Stats 101A Homework #4

Damien Ha

2023-02-24

Contents

Problem 1

Loading Data

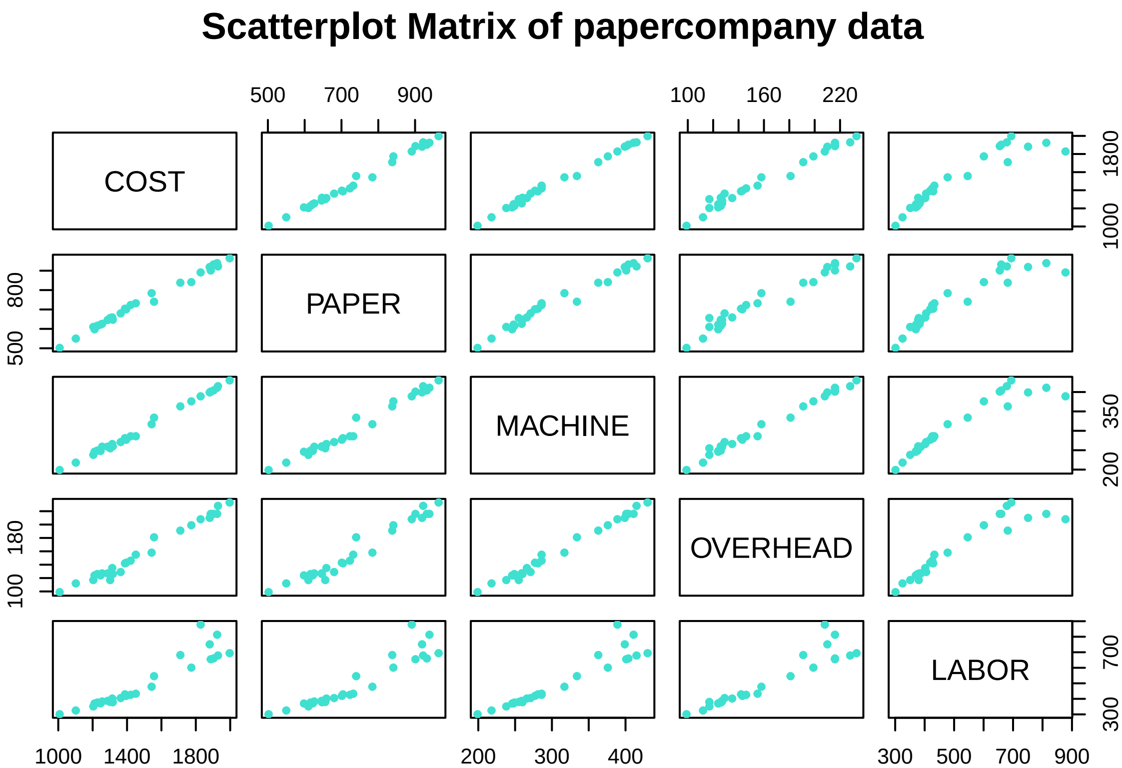
```
papercompany <- read.table("papercompany.txt", header = T)
```

Part A

```
cor(papercompany)
```

```
##               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(papercompany, main = "Scatterplot Matrix of papercompany data", col = "turquoise", pch = 20)
```



All pairings of variables seem to have a strong, positive linear relationship with correlation coefficients close to positive 1. Certain pairs appear to have some odd trends or possible outliers, such as the points branching off and forming a curve in labor vs overhead. Overall each variable seems to have a positive trend.

Part B

```
model <- lm(papercompany$COST ~ papercompany$PAPER + papercompany$MACHINE +
             papercompany$OVERHEAD + papercompany$LABOR)
summary(model)
```

```
##
## Call:
## lm(formula = papercompany$COST ~ papercompany$PAPER + papercompany$MACHINE +
##     papercompany$OVERHEAD + papercompany$LABOR)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.691  -7.407  -1.978   6.675  22.516
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           51.72314   21.70397   2.383   0.0262 *
## papercompany$PAPER     0.94794    0.12002   7.898 7.30e-08 ***
## papercompany$MACHINE   2.47104    0.46556   5.308 2.51e-05 ***
## papercompany$OVERHEAD  0.04834    0.52501   0.092   0.9275
## papercompany$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
```

