Compulsory Group HWA 1 - Regression Analysis

April 26, 2024

1. Problem 2.8 of textbook
Solution: For demeaned covariates, the intercept α means the mean of the response variable. Consider the model $y_i = \beta_0 + \beta_1 x_i + e_i$. The OLS estimator is

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}. \end{aligned}$$

When the covariates are demeaned, the covariate becomes $x_i - \bar{x}$. Hence,

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

The OLS estimators satisfy

$$\begin{aligned} \mathrm{Var}\left(\hat{\beta}_0 \mid \boldsymbol{X}\right) &= \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right], \\ \mathrm{Var}\left(\hat{\beta}_1 \mid \boldsymbol{X}\right) &= \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \\ \mathrm{Cov}\left(\hat{\beta}_0, \hat{\beta}_1 \mid \boldsymbol{X}\right) &= -\frac{\sigma^2 \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}. \end{aligned}$$

With $x_i - \bar{x}$ as our new covariate, we get

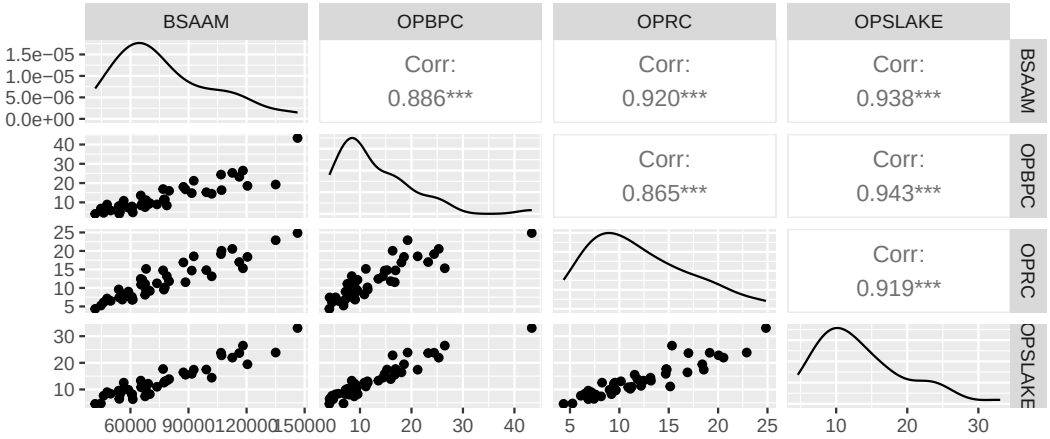
$$\begin{aligned} \mathrm{Var}\left(\hat{\alpha} \mid \boldsymbol{X}\right) &= \frac{1}{n}\sigma^2, \\ \mathrm{Var}\left(\hat{\beta}_1 \mid \boldsymbol{X}\right) &= \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \\ \mathrm{Cov}\left(\hat{\alpha}, \hat{\beta}_1 \mid \boldsymbol{X}\right) &= 0. \end{aligned}$$

2. Problem 3.6 of textbook
Solution: The scatter plot is

```
data(water, package = "alr4")
library(GGally)

## Loading required package:  ggplot2
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2

ggpairs(water[, c("BSAAM", "OPBPC", "OPRC", "OPSLAKE")])
```



The scatter plot shows quite strong positive correlations, which is in line with the correlation coefficients in the output. The regression results are

```
LR <- lm(BSAAM ~ OPBPC + OPRC + OPSLAKE, data = water)
summary(LR)

##
## Call:
## lm(formula = BSAAM ~ OPBPC + OPRC + OPSLAKE, data = water)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15964.1  -6491.8   -404.4   4741.9  19921.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 22991.85    3545.32   6.485  1.1e-07 ***
## OPBPC          40.61     502.40   0.081  0.93599
## OPRC         1867.46     647.04   2.886  0.00633 **
## OPSLAKE      2353.96     771.71   3.050  0.00410 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 8304 on 39 degrees of freedom
## Multiple R-squared:  0.9017,Adjusted R-squared:  0.8941
## F-statistic: 119.2 on 3 and 39 DF,  p-value: < 2.2e-16
```

The column of "t value" computes the value of the t statistic given by

$$\frac{\hat{\beta}_j / \sqrt{\left[(\boldsymbol{X}^T \boldsymbol{X})^{-1}\right]_{jj}}}{\sqrt{\hat{\boldsymbol{e}}^T \hat{\boldsymbol{e}} / (n-p)}},$$

which follows $t(n-p)$ if $\beta_j = 0$.

