Regression analysis_Homework Assignment 3

心理所碩二 R08227112 林子堯

2020/10/19

1. Consider the following two models where $E(\boldsymbol{\varepsilon}) = 0$ and $Var(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}$:

- **Model A: $\boldsymbol{y} = \boldsymbol{X_1\beta_1} + \boldsymbol{\varepsilon}$**
- **Model B: $\boldsymbol{y} = \boldsymbol{X_1\beta_1} + \boldsymbol{X_2\beta_2} + \boldsymbol{\varepsilon}$**

Show that $R_A^2 \leq R_B^2$.

By the definition, the R^2 is

$$
\begin{aligned}
R^2 = \frac{SSR}{SST} &= 1 - \frac{SSE}{SST} \\
&= 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\
&= 1 - \frac{\hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}}}{\boldsymbol{yCy}}
\end{aligned}
$$

where $\boldsymbol{C} = \boldsymbol{I} - \frac{1}{n}\boldsymbol{11}^\top$. Let $\hat{\boldsymbol{\beta}}_{\boldsymbol{1A}} = \arg\min_{\boldsymbol{\beta_1}} \|\boldsymbol{y} - \boldsymbol{x_1\beta_1}\|^2$ and $(\hat{\boldsymbol{\beta}}_{\boldsymbol{1B}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{2B}})^\top = \arg\min_{\boldsymbol{\beta_1},\boldsymbol{\beta_2}} \|\boldsymbol{y} - (\boldsymbol{x_1\beta_1} - \boldsymbol{x_2\beta_2})\|^2$. By the definition of least square estimation, one has

$$
\begin{aligned}
\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{B}}^\top \hat{\boldsymbol{\varepsilon}}_{\boldsymbol{B}} &= \min_{(\boldsymbol{\beta_1},\boldsymbol{\beta_2})^\top} \|\boldsymbol{y} - (\boldsymbol{x_1\beta_1} + \boldsymbol{x_2\beta_2})\|^2 \\
&\leq \|\boldsymbol{y} - (\boldsymbol{x_1\beta_1} + \boldsymbol{x_2 0})\|^2
\end{aligned}
$$

for any $\boldsymbol{\beta_1}$ in the last part of the inequation. Therefore one has

$$
\begin{aligned}
\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{B}}^\top \hat{\boldsymbol{\varepsilon}}_{\boldsymbol{B}} &= \left\|\boldsymbol{y} - (\boldsymbol{x_1}\hat{\boldsymbol{\beta}}_{\boldsymbol{1B}} + \boldsymbol{x_2}\hat{\boldsymbol{\beta}}_{\boldsymbol{2B}})\right\|^2 \\
&\leq \left\|\boldsymbol{y} - (\boldsymbol{x_1}\hat{\boldsymbol{\beta}}_{\boldsymbol{1A}})\right\|^2 = \hat{\boldsymbol{\varepsilon}}_{\boldsymbol{A}}^\top \hat{\boldsymbol{\varepsilon}}_{\boldsymbol{A}}
\end{aligned}
$$

$$
\Rightarrow R_A^2 = 1 - \frac{\hat{\boldsymbol{\varepsilon_A}}^\top \hat{\boldsymbol{\varepsilon_A}}}{\boldsymbol{yCy}} \leq 1 - \frac{\hat{\boldsymbol{\varepsilon_B}}^\top \hat{\boldsymbol{\varepsilon_B}}}{\boldsymbol{yCy}} = R_B^2
$$

2. Suppose we need to compare the effects of two drugs each administered to n subjects. The model for the effect of the first drug is

$$
y_{i1} = \beta_0 + \beta_1 x_{i1} + \varepsilon_{i1}
$$

while for the second drug it is

$$
y_{i2} = \beta_0 + \beta_2 x_{i2} + \varepsilon_{i2}
$$

and in each case $i = 1, \ldots, n$ and $\bar{x}_1 = \bar{x}_2 = 0$. Assume that all observations are independent and that for each i both ε_{1i} and ε_{2i} are normally distributed with mean 0 and variance σ^2.

a. Obtain the least squares estimator for $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^\top$ and its covariance matrix.

