Homework 4

Juwon Lee, Economics and Statistics, UCLA

2023-02-14

tinytex::install_tinytex()

1.

A paper company is interested in making its operations more efficient. They collect data on the total manufacturing **cost** per month (in dollars), the total production of **paper** per month (in tons), the total number of **machine** hours per month, the total variable **overhead** cost per month (in thousands of dollars) and the total number of **labor** hours each month.

(a) Explore the relationship among the variables:

- Generate a correlation coefficient matrix. Try cor(data) in R.

- Generate a scatter plot matrix. Try pairs(data) in R.

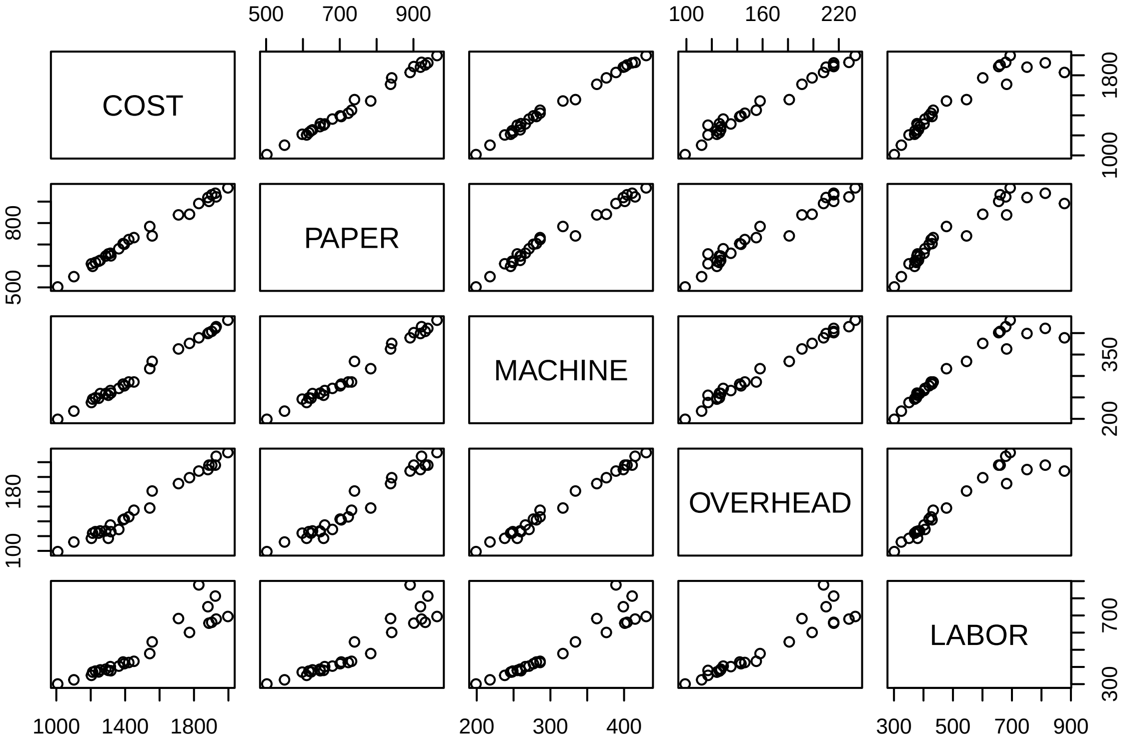
- What do you feel about the relationship among the variables? Comment.

```
paper <- read.table("/Users/user/Desktop/Yonsei/Junior/3-2/Introduction to Data Analysis and Regression/

cor(paper)
```

```
##                COST     PAPER   MACHINE  OVERHEAD     LABOR
## COST     1.0000000 0.9959338 0.9973885 0.9893730 0.9384741
## PAPER    0.9959338 1.0000000 0.9893982 0.9780120 0.9329742
## MACHINE  0.9973885 0.9893982 1.0000000 0.9943632 0.9447326
## OVERHEAD 0.9893730 0.9780120 0.9943632 1.0000000 0.9380474
## LABOR    0.9384741 0.9329742 0.9447326 0.9380474 1.0000000
```

```
pairs(paper)
```



It means that, LABOR does something different role, whereas four variables are correlated.

