Q1.

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{\sum(x_i - \overline{X})(y_i - \overline{Y})}{S_{XX}} = \frac{\sum(x_i y_i - x_i \overline{Y} - \overline{X} y_i + \overline{XY})}{S_{XX}}$$

$$= \frac{\sum(x_i y_i) - \sum(x_i \overline{Y}) - \sum(\overline{X} y_i) + \sum(\overline{XY})}{S_{XX}}$$

$$= \frac{\sum(x_i y_i) - n\overline{XY} - n\overline{XY} + n\overline{XY}}{S_{XX}} = \frac{\sum(x_i y_i) - n\overline{XY}}{S_{XX}} = \frac{\sum(x_i - \overline{X}) y_i}{S_{XX}}$$

$$\mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}\left(\frac{\sum(x_i - \overline{X}) y_i}{S_{XX}}\right) = \frac{[\sum(x_i - \overline{X})]^2}{(S_{XX})^2} \mathrm{Var}(y_i) = \frac{[\sum(x_i - \overline{X})]^2}{[\sum(x_i - \overline{X})^2]^2} \mathrm{Var}(y_i) = \frac{1}{S_{XX}} \sigma^2$$

(a).

$$\mathrm{cov}(\hat{\beta}_0, \hat{\beta}_1) = \mathrm{cov}(\overline{Y} - \hat{\beta}_1 \overline{X}, \hat{\beta}_1) = \mathrm{cov}(\overline{Y}, \hat{\beta}_1) - \mathrm{cov}(\hat{\beta}_1 \overline{X}, \hat{\beta}_1) = 0 - \overline{X}\mathrm{cov}(\hat{\beta}_1, \hat{\beta}_1)$$

$$= -\overline{X}\mathrm{Var}(\hat{\beta}_1) = \frac{-\overline{X}\sigma^2}{S_{XX}}$$

(b).

$$\mathrm{cov}(\overline{y}, \hat{\beta}_1) = \mathrm{cov}\left(\frac{\sum y_i}{n}, \frac{\sum(x_i - \overline{X}) y_i}{S_{XX}}\right) = \frac{1}{nS_{XX}} \mathrm{cov}\left(\sum y_i, \sum (x_i - \overline{X}) y_i\right)$$

$$= \frac{1}{nS_{XX}} cov(y_1 + y_2 + \cdots + y_n, (x_1 - \overline{X}) y_1 + (x_2 - \overline{X}) y_2 + \cdots + (x_n - \overline{X}) y_n)$$

$$= \frac{1}{nS_{XX}} [\mathrm{cov}(y_1, (x_1 - \overline{X}) y_1) + \mathrm{cov}(y_2, (x_2 - \overline{X}) y_2) + \cdots + \mathrm{cov}(y_n, (x_n - \overline{X}) y_n)]$$

$$= \frac{\sum \mathrm{cov}(y_i, (x_i - \overline{X}) y_i)}{nS_{XX}} \Rightarrow \frac{\sum(x_i - \overline{X})\mathrm{cov}(y_i, y_i)}{nS_{XX}} = \frac{(n\overline{X} - n\overline{X})\sigma^2}{nS_{XX}} = \frac{0 \times \sigma^2}{nS_{XX}} = 0$$

Q2.

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

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

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

Q3.

(a).

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

$$HH = [X(X^T X)^{-1} X^T][X(X^T X)^{-1} X^T] = X(X^T X)^{-1} X^T X (X^T X)^{-1} X^T$$

$$= X[(X^T X)^{-1}(X^T X)](X^T X)^{-1} X^T = X(X^T X)^{-1} X^T \quad \because [(X^T X)^{-1}(X^T X)] = 1$$

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

$$(1 - H)^2 = (1 - H)(1 - H) = I - 2H + H^2 = I - 2H + H \quad \therefore H^2 = HH = 1$$

$$= 1 - H$$

According to (a) and (b), H is idempotent.

(b).

$$Define\ \hat{Y} = X\left(X^T X\right)^{-1} X^T Y = HY$$

$$Var\left(\hat{Y}\right) = Var(HY) = HH \times Var(Y) = H \times Var(Y)$$

$$\because Y = x\beta + \varepsilon \ \text{, } x\beta \text{ is a constant } \text{, } \varepsilon \sim N(0, \sigma^2)$$

$$\therefore Var(Y) = Var(x\beta + \varepsilon) = Var(x\beta + \varepsilon) = Var(\varepsilon) = \sigma^2$$

$$\Rightarrow Var\left(\hat{Y}\right) = H \times Var(Y) = H\sigma^2$$

Q4.

$$R^2 = \frac{SS_{Regression}}{SS_{Total}} = \frac{SS_R}{SS_{TO}}$$

$$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \text{, } \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$SS_{TO} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \text{, } y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Although we estimate the $\beta_0$and β_1 with a close or equal estimator $\hat{\beta}_0$and $\hat{\beta}_1$, we still cannot estimate the error correctly. Thus y_i will always larger than $\hat{y}_i$, by $\sum_{i=1}^{n} \varepsilon_i{}^2$. Hence, $(SS_{TO} - SS_R)$ must be a positive number smaller than SS_{TO}, leading R^2 smaller than 1.

Q5

X_i	VIF	R_j^2
CYLINDERS	10.738	0.907
DISPLACEMENT	21.837	0.954
HORSEPOWER	9.944	0.899
WEIGHT	10.831	0.908
ACCELERATION	2.626	0.619
MODEL YEAR	1.245	0.197
ORIGIN	1.772	0.436

While building a multiple liner regression model, we need to check whether model has collinearity(共線性) problem with, which leads to certain variables increase predictive power between each other, and reduce of the model accuracy.

In the chart we can see that cylinders, displacement, weight have VIF value larger than 10, namely, they are highly correlate to each other (not independent), and makes them have better predictive power than the other variables. We can combine them into one variable, or deleting two of them instead.

DA_HW_04

March 28, 2021

0.0.1 Q5

[7]:
```
import pandas as pd
from statsmodels.stats.outliers_influence import variance_inflation_factor
import statsmodels.api as sm
```

[8]:
```
#read the dataset
data = pd.read_csv(r"C:\Users\TerryYang\pythonwork\pythonwork\Data Analytics␣
↪Homework\DA_Demo.csv")
#variables column
X = data[['origin', 'model year', 'acceleration', 'weight', 'horsepower',␣
↪'displacement', 'cylinders']]
#add constant
X = sm.add_constant(X)
# VIF dataframe
vif_data = pd.DataFrame()
vif_data["Variables"] = X.columns
# calculating VIF for each feature
vif_data["VIF"] = [variance_inflation_factor(X.values, i) for i in range(len(X.
↪columns))]
print(vif_data)
```

```
      Variables         VIF
0         const  763.557531
1        origin    1.772386
2    model year    1.244952
3  acceleration    2.625806
4        weight   10.831260
5    horsepower    9.943693
6  displacement   21.836792
7     cylinders   10.737535
```