Since β_0's in both models are equal, I convert them to a multiple linear regression model,

$$
\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}
$$

where

$$
\boldsymbol{y}=\begin{pmatrix} y_{11}\\ \vdots\\ y_{n1}\\ y_{12}\\ \vdots\\ y_{n2}\end{pmatrix},\quad \boldsymbol{X}=\begin{pmatrix} 1 & x_{11} & 0\\ \vdots & \vdots & \vdots\\ 1 & x_{n1} & 0\\ 1 & 0 & x_{12}\\ \vdots & \vdots & \vdots\\ 1 & 0 & x_{n2}\end{pmatrix},\quad \boldsymbol{\beta}=\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\end{pmatrix},\ \text{and}\ \boldsymbol{\varepsilon}=\begin{pmatrix} \varepsilon_{11}\\ \vdots\\ \varepsilon_{n1}\\ \varepsilon_{12}\\ \vdots\\ \varepsilon_{n2}\end{pmatrix}\sim N_{2n}(\boldsymbol{0},\sigma^2\boldsymbol{I})
$$

The least squares estimator for $\boldsymbol{\beta}$ is

$$
\hat{\boldsymbol{\beta}}=(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}
$$

and the covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$
\begin{aligned}
Cov(\hat{\boldsymbol{\beta}}) &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top Cov(\boldsymbol{y})((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)^\top\\
&=\sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}
\end{aligned}
$$

b. Estimate σ^2.

The residual is $\hat{\boldsymbol{\varepsilon}}=\boldsymbol{y}-\hat{\boldsymbol{y}}=\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}=(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}$ and the expectation of sum of square residual is

$$
\begin{aligned}
E(\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}) &= E(\boldsymbol{y}^\top(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y})\\
&= E(\boldsymbol{y}^\top)(\boldsymbol{I}-\boldsymbol{H})E(\boldsymbol{y})+trace((\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Var}(\boldsymbol{y}))\\
&= (\boldsymbol{X\beta})^\top(\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{X\beta})+trace((\boldsymbol{I}-\boldsymbol{H})\sigma^2\boldsymbol{I})\\
&= (2n-3)\sigma^2
\end{aligned}
$$

so, we can let $\frac{\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}}{2n-3}$ be an unbiased estimator of σ^2.

c. Write the test statistic for testing $\beta_1=\beta_2$ against the alternative that $\beta_1\neq\beta_2$.

The hypothesis test can rewrite as

$$
\boldsymbol{C\beta}=\boldsymbol{d}\quad vs.\boldsymbol{C\beta}\neq\boldsymbol{d}
$$

where $\boldsymbol{C}=\begin{pmatrix}0 & 1 & -1\end{pmatrix}$ with $rank(\boldsymbol{C})=1$ and $\boldsymbol{d}=0$

The full model is

$$
\boldsymbol{y}=\beta_0+\boldsymbol{x_1}\beta_1+\boldsymbol{x_2}\beta_2+\boldsymbol{\varepsilon}
$$

Let the least square estimator of $\boldsymbol{beta}$ be denoted as $\hat{\boldsymbol{\beta}}$. Then we know $SSE=\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}\sim\sigma^2\chi^2_{2n-3}$.

On the other side, the reduced model (under $H_0:\beta_1=\beta_2$) is

$$
\boldsymbol{y}=\beta_0+(\boldsymbol{x_1}+\boldsymbol{x_2})\beta_1+\boldsymbol{\varepsilon}
$$

the least square estimate of $\boldsymbol{beta}$ is $\hat{\boldsymbol{\beta}}_{\boldsymbol{H_0}}=\hat{\boldsymbol{\beta}}-(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{C}^\top(\boldsymbol{C}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{C}^\top)^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}}-\boldsymbol{d})$ and the residual is $\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{H_0}}=\boldsymbol{y}-\hat{\boldsymbol{y}}_{\boldsymbol{H_0}}=\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{H_0}}$

By some calculation, one can get

$$
\begin{aligned}
\Delta SSE=SSE-SSE_{H_0} &= \hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}-\hat{\boldsymbol{\varepsilon}}^\top_{\boldsymbol{H_0}}\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{H_0}}\\
&=(\boldsymbol{C}\hat{\boldsymbol{\beta}}-\boldsymbol{d})^\top(\boldsymbol{C}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{C}^\top)^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}}-\boldsymbol{d})
\end{aligned}
$$

and under $H_0:\boldsymbol{C\beta}=\boldsymbol{d}$, $\Delta SSE\sim\sigma^2\chi^2_{r=1}$.

