1. (10%) Given a simple linear regression model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \dots, n$, where $\epsilon_i \sim_{iid} N(\mu, \sigma^2)$.
Prove that:
   a. $\operatorname{cov}(\hat\beta_0, \hat\beta_1) = -\bar{x}\sigma^2/S_{xx}$
   b. $\operatorname{cov}(\bar{y}, \hat\beta_1) = 0$

(a)

Let $S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$

Step 1: Normal Equation:

$$\frac{\partial S}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \hat\beta_0 - \hat\beta_1 x_i) = 0$$

$$\frac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i (y_i - \hat\beta_0 - \hat\beta_1 x_i) = 0$$

$$\Rightarrow \begin{cases} \sum_{i=1}^{n}(y_i - \hat\beta_0 - \hat\beta_1 x_i) = 0 \quad \cdots ① \\ \sum_{i=1}^{n}(y_i - \hat\beta_0 - \hat\beta_1 x_i)x_i = 0 \quad \cdots ② \end{cases}$$

By ①

$$\sum_{i=1}^{n} y_i - n\hat\beta_0 - \hat\beta_1 \sum_{i=1}^{n} x_i = 0 \Rightarrow \hat\beta_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat\beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\because \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \Rightarrow \hat\beta_0 = \bar{y} - \hat\beta_1 \bar{x}$$

Subtitute $\hat\beta_0$ to ②

$$\sum_{i=1}^{n}\left[y_i - (\bar{y} - \hat\beta_1 \bar{x}) - \hat\beta_1 x_i\right] x_i = 0$$

$$\Rightarrow \sum_{i=1}^{n}\left[y_i - \bar{y} + \hat\beta_1 \bar{x} - \hat\beta_1 x_i\right] x_i = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i x_i - \bar{y}\sum_{i=1}^{n} x_i = -\hat\beta_1 \left(\sum_{i=1}^{n} x_i \bar{x} + \sum_{i=1}^{n} x_i^2\right)$$

$$\Rightarrow \hat\beta_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat\beta_0 = \bar{y} - \hat\beta_1 \bar{x}$$

Step 2: $\because \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$$\text{cov}(\hat{\beta}_0, \hat{\beta}_1) = \text{cov}(\bar{y} - \hat{\beta}_1 \bar{x}, \hat{\beta}_1) = \underbrace{\text{cov}(\bar{y}, \hat{\beta}_1)}_{\underset{0 \text{ (Poved in (b))}}{\parallel}} - \bar{x}\,\text{Var}(\hat{\beta}_1)$$

$$\Rightarrow \text{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\bar{x}\,\text{Var}(\hat{\beta}_1)$$

- Derive $\text{Var}(\hat{\beta}_1)$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(X_i - \bar{x})^2} = \frac{1}{S_{xx}} \sum_{i=1}^{n}(X_i - \bar{x})y_i$$

$$= \frac{1}{S_{xx}} \sum_{i=1}^{n}(X_i - \bar{x})(\beta_0 + \beta_1 X + \varepsilon), \text{ only care } \varepsilon$$

$$= \frac{1}{S_{xx}} \sum_{i=1}^{n}(X_i - \bar{x})\,\varepsilon$$

$$\boxed{\because S_{xx} = \sum_{i=1}^{n}(X_i - \bar{x})}$$

$$\Rightarrow \text{Var}(\hat{\beta}_1) = \text{Var}\left(\frac{1}{S_{xx}} \sum_{i=1}^{n}(X_i - \bar{x})\varepsilon_i\right) = \frac{1}{S_{xx}^2} \sum_{i=1}^{n}(X_i - \bar{x})^2 \sigma^2$$

$$= \frac{1}{S_{xx}} \sigma^2$$

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\bar{x} \cdot \frac{\sigma^2}{S_{xx}} \quad \#$$

(b) $$\hat{\beta}_1 = \frac{1}{S_{xx}} \sum_{i=1}^{n}(X_i - \bar{x})y_i = \sum_{i=1}^{n} c_i y_i, \text{ where } c_i = \frac{x_i - \bar{x}}{S_{xx}}$$

$$\text{Cov}(\bar{y}, \hat{\beta}_1) = \text{Cov}\left(\frac{1}{n}\sum_{i=1}^{n} y_i, \sum_{j=1}^{n} c_i y_i\right) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} c_i\,\text{cov}(y_i, y_j)$$

$\because \varepsilon_i$ is independent, and $Var(\varepsilon_i) = \sigma^2$, $\varepsilon_i \sim N(0, \sigma^2)$

$$\Rightarrow cov(y_i, y_j) = \begin{cases} \sigma^2 & , \text{if } i = j \\ 0 & , \text{if } i \neq j \end{cases}$$