3. Problem 4.13 of textbook
Solution: We first define the variable log(perCapitaUse). There are so many ways of doing so.

```
data(MinnWater, package = "alr4")
## Alternative 1
MinnWater$logperCapitaUse <- log(10 ^ 6 * MinnWater$muniUse / MinnWater$muniPop)
## Alternative 2
MinnWater$logperCapitaUse <- with(MinnWater, log(10 ^ 6 * muniUse / muniPop))
## Alternative 3
library(dplyr)

##
## Attaching package:  'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

MinnWater <- MinnWater %>%
             mutate(logperCapitaUse = log(10 ^ 6 * muniUse / muniPop))
```

To fit the linear regression models, we use

```
LR1 <- lm(logperCapitaUse ~ year, data = MinnWater)
LR2 <- lm(logperCapitaUse ~ year + muniPrecip, data = MinnWater)
summary(LR1)

##
## Call:
## lm(formula = logperCapitaUse ~ year, data = MinnWater)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
```

```
## -0.115920 -0.033955  0.004258  0.038805  0.103466
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.6179787  3.7168607   0.973    0.341
## year        0.0000563  0.0018589   0.030    0.976
##
## Residual standard error: 0.06304 on 22 degrees of freedom
## Multiple R-squared:  4.17e-05,Adjusted R-squared:  -0.04541
## F-statistic: 0.0009174 on 1 and 22 DF,  p-value: 0.9761
```

```
summary(LR2)
```

```
##
## Call:
## lm(formula = logperCapitaUse ~ year + muniPrecip, data = MinnWater)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.09766 -0.03057  0.01086  0.02871  0.07577
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.5040361  2.5925575   1.352    0.191
## year         0.0002155  0.0012969   0.166    0.870
## muniPrecip  -0.0102590  0.0020845  -4.922 7.21e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04397 on 21 degrees of freedom
## Multiple R-squared:  0.5356,Adjusted R-squared:  0.4914
## F-statistic: 12.11 on 2 and 21 DF,  p-value: 0.0003176
```

In both models, year does not have a significant effect on the logperCapitaUse.

4. Problem 6.9 of textbook
Solution: To test $H_0 : \beta_5 = 0$ versus $H_1 : \beta_5 \neq 0$, we can use the t test from

```
data(cakes, package = "alr4")
LR1 <- lm(Y ~ X1 + I(X1 ^ 2) + X2  + I(X2 ^ 2) + I(X1 * X2), data = cakes)
summary(LR1)
```

```
##
## Call:
## lm(formula = Y ~ X1 + I(X1^2) + X2 + I(X2^2) + I(X1 * X2), data = cakes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4912 -0.3080  0.0200  0.2658  0.5454
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.204e+03  2.416e+02  -9.125 1.67e-05 ***
## X1           2.592e+01  4.659e+00   5.563 0.000533 ***
## I(X1^2)     -1.569e-01  3.945e-02  -3.977 0.004079 **
## X2           9.918e+00  1.167e+00   8.502 2.81e-05 ***
## I(X2^2)     -1.195e-02  1.578e-03  -7.574 6.46e-05 ***
## I(X1 * X2)  -4.163e-02  1.072e-02  -3.883 0.004654 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4288 on 8 degrees of freedom
## Multiple R-squared:  0.9487,Adjusted R-squared:  0.9167
## F-statistic:  29.6 on 5 and 8 DF,  p-value: 5.864e-05
```

It is seen that given significance level 0.05, the interaction term is not significant. We can also fit a model with the interaction term and perform an F test.

```
LR2 <- lm(Y ~ X1 + I(X1 ^ 2) + X2  + I(X2 ^ 2), data = cakes)
anova(LR2, LR1)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X1 + I(X1^2) + X2 + I(X2^2)
## Model 2: Y ~ X1 + I(X1^2) + X2 + I(X2^2) + I(X1 * X2)
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1      9 4.2430
## 2      8 1.4707  1    2.7722 15.079 0.004654 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It is equivalent to the t test since the squared t statistic is the F statistic. Test of $H_0 : \beta_2 = 0$ versus $H_1 : \beta_2 \neq 0$ can be done in a similar way.
To test $H_0 : \beta_1 = \beta_2 = \beta_5 = 0$ versus H_1 : some are not zero.

```
LR3 <- lm(Y ~ X2  + I(X2 ^ 2), data = cakes)
anova(LR3, LR1)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X2 + I(X2^2)
## Model 2: Y ~ X1 + I(X1^2) + X2 + I(X2^2) + I(X1 * X2)
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     11 11.4739
## 2      8  1.4707  3    10.003 18.137 0.0006293 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The result is against H_0.