Furthermore, SSE is independent of ΔSSE, so we can use the test statistic

$$
F=\frac{\Delta SSE/1}{SSE/(2n-3)}=\frac{\frac{\Delta SSE}{\sigma^2}/1}{\frac{SSE}{\sigma^2}/(2n-3)}\sim F_{1,2n-3}
$$

We could reject H_0 if the test statistic F value is greater then the critical value $F_{1,2n-3}(\alpha)$, where α is significance level. Otherwise, we would retain H_0.

3. Consider the two models $\boldsymbol{y_1} = \boldsymbol{X_1}\beta_1 + \boldsymbol{\varepsilon_1}$ and $\boldsymbol{y_2} = \boldsymbol{X_2}\beta_2 + \boldsymbol{\varepsilon_2}$ where the $\boldsymbol{X_i}$'s are $n_i \times p$ matrices. Suppose that $\boldsymbol{\varepsilon_i} \sim N(\boldsymbol{0}, \sigma_i^2\boldsymbol{I})$ where $i = 1, 2$ and that $\boldsymbol{\varepsilon_1}$ and $\boldsymbol{\varepsilon_2}$ are independent.

a. Assuming that the σ_i's are known, obtain a test for the hypothesis $\boldsymbol{\beta_1} = \boldsymbol{\beta_2}$.

From the least square estimation, we have

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{1}} = (\boldsymbol{X_1^\top X_1})^{-1}\boldsymbol{X_1^\top y_1} \quad \& \quad \hat{\boldsymbol{\beta}}_{\boldsymbol{2}} = (\boldsymbol{X_2^\top X_2})^{-1}\boldsymbol{X_2^\top y_2}$$

and since $\boldsymbol{y_i} \sim N_p(X_i\beta_i, \sigma_i^2 I)$ and $\boldsymbol{\beta_i}$ is a linear combination of $\boldsymbol{y_i}$ for $i = 1, 2$, so

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{1}} \sim N_p\left(\boldsymbol{\beta_1}, \sigma_1^2(\boldsymbol{X_1^\top X_1})^{-1}\right) \quad \& \quad \hat{\boldsymbol{\beta}}_{\boldsymbol{2}} \sim N_p\left(\boldsymbol{\beta_2}, \sigma_2^2(\boldsymbol{X_2^\top X_2})^{-1}\right)$$

Otherwise, $\boldsymbol{y_1}$ are mutually independent with $\boldsymbol{y_2}$, therefore $Cov(\boldsymbol{Ay_1}, \boldsymbol{By_2}) = \boldsymbol{A}Cov(\boldsymbol{y_1}, \boldsymbol{y_2})\boldsymbol{B}^\top = \boldsymbol{0}$ for any matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. We can get

$$\begin{aligned}
&\hat{\boldsymbol{\beta}}_{\boldsymbol{1}} - \hat{\boldsymbol{\beta}}_{\boldsymbol{2}} \sim N_p\left((\boldsymbol{\beta_1} - \boldsymbol{\beta_2}), (\sigma_1^2(\boldsymbol{X_1^\top X_1})^{-1} + \sigma_2^2(\boldsymbol{X_2^\top X_2})^{-1})\right) \\
\Rightarrow \quad & \frac{(\hat{\boldsymbol{\beta}}_{\boldsymbol{1}} - \hat{\boldsymbol{\beta}}_{\boldsymbol{2}}) - (\boldsymbol{\beta_1} - \boldsymbol{\beta_2})}{\sqrt{\sigma_1^2(\boldsymbol{X_1^\top X_1})^{-1} + \sigma_2^2(\boldsymbol{X_2^\top X_2})^{-1}}} \sim N_p\left(\boldsymbol{0}, \boldsymbol{I}\right) \\
\Rightarrow \quad & \frac{[(\hat{\boldsymbol{\beta}}_{\boldsymbol{1}} - \hat{\boldsymbol{\beta}}_{\boldsymbol{2}}) - (\boldsymbol{\beta_1} - \boldsymbol{\beta_2})]^\top [(\hat{\boldsymbol{\beta}}_{\boldsymbol{1}} - \hat{\boldsymbol{\beta}}_{\boldsymbol{2}}) - (\boldsymbol{\beta_1} - \boldsymbol{\beta_2})]}{\sigma_1^2(\boldsymbol{X_1^\top X_1})^{-1} + \sigma_2^2(\boldsymbol{X_2^\top X_2})^{-1}} \sim \chi_p^2
\end{aligned}$$