The double sum reduces to:

$$\Rightarrow Cov(\bar{y}, \hat{\beta}) = \frac{1}{n} \sum_{i=1}^{n} c_i \sigma^2 = \frac{\sigma^2}{n} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{S_{xx}} \quad , (\because \sum_{i=1}^{n} (x_i - \bar{x}) = 0)$$

$$Cov(\bar{y}, \hat{\beta}_1) = \frac{\sigma^2}{n S_{xx}} \cdot 0 = 0 \quad ※$$

2. (10%) Demonstrate that the regression sum of squares (SSR) can be computed using the following formula:

$$SS_R = \left(\sum_{i=1}^{n} \hat{y}_i^2\right) - n\bar{y}^2$$

The defination of SSR is $\sum_{i=1}^{n} (\hat{y} - \bar{y})^2$

$$\Rightarrow SS_R = \sum_{i=1}^{n} \hat{y}_i^2 - 2\bar{y}\sum_{i=1}^{n} \hat{y}_i + \sum_{i=1}^{n} \bar{y}^2 \quad (\because \sum_{i=1}^{n} \bar{y}^2 = n\bar{y}^2)$$

$$= \sum_{i=1}^{n} \hat{y}_i^2 - 2\bar{y} \sum_{i=1}^{n} \hat{y}_i + n\bar{y}^2$$

Derive $\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i) = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i$

$$= n(\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 \sum_{i=1}^{n} x_i = n\bar{y} + \hat{\beta}_1 \underbrace{\left(\sum_{i=1}^{n} x_i - n\bar{x}\right)}_{=0} = n\bar{y}$$

$$\Rightarrow SSR = \sum_{i=1}^{n} \hat{y}_i - 2\bar{y} \cdot n\bar{y} + n\bar{y}^2 = \sum_{i=1}^{n} \hat{y}_i - n\bar{y} \quad ※$$

3. (10%) Given a multiple regression model: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Prove that the LSE can be also expressed as $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \mathbf{R}\boldsymbol{\epsilon}$, where $\mathbf{R} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

By Normal Equation: $\hat{\beta} = (X^T X)^{-1} X^T y$

Substitute $y = X\beta + \varepsilon$

$$\Rightarrow \hat{\beta} = (X^TX)^{-1}X^T(X\beta+\varepsilon) = (X^TX)^{-1}X^T(X\beta+\varepsilon)$$
$$= (X^TX)^{-1}X^TX\beta + (X^TX)^{-1}X^T\varepsilon$$
$$= \beta + (X^TX)^{-1}X^T\varepsilon \quad , \text{ where } R = (X^TX)^{-1}X^T \quad \#$$

4. (10%) The matrix, $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, derived in multiple regression is usually defined as $\mathbf{H}$ and called the hat matrix. Show that:
   a. $\mathbf{H}$ is idempotent, i.e., $\mathbf{HH} = \mathbf{H}$ and $(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H}) = \mathbf{I}-\mathbf{H}$
   b. $\mathrm{V}(\hat{\mathbf{y}}) = \sigma^2\mathbf{H}$

a.
$$H = X(X^TX)^{-1}X^T ,$$
$$H^2 = [X(X^TX)^{-1}X^T][X(X^TX)^{-1}X^T]$$
$$= X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T$$
$$= X(X^TX)^{-1}X^T = H \quad \Rightarrow H \text{ is idempotent.}$$
$$(I-H)(I-H) = I^2 - 2H + H = I - H$$
$$\Rightarrow (I-H) \text{ is idempotent}$$

b.
$$\hat{y} = Hy, \quad y = X\beta + \varepsilon \qquad V(\varepsilon) = \sigma^2 I \Rightarrow V(y) = \sigma^2 I$$
$$\Rightarrow V(\hat{y}) = V(Hy) = HV(y)H^T = H(\sigma^2 I)H^T = \sigma^2 HH^T$$
$$\because H \text{ is idempotent} \quad \therefore V(\hat{y}) = \sigma^2 HH = \sigma^2 H \quad \#$$

5. (20%) Given a **simple** linear regression model,
   a. show that the elements of the hat matrix can be expressed as:
$$h_{ii} = \frac{1}{n} + \frac{(x_i-\bar{x})^2}{S_{xx}} \text{ and } h_{ij} = \frac{1}{n} + \frac{(x_i-\bar{x})(x_j-\bar{x})}{S_{xx}}.$$
   b. Analyze the behavior of h_{ii} and h_{ij} when x_i deviates from $\bar{x}$.

