Linear Model Assignment2

110024516 統研碩一邱繼賢

2021 年 10 月 16 日

1. For the data in the problem 2 in Assignment 1. Fit a regression model with the durable press rating (i.e., press) as the response and the four other variables as predictors. Present the output.

```
data = read.table("wrinkle.txt", header = T)
fit = lm(press ~ HCHO + catalyst + temp + time, data = data)
summary(fit)
```

```
##
## Call:
## lm(formula = press ~ HCHO + catalyst + temp + time, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.07876 -0.63939 -0.08531  0.36236  1.65332
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.912212   0.875484  -1.042   0.3074
## HCHO         0.160726   0.066166   2.429   0.0227 *
## catalyst     0.219783   0.034062   6.452 9.33e-07 ***
## temp         0.011226   0.004973   2.257   0.0330 *
## time         0.101974   0.058735   1.736   0.0948 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8365 on 25 degrees of freedom
## Multiple R-squared:  0.6924, Adjusted R-squared:  0.6432
## F-statistic: 14.07 on 4 and 25 DF,  p-value: 3.845e-06
```

a. What percentage of variation in the response is explained by these predictors?

```
summary(fit)$r.squared
```

```
## [1] 0.6923783
```

The percentage of variation in the response is explained by these predictors is about

$$R^2 \approx 69.24\%$$

b. Which observation has the largest (positive) residual? Give the case number.

```
res = summary(fit)$residuals
res[res == max(res)]
```

```
##        9
## 1.653322
```

residual 的最大值為：第九個觀察值的 $residual = 1.653322$

c. Compute the mean and median of the residuals.

```
mean(res)
```

```
## [1] 1.212292e-16
```

```
median(res)
```

```
## [1] -0.08531249
```

The mean of the residuals is very small and closed to zero.
The median of the residuals is about -0.0853.

d. Compute the correlation of the residuals with the fitted values.

```
fitted_value = fit$fitted.values
cor(res, fitted_value)
```

```
## [1] 1.38365e-16
```

The correlation of the residuals with the fitted values is very small and closed to zero.

e. Compute the correlation of the residuals with the formaldehyde concentration (i.e., HCHO).

```
cor(res, data$HCHO)
```

```
## [1] 4.030718e-17
```

The correlation of the residuals with the formaldehyde concentration is very small and closed to zero.

f. Suppose the temperature was increased by 10 while the other predictors were held constant. Predict the change in the press rating.

```
fit$coefficients[4] * 10
```

```
##      temp
## 0.1122556
```

預測 press rating 會上升 10 倍的 estimated coefficient of temperature，大約為 0.1123。

g. Add the variable "HCHC-catalyst" to the model as a predictor. Show the regression output. Add the variable "HCHO/catalyst" to the (original) model as a predictor. Show the output. Why is there no real change in the fit for former model but there is change for the latter model?

```r
fit2 = lm(press ~ HCHO + catalyst + temp + time + (HCHO-catalyst), data = data)
summary(fit2)
```

```
##
## Call:
## lm(formula = press ~ HCHO + catalyst + temp + time + (HCHO -
##     catalyst), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.07876 -0.63939 -0.08531  0.36236  1.65332
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.912212   0.875484  -1.042   0.3074
## HCHO         0.160726   0.066166   2.429   0.0227 *
## catalyst     0.219783   0.034062   6.452 9.33e-07 ***
## temp         0.011226   0.004973   2.257   0.0330 *
## time         0.101974   0.058735   1.736   0.0948 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8365 on 25 degrees of freedom
## Multiple R-squared:  0.6924, Adjusted R-squared:  0.6432
## F-statistic: 14.07 on 4 and 25 DF,  p-value: 3.845e-06
```

```r
fit3 = lm(press ~ HCHO + catalyst + temp + time + (HCHO/catalyst), data = data)
summary(fit3)
```

```
##
## Call:
## lm(formula = press ~ HCHO + catalyst + temp + time + (HCHO/catalyst),
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0975 -0.6315 -0.0528  0.3493  1.6548
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.047988   1.102722  -0.950   0.3514
## HCHO           0.184054   0.130085   1.415   0.1699
## catalyst       0.239330   0.099457   2.406   0.0242 *
## temp           0.011147   0.005085   2.192   0.0383 *
## time           0.103589   0.060384   1.716   0.0991 .
## HCHO:catalyst -0.003202   0.015267  -0.210   0.8356
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.853 on 24 degrees of freedom
## Multiple R-squared:  0.6929, Adjusted R-squared:  0.629
## F-statistic: 10.83 on 5 and 24 DF,  p-value: 1.546e-05
```