When the the σ_i's are known, the test statistic for $\boldsymbol{\beta_1} = \boldsymbol{\beta_2}$ is $[(\hat{\boldsymbol{\beta}}_{\boldsymbol{1}} - \hat{\boldsymbol{\beta}}_{\boldsymbol{2}})]^\top[\sigma_1^2(\boldsymbol{X_1^\top X_1})^{-1} + \sigma_2^2(\boldsymbol{X_2^\top X_2})^{-1}]^{-1}[(\hat{\boldsymbol{\beta}}_{\boldsymbol{1}} - \hat{\boldsymbol{\beta}}_{\boldsymbol{2}})]$, which is distributed at chi-square distribution with $df = p$. We could reject the null hypothesis $H_0 : \boldsymbol{\beta_1} = \boldsymbol{\beta_2}$ if the the test statistic is greater then $\chi_p^2(\alpha)$, where α is significance level. Otherwise, we would retain H_0.

b. Assume that $\sigma_1 = \sigma_2$ but they are unknown. Derive a test for the hypothesis $\beta_1 = \beta_2$.

Let $\sigma = \sigma_1 = \sigma_2$ are unknown, the pooled sample variance is

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{1}}^\top\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{1}} + \hat{\boldsymbol{\varepsilon}}_{\boldsymbol{2}}^\top\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{2}}}{n_1 + n_2 - 2} = \frac{\boldsymbol{y_1^\top(I - H_1)y_1 + y_2^\top(I - H_2)y_2}}{n_1 + n_2 - 2}$$

since $\frac{\boldsymbol{y_i^\top(I-H_i)y_i}}{\sigma^2} \sim \chi_{ni-p}^2$ for $i = 1, 2$ and y_{i1}'s are independent of y_{i2}'s, one has

$$\begin{aligned}
&\frac{\boldsymbol{y_1^\top(I - H_1)y_1 + y_2^\top(I - H_2)y_2}}{\sigma^2} \sim \chi_{n1+n2-2p}^2 \\
\Rightarrow \quad & \frac{(n_1 + n_2 - 2)s_p^2}{\sigma^2} \sim \chi_{n1+n2-2p}^2
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&E\left(\frac{(n_1 + n_2 - 2)s_p^2}{\sigma^2}\right) = n_1 + n_2 - 2p \\
\Rightarrow \quad & E(s_p^2) = \frac{\sigma^2}{n_1 + n_2 - 2p}(n_1 + n_2 - 2p) = \sigma^2
\end{aligned}$$

so, the pooled sample variance is a unbiased estimator for σ^2.

Combine the result form part a., the random vector

$$\frac{(\hat{\boldsymbol{\beta}}_{\boldsymbol{1}} - \hat{\boldsymbol{\beta}}_{\boldsymbol{2}}) - (\boldsymbol{\beta_1} - \boldsymbol{\beta_2})}{s_p\sqrt{(\boldsymbol{X_1^\top X_1})^{-1} + (\boldsymbol{X_2^\top X_2})^{-1}}} = \frac{\frac{(\hat{\boldsymbol{\beta}}_{\boldsymbol{1}} - \hat{\boldsymbol{\beta}}_{\boldsymbol{2}}) - (\boldsymbol{\beta_1} - \boldsymbol{\beta_2})}{\sqrt{\sigma^2(\boldsymbol{X_1^\top X_1})^{-1} + \sigma^2(\boldsymbol{X_2^\top X_2})^{-1}}}}{\sqrt{\frac{(n_1 + n_2 - 2p)s_p^2/\sigma^2}{n_1 + n_2 - 2p}}} \sim t_{n1+n2-2p}$$

