資料分析方法

經研一 江彥亨 R11323040

WAY TO WIN

1.a. $\text{Cov}(\hat\beta_0, \hat\beta_1) = \text{Cov}(\bar y - \hat\beta_1 \bar X, \hat\beta_1)$

$= \underbrace{\text{Cov}(\bar y, \hat\beta_1)}_{=0} - \bar X \,\text{Var}(\hat\beta_1)$

$= -\bar X \,\text{Var}\left(\frac{\sum_{i=1}^n (X_i-\bar X) y_i}{\sum_{i=1}^n (X_i-\bar X)^2}\right) = -\bar X\,\text{Var}\left(\frac{\sum_{i=1}^n (X_i-\bar X)(\beta_0+\beta_1 X_i+u_i)}{\sum_{i=1}^n (X_i-\bar X)^2}\right)$

$= -\bar X\,\text{Var}\left(\beta_1 + \frac{\sum_{i=1}^n (X_i-\bar X) u_i}{\sum_{i=1}^n (X_i-\bar X)^2}\right)$

$= -\bar X \frac{\sum_{i=1}^n (X_i-\bar X)^2}{\left(\sum_{i=1}^n (X_i-\bar X)^2\right)^2} \cdot \text{Var}(u_i) = -\bar X \sigma^2 \cdot \frac{n}{\sum_{i=1}^n (X_i-\bar X)^2}$

$= -\bar X \sigma^2 / S_{xx}$ #

b. $\text{Cov}(\bar y, \hat\beta_1) = \text{Cov}(\hat\beta_0 + \hat\beta_1 \bar X, \hat\beta_1)$

$= \text{Cov}(\hat\beta_0, \hat\beta_1) + \bar X\,\text{Var}(\hat\beta_1)$

$= -\bar X\,\text{Var}(\hat\beta_1) + \bar X\,\text{Var}(\hat\beta_1)$

$= 0$ #

2. $SSR = \sum_{i=1}^n (\hat y_i - \bar y)^2 = \sum_{i=1}^n \hat y_i^2 - 2\hat y_i \bar y + \bar y^2$

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

$\sum_{i=1}^n \hat y_i = \sum_{i=1}^n \hat\beta X_i = \hat\beta \sum_{i=1}^n X_i = n \cdot \hat\beta \bar X$

Since least square regression curve must pass sample average $\bar y$ and $\bar X$ $\Rightarrow$ $\bar y = \hat\beta \bar X$

$\Rightarrow \sum_{i=1}^n \hat y_i = n\hat\beta \bar X = n\bar y$

$\Rightarrow SSR = \sum_{i=1}^n \hat y_i^2 - 2n\bar y + n\bar y = \sum_{i=1}^n \hat y_i^2 - n\bar y$ #

3.a $HH = X(X^TX)^{-1}X^T X(X^TX)^{-1}X^T = X(X^TX)^{-1}(X^TX)(X^TX)^{-1}X^T$

$= XI(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T = H$

$(I-H)(I-H) = II - HI - IH + HH = I - 2H + H = I - H$

3.b $V(\hat{y}) = V(X\hat{\beta}) = X\,Var(\hat{\beta})X^T$

$Var(\hat{\beta}) = Var((X^TX)^{-1}X^Ty) = (X^TX)^{-1}X^T\,Var(y)\,X(X^TX)^{-1}$

$= (X^TX)^{-1}X^T\sigma^2 I X(X^TX)^{-1}$

$\Rightarrow Var(\hat{y}) = X(X^TX)^{-1}X^T\sigma^2 I X(X^TX)^{-1}X^T = H\sigma^2 I H$

$= \sigma^2 HH = \sigma^2 H$

4. By definition, $R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$

and $\sum_{i=1}^{n}(y_i-\bar{y})^2$ can be divided into

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

$\Rightarrow R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}$

Thus, R^2 must be smaller than one. In addition, $\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$ must be larger or equal to zero, So R^2 must not be smaller than zero. Intuitively, R^2 means how much variance in the explained variable can be explained by the explanatory variables. Hence, R^2 must falls between zero and one.

5.

```python
y = data['mpg']
xvar = ['cylinders', 'displacement', 'horsepower', 'weight', 'acceleration', 'model year', 'origin']
X = data[xvar]
model = sm.OLS(y,X).fit()
VIFs = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
for x,vif in zip(xvar,VIFs):
    print(f'the VIFs of {x} is {vif}')
```

✓ 0.0s

```
the VIFs of cylinders is 117.70854741316116
the VIFs of displacement is 96.90976248526793
the VIFs of horsepower is 67.07215430169894
the VIFs of weight is 139.45416214259953
the VIFs of acceleration is 69.69976893381958
the VIFs of model year is 115.7946144892812
the VIFs of origin is 8.469941669334569
```

由結果可知，除 origin 變數外，其餘變數之膨脹因子 VIFs 皆非常大，故這些

變數具有高度共線性問題，因去除其中幾項較不顯著的變數。