Simple Linear regression model,

(a) $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad , \quad i = 1, 2, \cdots, n$

The Hat matrix is defined : $H = X(X^T X)^{-1} X^T$

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} ,$$

$$X^T X = \begin{pmatrix} \sum_{i=1}^{n} 1 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix}$$

$$\Rightarrow (X^T X)^{-1} = \frac{1}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \begin{pmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{pmatrix}$$

$$= \frac{1}{n S_{xx}} \begin{pmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{pmatrix}$$

$$H = X(X^T X)^{-1} X^T , \Rightarrow h_{ij} = x_i^T (X^T X)^{-1} x_j$$

$$= (1 \quad x_i) \left[\frac{1}{n S_{xx}} \begin{pmatrix} \sum_{k=1}^{n} x_k^2 & -\sum_{k=1}^{n} x_k \\ -\sum_{k=1}^{n} x_k & n \end{pmatrix} \right] \begin{pmatrix} 1 \\ x_j \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} \sum_{k=1}^{n} x_k^2 & -\sum_{k=1}^{n} x_k \\ -\sum_{k=1}^{n} x_k & n \end{pmatrix} \begin{bmatrix} 1 \\ x_j \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{n} x_k^2 - x_j \sum_{k=1}^{n} x_k \\ -\sum_{k=1}^{n} x_k + n x_j \end{bmatrix}$$

$$h_{ij} = \frac{1}{n S_{xx}} \left\{ \left[1 \cdot \left(\sum_{k=1}^{n} x_k^2 - x_j \sum_{k=1}^{n} x_k \right) \right] + \left[x_i \cdot \left(-\sum_{k=1}^{n} x_k + n x_j \right) \right] \right\}$$

$$= \frac{1}{n S_{xx}} \left[\sum_{k=1}^{n} x_k^2 - x_j \sum_{k=1}^{n} x_k - x_i \sum_{k=1}^{n} x_k + n x_i x_j \right]$$

- $\sum_{k=1}^{n} x_k = n\bar{x}$
- $\sum_{k=1}^{n} x_k^2 = S_{xx} + n\bar{x}$

$$\Rightarrow h_{ij} = \frac{1}{n S_{xx}} \left[(S_{xx} + n\bar{x}^2) - n\bar{x}(x_i + x_j) + n x_i x_j \right]$$

$$= \frac{1}{n S_{xx}} \left[S_{xx} + n\left(x_i x_j - \bar{x}(x_i + x_j) + \bar{x}^2\right) \right]$$

- $x_i x_j - \bar{x}(x_i + x_j) + \bar{x}^2 = (x_i - \bar{x})(x_j - \bar{x})$

$$= \frac{1}{n S_{xx}} \left[S_{xx} + n[(x_i - \bar{x})(x_j - \bar{x})] \right] = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}}$$

For $i = j$ $\quad h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$

(b) Analysis of the Behavior

For $i=j$

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$$

When x_i is far from $\bar{x}$, $\frac{(x_i - \bar{x})^2}{S_{xx}}$ become larger, so h_{ii} have a greater influence on the fitted regression line.

For $i \neq j$. h_{ij} reflect the relationship between pairs of observations, when both observations are similarly far from $\bar{x}$ in the same direction and can be smaller when the observations lie on opposite sides of $\bar{x}$.

6. Assuming that the true underlying model is

$$\mathbf{y} = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \cdots \beta_p\mathbf{x}_p + \boldsymbol{\epsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

Jack intentionally ignores the intercept term and fit the data with the following model:

$$\mathbf{y} = \hat{\beta}_1\mathbf{x}_1 + \hat{\beta}_2\mathbf{x}_2 + \cdots \hat{\beta}_p\mathbf{x}_p = \mathbf{X}'\widehat{\boldsymbol{\beta}}'.$$

a. (10%) Show that $\mathrm{E}[\hat{\beta}_1] \neq \beta_1, \mathrm{E}[\hat{\beta}_2] \neq \beta_2, \dots, \mathrm{E}[\hat{\beta}_p] \neq \beta_p$.

b. (5%) What can you conclude from (a)?

a.

The true model is: $y = \beta_0 \mathbb{1} + X'\beta + \varepsilon$

where $\beta = (\beta_1, \dots, \beta_p)^T$, and X' is $n \times p$ matrix

$$\hat{\beta} = [(X')^T X']^{-1}(X')^T y$$

$$= [(X')^T X']^{-1}(X')^T(\beta_0 \mathbb{1} + X'\beta + \varepsilon)$$

$$\hat{\beta} = [(X')^T X']^{-1}(X')^T(\beta_0 \mathbb{1}) + \underbrace{[(X')^T X']^{-1}(X')^T X'}_{= I}\beta + [(X')^T X']^{-1}(X')^T \varepsilon$$

$$\Rightarrow \hat{\beta} = \beta + \beta_0[(X')^T X']^{-1}(X')^T \mathbb{1} + [(X')^T X']^{-1}(X')^T \varepsilon$$

$\because E[\varepsilon] = 0$

$\therefore E[\hat{\beta}] = \beta + \beta_0[(X')^T X']^{-1}(X')^T \mathbb{1}$,

$\Rightarrow E[\hat{\beta}] \neq \beta \Rightarrow E[\hat{\beta}_j] \neq \beta_j$, for $j = 1, \dots, p$

b. The conclusion is that omitting the intercept from regression model (when the true model includes one) results in biased estimates of slope coef.