where $t_{n_1+n_2-2p}$ is a p-variate t distribution with $df = n_1 + n_2 - 2p$. Furthermore, we sum of square the random vector in the above equation and divide it by p, we get

$$\frac{[(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2) - (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)]^\top [(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2) - (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)]}{ps_p^2((\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} + (\boldsymbol{X}_2^\top \boldsymbol{X}_2)^{-1})} = \frac{\frac{[(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2) - (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)]^\top [(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2) - (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)]}{\sigma^2(\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} + \sigma^2(\boldsymbol{X}_2^\top \boldsymbol{X}_2)^{-1}} / p}{\frac{(n_1+n_2-2p)s_p^2/\sigma^2}{n_1+n_2-2p}} \sim F_{p, n_1+n_2-2p}$$

where the numerator is distributed at χ_p^2/p and the denominator is distributed at $\chi_{n_1+n_2-2p}$. When $\sigma_1 = \sigma_2$ is unknown, the test statistic for $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ is $\frac{1}{pS_p^2}[(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2)]^\top [(\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} + (\boldsymbol{X}_2^\top \boldsymbol{X}_2)^{-1})]^{-1}[(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2)]$, which is distributed at F distribution with $df_1 = p$ and $df_2 = n_1 + n_2 - 2p$. In conclusion, We could reject the null hypothesis $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ if the the test statistic is greater then $F_{p,n_1+n_2-2p}(\alpha)$, where α is significance level. Otherwise, we would retain H_0.

4. Moore (1975) reported the results of an experiment to construct a model for total oxygen demand in dairy wastes as a function of five laboratory measurements (Data is attached in the mail). Data were collected on samples kept in suspension in water in a laboratory for 220 days. Although all observations reported here were taken on the same sample over time, assume that they are independent. The measured variables are:

- **y log(oxygen demand, mg oxygen per minute)**
- **x_1 biological oxygen demand, mg/liter**
- **x_2 total Kjeldahl nitrogen, mg/liter**
- **x_3 total solids, mg/liter**
- **x_4 total volatile solids, a component of x_3, mg/liter**
- **x_5 chemical oxygen demand, mg/liter**

a. Fit a multiple regression model using y as the dependent variable and all x_j's as the independent variables.

First of all, we should load the data to R

```
data ← readxl::read_excel("E3.7.xlsx", col_names =TRUE)
knitr::kable(data)
```

Day	x.1	x.2	x.3	x.4	x.5	y
0	1125	232	7160	85.9	8905	1.5563
7	920	268	8804	86.5	7388	0.8976
15	835	271	8108	85.2	5348	0.7482
22	1000	237	6370	83.8	8056	0.7160
29	1150	192	6441	82.1	6960	0.3130
37	990	202	5154	79.2	5690	0.3617
44	840	184	5896	81.2	6932	0.1139
58	650	200	5336	80.6	5400	0.1139
65	640	180	5041	78.4	3177	-0.2218
72	583	165	5012	79.3	4461	-0.1549
80	570	151	4825	78.7	3901	0.0000
86	570	171	4391	78.0	5002	0.0000
93	510	243	4320	72.3	4665	-0.0969
100	555	147	3709	74.9	4642	-0.2218
107	460	286	3969	74.4	4840	-0.3979
122	275	198	3558	72.5	4479	-0.1549
129	510	196	4361	57.7	4200	-0.2218
151	165	210	3301	71.8	3410	-0.3979

Day	x.1	x.2	x.3	x.4	x.5	y
171	244	327	2964	72.5	3360	-0.5229
220	79	334	2777	71.9	2599	-0.0458

Our model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \varepsilon_i$$

where ε_i's are mutually independent with mean 0 and variance σ^2 for all i.