(b) Fit a multiple regression using cost as the response variable, and the other four variables as explanatory variables. Write down the regression equation.

Let $Y = \text{COST},\ X_1 = \text{PAPER},\ X_2 = \text{MACHINE},\ X_3 = \text{OVERHEAD},\ X_4 = \text{LABOR}$

```
lm_data_hw4_1 <- lm(COST~PAPER+MACHINE+OVERHEAD+LABOR, data=paper)

summary(lm_data_hw4_1)
```

```
##
## Call:
## lm(formula = COST ~ PAPER + MACHINE + OVERHEAD + LABOR, data = paper)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.691  -7.407  -1.978   6.675  22.516
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 51.72314   21.70397   2.383   0.0262 *
## PAPER        0.94794    0.12002   7.898 7.30e-08 ***
## MACHINE      2.47104    0.46556   5.308 2.51e-05 ***
## OVERHEAD     0.04834    0.52501   0.092   0.9275
## LABOR       -0.05058    0.04030  -1.255   0.2226
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.08 on 22 degrees of freedom
## Multiple R-squared:  0.9988, Adjusted R-squared:  0.9986
## F-statistic:  4629 on 4 and 22 DF,  p-value: < 2.2e-16
```

Thus, $Y = 51.72314 + 0.94794\hat{\beta}_1 + 2.47104\hat{\beta}_2 + 0.04834\hat{\beta}_3 - 0.05058\hat{\beta}_4$.

(c) Conduct the F-test for the overall fit of the regression. What conclusions can you draw?

```
anova(lm_data_hw4_1)
```

```
## Analysis of Variance Table
##
## Response: COST
##           Df  Sum Sq Mean Sq    F value    Pr(>F)
## PAPER      1 2255666 2255666 18388.2129 < 2.2e-16 ***
## MACHINE    1   15561   15561   126.8547  1.33e-10 ***
## OVERHEAD   1       3       3     0.0269    0.8711
## LABOR      1     193     193     1.5755    0.2226
## Residuals 22    2699     123
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Thus, for X_1(PAPER) and X_2(MACHINE), the null hypotheses $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ are rejected.

(d) What proportion of the variation in cost has been explained by the regression?

In (b), $R^2 = 0.9988$. Thus, 99.88% can be explained by the regression.

(e) Test each of the individual regression coefficients. Interpret the result for each test. Which of the variables are significant?

In (b), for each X_i for $i = 1,\ 2,\ 3,\ 4$,

$H_0 : \beta_1|x_2, x_3, x_4 = 0$ vs. $H_1 : \beta_1|x_2, x_3, x_4 \neq 0$

$H_0 : \beta_2|x_1, x_3, x_4 = 0$ vs. $H_1 : \beta_2|x_1, x_3, x_4 \neq 0$

$H_0 : \beta_3|x_1, x_2, x_4 = 0$ vs. $H_1 : \beta_3|x_1, x_2, x_4 \neq 0$

$H_0 : \beta_4|x_1, x_2, x_3 = 0$ vs. $H_1 : \beta_4|x_1, x_2, x_3 \neq 0$

Then, we can't say that $\beta_1|x_2, x_3, x_4 = 0$ and $\beta_2|x_1, x_3, x_4 = 0$ for $\alpha = 0.001$.

(f) Perform a partial F-test to determine whether the variables associated with 'overhead' and 'laber' hours can be removed from the model. Comment on the results of the test.

First of all, $H_0 : Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + e$ vs.
$H_1 : Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + e$.

```r
lm_data_hw4_4 <- lm(COST~PAPER+MACHINE+LABOR, data=paper)

anova(lm_data_hw4_4, lm_data_hw4_1)
```

```
## Analysis of Variance Table
##
## Model 1: COST ~ PAPER + MACHINE + LABOR
## Model 2: COST ~ PAPER + MACHINE + OVERHEAD + LABOR
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     23 2699.8
## 2     22 2698.7  1    1.0399 0.0085 0.9275
```

Thus, p-value $= 0.9275 > 0.05$, so that we can't reject the null.
It means that we cannot say that the full model is better.

If we remove the 'Labor', $H_0 : Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$ vs.
$H_1 : Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + e$.

