{\beta} &= \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} - \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2} \\
\hat{\beta}_1 - \hat{\beta} &= \frac{\sum_{i=1}^{n}(y_i x_i - y_i\bar{x} - x_i\bar{y} + \bar{y}\bar{x})}{\sum_{i=1}^{n} x_i^2 - 2x_i\bar{x} + \bar{x}^2} - \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2} \\
\hat{\beta}_1 - \hat{\beta} &= \frac{(\sum_{i=1}^{n} y_i x_i) - n\bar{y}\bar{x}}{(\sum_{i=1}^{n} x_i^2) - n\bar{x}^2} - \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}
\end{aligned} \qquad \sum_{i=1}^{n} x_i = n\bar{x}$$

$\hat{\beta}_1$ and $\hat{\beta}$ are only the same if all $x_i$ are centered on 0. If only centering $y_i$, the bias persists.