因為變數 HCHO-catalyst 和原先的變數有共線性，所以做出來的模型會跟原本的一模一樣；而變數 HCHO/catalyst 和原先的變數之間並沒有共線性，所以做出來的模型會有所不同。

2.

a. Fit a regression model with Fertility as the response and all the other variables as predictors. Compute the estimated covariance matrix of the regression coefficients.

```
df = read.table("swiss.txt", header = T)
fit = lm(Fertility ~ Agriculture + Examination + Education +
             Catholic + Mortality, data = df)
summary(fit)
```

```
##
## Call:
## lm(formula = Fertility ~ Agriculture + Examination + Education +
##     Catholic + Mortality, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.2723  -5.2643   0.5014   4.1177  15.3179
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 66.91040   10.70518   6.250 1.91e-07 ***
## Agriculture -0.17210    0.07030  -2.448  0.01873 *
## Examination -0.25778    0.25387  -1.015  0.31587
## Education   -0.87095    0.18300  -4.759 2.42e-05 ***
## Catholic     0.10414    0.03525   2.954  0.00517 **
## Mortality    1.07699    0.38168   2.822  0.00733 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.165 on 41 degrees of freedom
## Multiple R-squared:  0.7068, Adjusted R-squared:  0.671
## F-statistic: 19.77 on 5 and 41 DF,  p-value: 5.574e-10
```

```
summary(fit)$cov * (summary(fit)$sigma^2) # cov matrix of beta hat
```

```
##             (Intercept)  Agriculture  Examination   Education      Catholic
## (Intercept) 114.6008309 -0.4848505096 -1.2025717683 -0.281121045 -0.0222242006
## Agriculture  -0.4848505  0.0049414193  0.0043716216  0.004787172 -0.0005106843
## Examination  -1.2025718  0.0043716216  0.0644481217 -0.027302590  0.0051328937
## Education    -0.2811210  0.0047871724 -0.0273025899  0.033487979 -0.0029982134
## Catholic     -0.0222242 -0.0005106843  0.0051328937 -0.002998213  0.0012425131
## Mortality    -3.2651742  0.0065633502  0.0003487616  0.012260841 -0.0027453427
##                 Mortality
## (Intercept) -3.2651741723
## Agriculture  0.0065633502
```

```
## Examination  0.0003487616
## Education    0.0122608414
## Catholic    -0.0027453427
## Mortality    0.1456821811
```

b. Use the residuals from the model in part a as the response in a new model with the same predictors. Compare the regression summary for this new model with the previous summary. Identify the similarities and differences and explain mathematically why this occurred.

```
res = fit$residuals
fit2 = lm(res ~ Agriculture + Examination + Education +
                 Catholic + Mortality, data = df)
summary(fit2)
```

```
##
## Call:
## lm(formula = res ~ Agriculture + Examination + Education + Catholic +
##     Mortality, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.2723  -5.2643   0.5014   4.1177  15.3179
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.334e-15  1.071e+01       0        1
## Agriculture -6.773e-18  7.030e-02       0        1
## Examination -3.625e-17  2.539e-01       0        1
## Education    4.377e-17  1.830e-01       0        1
## Catholic    -1.391e-17  3.525e-02       0        1
## Mortality    2.839e-16  3.817e-01       0        1
##
## Residual standard error: 7.165 on 41 degrees of freedom
## Multiple R-squared:  9.401e-32,  Adjusted R-squared:  -0.122
## F-statistic: 7.709e-31 on 5 and 41 DF,  p-value: 1
```

(i) 各變數的係數估計值都呈現非常接近 0 的數值，這是因為 residual 在向量空間中和變數所形成的空間處於直交，所以將 residual 投影到該空間會非常接近一個點，故造成此現象。