```r
lm_data_hw4_5 <- lm(COST~PAPER+MACHINE+OVERHEAD, data=paper)

anova(lm_data_hw4_5, lm_data_hw4_1)
```

```
## Analysis of Variance Table
##
## Model 1: COST ~ PAPER + MACHINE + OVERHEAD
## Model 2: COST ~ PAPER + MACHINE + OVERHEAD + LABOR
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     23 2892.0
## 2     22 2698.7  1    193.26 1.5755 0.2226
```

Then p-value $= 0.2226 > 0.05$, so that we can't reject the null.
It means that we cannot say that the full model is better.

2. Exercise 5.4.3

Quality of vintage(Y) = on a scale from 1 (worst) to 5 (best) with some half points.
End of harvest(X_1) = measured as the number days since August 31.
Rain(X_2) = a dummy variable for unwanted rain at harvest = 1 if yes.

The model is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 * X_2 + e$.

(a) Show that the coefficient of the interaction term in moedl is statistically significant. In other words, show that the rate of change in quality rating depends on whether there has been any unwanted rain at vintage.

```
wine <- read.table("/Users/user/Desktop/Yonsei/Junior/3-2/Introduction to Data Analysis and Regression/

lm_data_hw4_2 <- lm(Quality~EndofHarvest+Rain+EndofHarvest*Rain, data=wine)
summary(lm_data_hw4_2)
```

```
##
## Call:
## lm(formula = Quality ~ EndofHarvest + Rain + EndofHarvest * Rain,
##     data = wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6833 -0.5703  0.1265  0.4385  1.6354
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        5.16122    0.68917   7.489 3.95e-09 ***
## EndofHarvest      -0.03145    0.01760  -1.787   0.0816 .
## Rain               1.78670    1.31740   1.356   0.1826
## EndofHarvest:Rain -0.08314    0.03160  -2.631   0.0120 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7578 on 40 degrees of freedom
## Multiple R-squared:  0.6848, Adjusted R-squared:  0.6612
## F-statistic: 28.97 on 3 and 40 DF,  p-value: 4.017e-10
```

Thus, the coefficient of the interaction term(β_3) has p-value $0.012 < 0.025$, so that it is statistically significant for $\alpha = 0.05$.

(b) Estimate the number of days of delay to the end of harvest it takes to decrease the quality rating by 1 point when there is:

(i) No unwanted rain at harvest($X_4 = 0$)
(ii) Some unwanted rain at harvest($X_4 \neq 0$)

```
lm_data_hw4_3 <- lm(Quality~EndofHarvest+Rain, data=wine)
summary(lm_data_hw4_3)
```

```
##
## Call:
## lm(formula = Quality ~ EndofHarvest + Rain, data = wine)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max 
## -1.4563 -0.7366  0.1430  0.6413  1.7652 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   6.14633    0.61896   9.930  1.8e-12 ***
## EndofHarvest -0.05723    0.01564  -3.660 0.000713 ***
## Rain         -1.62219    0.25478  -6.367  1.3e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.8107 on 41 degrees of freedom
## Multiple R-squared:  0.6303, Adjusted R-squared:  0.6123 
## F-statistic: 34.95 on 2 and 41 DF,  p-value: 1.383e-09
```

(i) Thus, $Y = 6.14633 - 0.05723X_1 - 1.62219X_2$.
Then, claim $Y - 1 = 6.14633 - 0.05723(X_1 + k) - 1.62219X_2$.
$\rightarrow\ 1 = 0.05723k,\ k \approx 17.473$.

(ii) $Y = 5.16122 - 0.03145X_1 + 1.7867X_2 - 0.08314X_3$.
Then, claim $Y - 1 = 5.16122 - 0.03145(X_1 + k) + 1.7867X_2 - 0.08314X_3$.
$\rightarrow\ 1 = 0.03145k,\ k \approx 31.797$.