I use the `lm()` function to fit multiple regression model in R

```
full_lm <- lm(y ~ 1 + x.1 + x.2 + x.3 + x.4 + x.5, data)
summary(full_lm)
```

```


Call:
lm(formula = y ~ 1 + x.1 + x.2 + x.3 + x.4 + x.5, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.39447 -0.11847  0.00053  0.08313  0.56232 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)  
(Intercept) -2.156e+00  9.135e-01  -2.360   0.0333 *
x.1         -9.012e-06  5.184e-04  -0.017   0.9864  
x.2          1.316e-03  1.263e-03   1.041   0.3153  
x.3          1.278e-04  7.690e-05   1.662   0.1188  
x.4          7.899e-03  1.400e-02   0.564   0.5815  
x.5          1.417e-04  7.375e-05   1.921   0.0754 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2618 on 14 degrees of freedom
Multiple R-squared:  0.8107,	Adjusted R-squared:  0.743 
F-statistic: 11.99 on 5 and 14 DF,  p-value: 0.0001184
```

then the fitted model is

$$\begin{aligned}\hat{E}(y_i|x_{i1},\ldots,x_{i5}) = \hat{y}_i =& -2.16 - 9.01\times 10^{-6}x_{i1} + 1.32\times 10^{-3}x_{i2} \\ &+ 1.28\times 10^{-4}x_{i3} + 7.90\times 10^{-3}x_{i4} + 1.42\times 10^{-4}x_{i5}\end{aligned}$$

b. Now fit a regression model with only the independent variables x_3 and x_5. How do the new parameters, the corresponding value of R^2 and the t-values compare with those obtained from the full model?

Our reduced model now is

$$y_i = \beta_0 + \beta_3 x_{i3} + \beta_5 x_{i5} + \varepsilon_i$$

```
reduced_lm <- lm(y ~ 1 + x.3 + x.5, data)
summary(reduced_lm)
```

```
Call:
lm(formula = y ~ 1 + x.3 + x.5, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.37768 -0.09357 -0.04241  0.06230  0.59623 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.371e+00  1.963e-01  -6.988 2.19e-06 ***
x.3          1.492e-04  5.473e-05   2.726   0.0144 *  
x.5          1.419e-04  5.302e-05   2.676   0.0160 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2519 on 17 degrees of freedom
Multiple R-squared:  0.7872,    Adjusted R-squared:  0.7621 
F-statistic: 31.44 on 2 and 17 DF,  p-value: 1.942e-06
```

then the fitted model is

$$\hat{E}(y_i|x_{i3}, x_{i5}) = \hat{y}_i = -1.37 + 1.49 \times 10^{-4} x_{i3} + 1.42 \times 10^{-4} x_{i5}$$

The coefficient of determination in the reduced model ($R^2 = 0.79$) is smaller the the full model ($R^2 = 0.81$). The t-value of β_0 is smaller in the reduced model and the t-value of β_3 & β_5 is greater in the reduced model. However, there have smaller p-value in the reduced model than in the full model, so triple of them are significance in the t test (under the significance level $\alpha = 0.05$).

5. Consider the data given in 4. Suppose the model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \varepsilon_i$$

where $i = 1, \ldots, n$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top \sim N(0, \sigma^2 I_n)$.

a. Test the hypothesis $\beta_2 = \beta_4 = 0$ at the 5 per cent level of significance.

Our reduced model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_3 x_{i3} + \beta_5 x_{i5} + \varepsilon_i$$

```
reduced2_lm <- lm(y ~ 1 + x.1 + x.3 + x.5, data)
summary(reduced2_lm)
```

```
Call:
lm(formula = y ~ 1 + x.1 + x.3 + x.5, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.38529 -0.10424 -0.03769  0.03625  0.58651 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.438e+00  2.353e-01  -6.111  1.5e-05 ***
x.1         -2.416e-04  4.472e-04  -0.540   0.5965    
x.3          1.683e-04  6.613e-05   2.544   0.0216 *  
x.5          1.656e-04  6.975e-05   2.375   0.0304 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2573 on 16 degrees of freedom
Multiple R-squared:  0.791, Adjusted R-squared:  0.7518 
F-statistic: 20.18 on 3 and 16 DF,  p-value: 1.097e-05
```