5. Consider multiple linear regression. To what extent do we need the normal assumption of e? For example, which of the following rely on the normal assumption: unbiasedness of$\hat{\boldsymbol{\beta}}$, standard errors, confidence interval, F test?
Solution: The OLS estimator $\hat{\boldsymbol{\beta}}$ remains unbiased even if e is not normally distributed. The standard errors do not depend on the normality assumption either since the result

$$\text{Var}\left(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}$$

does not depend on the normality assumption. However, the confidence interval, the F test, and the t test depend on the normality assumption. If the sample size is small, the consequence can be large. If the sample size is large, the consequence may not be so severe due to asymptotic normality.

6. Suppose that $Y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + e$. Statistician A regresses Y directly on X by OLS and obtained the estimator of the regression coefficients. Statistician B standardizes the covariates X_j's first such that the standardized covariates (denoted by X_j^*) have a zero sample mean and variance 1. What is the connection between the estimator by Statistician A and the estimator by Statistician B?
Solution: Standardization means that

$$x_j^* = \frac{x_j - \bar{x}_j}{s\left(x_j\right)},$$

where $s\left(x_j\right)$ is the standard deviation of x_j. Hence, the model becomes

$$\begin{aligned} Y &= \beta_0 + \sum_{j=1}^{p} \beta_j x_j + e \\ &= \beta_0 + \sum_{j=1}^{p} s\left(x_j\right) \beta_j \left(x_j^* + \frac{\bar{x}_j}{s\left(x_j\right)}\right) + e \\ &= \beta_0 + \sum_{j=1}^{p} \beta_j \bar{x}_j + \sum_{j=1}^{p} s\left(x_j\right) \beta_j x_j^* + e \end{aligned}$$

This means that the new intercept becomes $\beta_0 + \sum_{j=1}^{p} \beta_j \bar{x}_j$ and the new slopes become $s\left(x_j\right) \beta_j$. The results obtained by A and B are simply change of scales of the regressors.

7. Consider the data set HWA1 on Studium. We have measured the weight of protein of one type of meat produced by five brands. For each brand, 100 samples are measured.

   (a) Test whether different brands have the same weights.
   Solution: We can apply ANOVA to compare mean of different groups

```
load("HWA1.RData")
```

```r
ANOVA <- aov(Protein ~ Brand, data = HWA)
summary(ANOVA)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Brand         4     57  14.248   1.721  0.144
## Residuals   495   4099   8.281
```

We cannot reject the null hypothesis that they have the same weight under the significance level 0.05.

(b) Test the statement: the weight of Protein of Brand 2 is twice the weight of Protein of Brand 3. Write your own script for this task.
Solution: The linear model that we have fitted is

$$E\left(Y \mid \text{brand}\right) \;=\; \beta_1 + \sum_{i=2}^{5} \beta_i U_i + e_i.$$

Then, the hypothesis corresponds to

$$\beta_1 + \beta_2 \;=\; 2\left(\beta_1 + \beta_3\right),$$

which is equivalent to

$$\begin{bmatrix} 1 & -1 & 2 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} \;=\; 0$$

Our script is

```r
LR <- lm(Protein ~ Brand, data = HWA)
RSS1 <- sum(resid(LR) ^ 2)
L <- matrix(c(1, -1, 2, 0, 0), nrow = 1, ncol = 5)
X <- model.matrix(LR)
Beta_L <- coef(LR) - solve(t(X) %*% X) %*% t(L) %*%
              solve(L %*% solve(t(X) %*% X) %*% t(L)) %*% L %*% coef(LR)
RSS0 <- sum((HWA$Protein - X %*% Beta_L) ^ 2)
n <- 500
p <- 5
p0 <- 4 # Matrix::rankMatrix(rbind(X, L))[1] - Matrix::rankMatrix(L)[1]
Num <- (RSS0 - RSS1) / (p - p0)
Den <- RSS1 / (n - p)
Num / Den
```

```
## [1] 847.1416
```

The F test critical value at the significance level 0.05 is

```r
df(0.95, p - p0, n - p)
```

```
## [1] 0.2542848
```

Hence, we reject the null hypothesis that the weight of Protein of Brand 2 is twice the weight of Protein of Brand 3.
Note that we have used the F test for this task. You can also use the t test for linear combination.

8. Derive the GLS estimator that minimizes

$$\mathrm{RSS}(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}),$$

where $\boldsymbol{\Sigma} > 0$.
Solution: Note that

$$\mathrm{RSS}(\boldsymbol{\beta}) = \boldsymbol{y}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y} + 2\boldsymbol{y}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\boldsymbol{\beta}.$$

Hence,

$$\frac{\partial \mathrm{RSS}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\boldsymbol{\beta},$$

leading to the stationary point

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}.$$