(ii) R^2 的數值非常小，因為 R^2 的意義為模型對觀測值的可解釋比例，由於 (i) 所說的原因，此模型對 residual 並不能有很好的解釋。

(iii) 此報表和 a. 小題報表中的 residual standard error 一致，是因為兩個模型的 residual sum of square 和其可自由變動的維度都一樣。

c. Now use the fitted values from the model in part a as the response in a new model with the same predictors. Compare the regression summary for this new model with the first summary. Identify the similarities and differences and explain mathematically why this occurred.

```
fitted_vl = fit$fitted.values
fit3 = lm(fitted_vl ~ Agriculture + Examination + Education +
              Catholic + Mortality, data = df)
summary(fit3)
```

```
## Warning in summary.lm(fit3): essentially perfect fit: summary may be unreliable
```

```
##
## Call:
## lm(formula = fitted_vl ~ Agriculture + Examination + Education +
##     Catholic + Mortality, data = df)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -4.944e-14 -1.896e-15 -3.270e-16  4.276e-15  2.108e-14
##
## Coefficients:
##               Estimate Std. Error    t value Pr(>|t|)
## (Intercept)  6.691e+01  1.433e-14  4.669e+15   <2e-16 ***
## Agriculture -1.721e-01  9.409e-17 -1.829e+15   <2e-16 ***
## Examination -2.578e-01  3.398e-16 -7.586e+14   <2e-16 ***
## Education   -8.709e-01  2.450e-16 -3.556e+15   <2e-16 ***
## Catholic     1.041e-01  4.718e-17  2.207e+15   <2e-16 ***
## Mortality    1.077e+00  5.109e-16  2.108e+15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.59e-15 on 41 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.103e+31 on 5 and 41 DF,  p-value: < 2.2e-16
```

(i) 此報表中的 estimated coefficients 和 a. 小題中所呈現的一模一樣，是因為此題所使用的 response variable 就是全部落在 a. 的回歸線上的 predicted values，所以此題的 estimated coefficients 不會改變。

(ii) 此報表的 residual standard error 非常接近 0，而且 $R^2 = 1$，皆是因為所有的觀測值都落在回歸線上，回歸線可以完美解釋，不會有誤差，所有的變數對模型的貢獻都極為顯著也是同樣的原因。

3. The data set gives information on capital, labor and value added for each of three economic sectors: Food and kindred products (20), electrical and electronic machinery, equipment and supplies (36) and transportation equipment (37). For each sector:

(1) For food and kindred products (20)
a.

```
fit1_20 = lm(log(v_20) ~ log(k_20) + log(l_20), data = data)
summary(fit1_20)$coef
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 25.4928845  6.0876737  4.187623 0.0012593120
## log(k_20)    0.2268538  0.2536026  0.894525 0.3886307189
## log(l_20)   -1.4584782  0.2733979 -5.334636 0.0001780019
```

The estimation of β_1 is about 0.2269, and the estimation of β_2 is about -1.4585.

b.

```
fit2_20 = lm(log(v_20) ~ log(k_20/l_20), offset = log(l_20), data = data)
summary(fit2_20)$coef
```

```
##                 Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)    -3.483744  0.1657144 -21.022574 2.022779e-11
## log(k_20/l_20)  1.289695  0.1964176   6.566088 1.807863e-05
```

The estimation of β_1 is about 1.2897, and the estimation of β_2 is about -0.2897.

c.

```
fit3_20 = lm(log(v_20) ~ log(k_20) + log(l_20) + year, data = data)
summary(fit3_20)$coef
```

```
##                Estimate  Std. Error     t value  Pr(>|t|)
## (Intercept) 19.55432670 16.36468879  1.19490978 0.2572497
## log(k_20)    0.04436007  0.53332801  0.08317597 0.9352059
## log(l_20)   -0.90823598  1.42732510 -0.63632033 0.5375830
## year         0.01095197  0.02784303  0.39334670 0.7015801
```

The estimation of β_1 is about 0.0444, and the estimation of β_2 is about -0.9082.

d.

```
fit4_20 = lm(log(v_20) ~ log(k_20/l_20) + year, offset = log(l_20), data = data)
summary(fit4_20)$coef
```

```
##                   Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)    -9.28894855 1.48528193 -6.253997 4.228991e-05
## log(k_20/l_20) -0.49470246 0.47489645 -1.041706 3.180812e-01
## year            0.05464355 0.01393935  3.920094 2.034775e-03
```

The estimation of β_1 is about -0.4947, and the estimation of β_2 is about 1.4947.