We can use the F test to test the hypothesis

```
anova(reduced2_lm, full_lm)
```

```
Analysis of Variance Table

Model 1: y ~ 1 + x.1 + x.3 + x.5
Model 2: y ~ 1 + x.1 + x.2 + x.3 + x.4 + x.5
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     16 1.05927                            
2     14 0.95953  2  0.099733 0.7276 0.5005
```

At the significance level $\alpha = 0.05$, the p-value of the F test is $0.50 > \alpha$, fail to reject $H_0 : \beta_2 = \beta_4 = 0$. It shows that at least one of β_2 and β_4 is not equal to 0. From the above result, therefore, we would retain the original full model $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \varepsilon_i$.

b. Find a 95 per cent C.I. for β_1.

c. Find a 95 per cent C.I. for $\beta_3 + 2\beta_5$.

I calculate both of 95% C.I for β_1 and $\beta_3 + 2\beta_5$ at the same time. The R code is as follows

```
library(multcomp)
contrast <-  rbind("β1" = c(0, 1, 0, 0, 0, 0),
                   "β3+2*β5" = c(0, 0, 0, 1, 0, 2))
full_glht <- glht(full_lm, linfct = contrast)
summary(full_glht)
```

```
	 Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = y ~ 1 + x.1 + x.2 + x.3 + x.4 + x.5, data = data)

Linear Hypotheses:
               Estimate Std. Error t value Pr(>|t|)
β1 == 0       -9.012e-06  5.184e-04  -0.017    1.000
β3+2*β5 == 0  4.111e-04  1.642e-04   2.504    0.039 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

```
full_confint <- confint(full_glht,
                        level = 0.95,
                        calpha = univariate_calpha()) # specify univariate confiden
ce intervals
full_confint
```

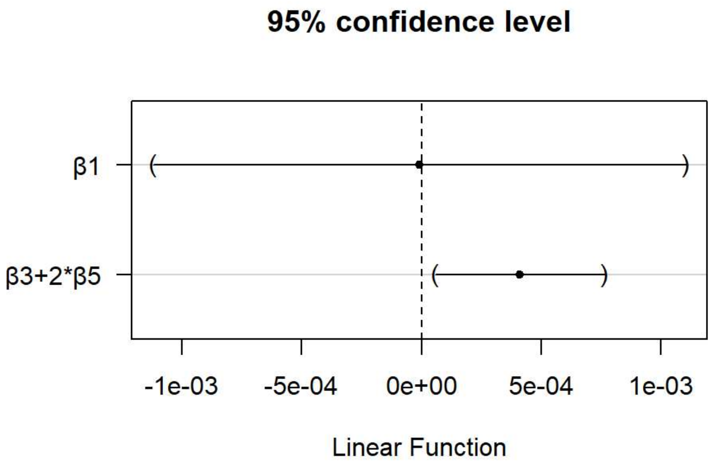
```
	 Simultaneous Confidence Intervals

Fit: lm(formula = y ~ 1 + x.1 + x.2 + x.3 + x.4 + x.5, data = data)

Quantile = 2.1448
95% confidence level


Linear Hypotheses:
             Estimate   lwr        upr
β1 == 0      -9.012e-06 -1.121e-03  1.103e-03
β3+2*β5 == 0  4.111e-04  5.899e-05  7.632e-04
```

```
plot(full_confint)
```



In part b., the estimate of β_0 is -9.01×10^{-6} and the 95% C.I. is $[-1.12 \times 10^{-3}, 1.10 \times 10^{-3}]$. Otherwise, the 95% C.I contains 0, so we fail to reject the null hypothesis of $\beta_0 = 0$.

In part c., the estimate of $\beta_3 + 2\beta_5$ is 4.11×10^{-4} and the 95% C.I. is $[5.90 \times 10^{-5}, 7.63 \times 10^{-4}]$. On the contrary, this 95% C.I does not contained 0, so we can reject the null hypothesis of $\beta_3 + 2\beta_5 = 0$.

Note: In part b. and c., the significance level of per contrast is 0.05. But we test both of tests simultaneously, actually, we should consider the family-wise significance level ($\alpha_{FW} = 0.05$) then refine the significance level of per contrast ($\alpha_{PC} < 0.05$) in each test.