When we centered each column of X', $E[\hat{\beta}]$ can equal to β.

7. (15%) In a multiple regression model: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, it is critical to know if $(\mathbf{X}^T\mathbf{X})^{-1}$ exists. The diagonal elements of $(\mathbf{X}^T\mathbf{X})^{-1}$ in the correlation form, where $\mathbf{X}$ is standardized, are known as Variance Inflation Factors (VIFs). They are crucial to diagnose multicollinearity. VIF for the j^{th} regression coefficient is expressed as

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2},$$

where R_j^2 is the coefficient of multiple determination obtained from regressing $\mathbf{x}_j$ on the other regressor variables ($\mathbf{x}_1$ to $\mathbf{x}_p$, except $\mathbf{x}_j$). Calculate all the VIFs in the "autompg" dataset and discuss your observation.

HW3 Problem7

Howard

2025-03-16

Problem 7

- Calculate all the `VIFs` in the “autompg” dataset and discuss your observation.

```
# Load required packages
library(readxl)   # to read Excel files
library(dplyr)    # for data manipulation
```

```
##
## 載入套件：'dplyr'
```

```
## 下列物件被遮斷自 'package:stats':
##
##     filter, lag
```

```
## 下列物件被遮斷自 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(car)      # for computing VIFs
```

```
## 載入需要的套件：carData
```

```
##
## 載入套件：'car'
```

```
## 下列物件被遮斷自 'package:dplyr':
##
##     recode
```

```
# Read the auto-mpg dataset from the Excel file
df <- read_excel("Auto-mpg/auto-mpg.xlsx")

# Inspect the structure of the dataset
str(df)
```

```
## tibble [392 × 9] (S3: tbl_df/tbl/data.frame)
##  $ mpg         : num [1:392] 18 15 18 16 17 15 14 14 14 15 ...
##  $ cylinders   : num [1:392] 8 8 8 8 8 8 8 8 8 8 ...
##  $ displacement: num [1:392] 307 350 318 304 302 429 454 440 455 390 ...
##  $ horsepower  : num [1:392] 130 165 150 150 140 198 220 215 225 190 ...
##  $ weight      : num [1:392] 3504 3693 3436 3433 3449 ...
##  $ acceleration: num [1:392] 12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ model year  : num [1:392] 70 70 70 70 70 70 70 70 70 70 ...
##  $ origin      : num [1:392] 1 1 1 1 1 1 1 1 1 1 ...
##  $ car name    : chr [1:392] "chevrolet chevelle malibu" "buick skylark 320" "plymouth satellite" "amc rebel sst"
...
```

```r
# If there is a non-numeric column (e.g., 'name'), remove it.
# Here we keep only numeric columns.
df <- df %>% select_if(is.numeric)

# For a typical auto-mpg dataset, assume 'mpg' is the response variable.
# Fit a linear regression model with mpg as response and the rest as predictors.
model <- lm(mpg ~ ., data = df)

# Display a summary of the model
summary(model)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5903 -2.1565 -0.1169  1.8690 13.0604
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
## cylinders     -0.493376   0.323282  -1.526  0.12780
## displacement   0.019896   0.007515   2.647  0.00844 **
## horsepower    -0.016951   0.013787  -1.230  0.21963
## weight        -0.006474   0.000652  -9.929  < 2e-16 ***
## acceleration   0.080576   0.098845   0.815  0.41548
## `model year`   0.750773   0.050973  14.729  < 2e-16 ***
## origin         1.426141   0.278136   5.127 4.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.328 on 384 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8182
## F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

```r
# Compute the Variance Inflation Factors (VIFs) for all predictors
vif_values <- vif(model)
print(vif_values)
```

```
##    cylinders displacement   horsepower       weight acceleration `model year`
##    10.737535    21.836792     9.943693    10.831260     2.625806     1.244952
##       origin
##     1.772386
```

Discussion of findings

- **High VIFs**: Variables such as `cylinders` , `displacement` , `horsepower` , and `weight` are often highly correlated. High VIF values (commonly above 5 or 10) for these predictors indicate strong multicollinearity. This suggests that their estimated coefficients may have inflated standard errors and be less reliable for inference.
- **Low VIFs**: Predictors such as `acceleration` , `model_year` , or `origin` may exhibit lower VIFs, implying they are less collinear with the other predictors.
- **Implication**: High multicollinearity does not affect prediction accuracy much but makes it difficult to assess the individual contribution of each predictor. Remedies might include removing or combining collinear variables.