(2) For electrical and electronic machinery, equipment and supplies (36)
a.

```
fit1_36 = lm(log(v_36) ~ log(k_36) + log(l_36), data = data)
summary(fit1_36)$coef
```

```
##               Estimate Std. Error    t value  Pr(>|t|)
## (Intercept) -1.2332115  4.0441871 -0.3049343 0.7656403
## log(k_36)    0.5260689  0.6556094  0.8024121 0.4379179
## log(l_36)    0.2543206  0.3837468  0.6627301 0.5200301
```

The estimation of β_1 is about 0.5261, and the estimation of β_2 is about 0.2543.

b.

```
fit2_36 = lm(log(v_36) ~ log(k_36/l_36), offset = log(l_36), data = data)
summary(fit2_36)$coef
```

```
##                   Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)     -3.8170529  0.3085880 -12.369414 1.451410e-08
## log(k_36/l_36)   0.9000888  0.2918307   3.084284 8.706196e-03
```

The estimation of β_1 is about 0.9001, and the estimation of β_2 is about 0.0999.

c.

```
fit3_36 = lm(log(v_36) ~ log(k_36) + log(l_36) + year, data = data)
summary(fit3_36)$coef
```

```
##                  Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) -15.41454402 2.644064203 -5.829868 0.0001141938
## log(k_36)     0.82098254 0.289192770  2.838876 0.0161142284
## log(l_36)     0.88248951 0.188885139  4.672096 0.0006802239
## year          0.02496758 0.003464598  7.206488 0.0000173807
```

The estimation of β_1 is about 0.8210, and the estimation of β_2 is about 0.8825.

d.

```
fit4_36 = lm(log(v_36) ~ log(k_36/l_36) + year, offset = log(l_36), data = data)
summary(fit4_36)$coef
```

```
##                    Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)     -6.06760912 0.529372148 -11.461897 8.049011e-08
## log(k_36/l_36)   0.03450154 0.263737020   0.130818 8.980868e-01
## year             0.01692118 0.003703154   4.569398 6.441497e-04
```

The estimation of β_1 is about 0.0345, and the estimation of β_2 is about 0.9655.

(3) For transportation equipment (37)

a.

```r
fit1_37 = lm(log(v_37) ~ log(k_37) + log(l_37), data = data)
summary(fit1_37)$coef
```

```
##                Estimate Std. Error    t value    Pr(>|t|)
## (Intercept) -9.6259339  2.8996263 -3.3197153 0.006113404
## log(k_37)    0.5056509  0.5060702  0.9991715 0.337434030
## log(l_37)    0.8454644  0.4215675  2.0055258 0.067992424
```

The estimation of β_1 is about 0.5057, and the estimation of β_2 is about 0.8455.

b.

```r
fit2_37 = lm(log(v_37) ~ log(k_37/l_37), offset = log(l_37), data = data)
summary(fit2_37)$coef
```

```
##                    Estimate Std. Error       t value     Pr(>|t|)
## (Intercept)    -4.712885451  0.0207534 -227.08984793 8.831248e-25
## log(k_37/l_37)  0.009608932  0.4415073    0.02176392 9.829668e-01
```

The estimation of β_1 is about 0.0096, and the estimation of β_2 is about 0.9904.

c.

```r
fit3_37 = lm(log(v_37) ~ log(k_37) + log(l_37) + year, data = data)
summary(fit3_37)$coef
```

```
##                  Estimate  Std. Error    t value    Pr(>|t|)
## (Intercept) -10.027158583 3.075131787 -3.2607248 0.007589203
## log(k_37)     0.158555457 0.817862946  0.1938656 0.849814761
## log(l_37)     1.195294252 0.769390154  1.5535606 0.148570443
## year          0.004579341 0.008312951  0.5508683 0.592735809
```

The estimation of β_1 is about 0.1586, and the estimation of β_2 is about 1.1953.

d.

```r
fit4_37 = lm(log(v_37) ~ log(k_37/l_37) + year, offset = log(l_37), data = data)
summary(fit4_37)$coef
```

```
##                    Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)    -5.050476290 0.705314602 -7.1606008 1.147324e-05
## log(k_37/l_37) -0.316815696 0.819688574 -0.3865074 7.058870e-01
## year            0.004259221 0.008894519  0.4788591 6.406461e-01
```

The estimation of β_1 is about -0.3168 and the estimation of β_2 is about 1.3168.