The regression equation is $Y = 51.72314 + 0.94794X_1 + 2.47104X_2 + 0.04834X_3 - 0.05058X_4$ where Y is cost and X_1, X_2, X_3, and X_4 are paper, machine, overhead, and labor respectively

Part C

```r
anova(model)
```

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

Paper has F = 18388.2129 with p = $2.2 \cdot 10^{-16}$ and machine has F = 126.8547 with p = $1.33 \cdot 10^{-10}$ suggesting that at $\alpha = 0.05$, these variables help better fit the model to the data. However, overhead has a p-value if 0.8711 and labor has a p-value of 0.2226, which, if we consider at the level $\alpha = 0.05$, may suggest these variables aren't significant

Part D

From out linear model summary, we can see the R^2 value is 0.998, so 99.8% of the variation in cost is explained by the regression model

Part E

```r
# R-squared for
# cost vs paper
summary(lm(papercompany$COST ~ papercompany$PAPER))[[9]]
```

```
## [1] 0.9915595
```

```r
summary(lm(papercompany$COST ~ papercompany$PAPER + papercompany$MACHINE +
             papercompany$OVERHEAD + papercompany$LABOR))[[4]][[2]]
```

```
## [1] 0.9479421
```

```r
# cost vs machine
summary(lm(papercompany$COST ~ papercompany$MACHINE))[[9]]
```

```
## [1] 0.9945752
```

```r
summary(lm(papercompany$COST ~ papercompany$PAPER + papercompany$MACHINE +
             papercompany$OVERHEAD + papercompany$LABOR))[[4]][[3]]
```

```
## [1] 2.47104
```

```r
# cost vs overhead
summary(lm(papercompany$COST ~ papercompany$OVERHEAD))[[9]]
```

```
## [1] 0.9780133
```

```r
summary(lm(papercompany$COST ~ papercompany$PAPER + papercompany$MACHINE +
             papercompany$OVERHEAD + papercompany$LABOR))[[4]][[4]]
```

```
## [1] 0.04833872
```

```r
# cost vs labor
summary(lm(papercompany$COST ~ papercompany$LABOR))[[9]]
```

```
## [1] 0.875963
```

```r
summary(lm(papercompany$COST ~ papercompany$PAPER + papercompany$MACHINE +
             papercompany$OVERHEAD + papercompany$LABOR))[[4]][[5]]
```

```
## [1] -0.05057756
```

Part F

```r
reduced <- lm(papercompany$COST ~ papercompany$PAPER + papercompany$MACHINE)
anova(reduced, model)
```

```
## Analysis of Variance Table
##
## Model 1: papercompany$COST ~ papercompany$PAPER + papercompany$MACHINE
## Model 2: papercompany$COST ~ papercompany$PAPER + papercompany$MACHINE +
##     papercompany$OVERHEAD + papercompany$LABOR
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     24 2895.3
## 2     22 2698.7  2    196.57 0.8012 0.4615
```

F = 0.8 and p = 0.4615 so at $\alpha = 0.05$ we fail to reject the null hypothesis and say overhead and labor don't contribute significant information to cost when considering paper and machine variables

Question 2

Loading Data

```r
latour <- read.table("Latour.txt", header = T)
```

Part A

```r
summary(lm(Quality ~ EndofHarvest + Rain + Rain:EndofHarvest, data = latour))
```

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

The interaction term has a p-value of 0.012 which, at $\alpha = 0.05$ is significant, suggesting the rate of change in quality rating depends on whether there has been any unwanted rain at vintage.

Part B

Part I

From the equation, letting Y = quality, X_1 = end of harvest days, and X_2 = rain, we have $Y = 5.1622 - 0.03145X_1 + 1.7867X_2 - 0.08314X_1X_2$

Setting $X_2 = 0$, we have $Y = 5.16122 - 0.03145X_1$

The derivative of the equation with respect to X_1 is -0.03145, and we want to find a quality drop of 1:

```r
paste0("The number of days for a decrease in quality by 1 without rain is ", -1/-0.03145)
```

```
## [1] "The number of days for a decrease in quality by 1 without rain is 31.7965023847377"
```

Part II

Set $X_2 = 1$

$Y = 6.94792 - 0.11459X_1$

The derivative with respect to X_1 is -0.11459, and we want to find a quality drop of 1

```r
paste0("The number of days for a decrease in quality by 1 with rain is ", -1/-0.11459)
```

```
## [1] "The number of days for a decrease in quality by 1 with rain is 8.72676498821887"
```