401-hw3

2024-10-14

```r
library(readr)
library(car)
```

```
## Warning: package 'car' was built under R version 4.3.3
```

```
## Loading required package: carData
```

```r
library(carData)
```

Problem 2

a. Here, y depends on both x and w, and x depends on w. This configuration represents a **fork** because w affects both x and y.

```r
# Set seed for reproducibility
set.seed(123)

# Generate the data
n <- 500
w <- runif(n, min = 0, max = 5)
delta <- rnorm(n, mean = 0, sd = 1)
x <- w + delta
epsilon <- rnorm(n, mean = 0, sd = 1)
y <- 4 + 2 * x - 3 * w + epsilon

# Correlation matrix
data <- data.frame(x = x, w = w, y = y)
cor_matrix <- cor(data)

# Basic descriptive statistics
summary_stats <- summary(data)

list(correlation_matrix = cor_matrix, summary_statistics = summary_stats)
```

```
## $correlation_matrix
##                 x          w             y
## x  1.0000000000  0.8189365 -0.0006478424
## w  0.8189364870  1.0000000 -0.5304605791
## y -0.0006478424 -0.5304606  1.0000000000
##
## $summary_statistics
##        x                w                  y
##  Min.   :-1.546   Min.   :0.002327   Min.   :-7.3449
##  1st Qu.: 1.138   1st Qu.:1.229984   1st Qu.:-0.1165
##  Median : 2.472   Median :2.382781   Median : 1.6234
##  Mean   : 2.498   Mean   :2.476418   Mean   : 1.5896
##  3rd Qu.: 3.908   3rd Qu.:3.664487   3rd Qu.: 3.3221
##  Max.   : 6.943   Max.   :4.997023   Max.   :10.1170
```

```
# Linear regression of y on x
model1 <- lm(y ~ x, data = data)

# Summary of the model to check coefficients
summary(model1)
```

```
##
## Call:
## lm(formula = y ~ x, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9352 -1.7057  0.0338  1.7308  8.5278
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.5920148  0.2064629   7.711 6.85e-14 ***
## x           -0.0009786  0.0676899  -0.014    0.988
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.649 on 498 degrees of freedom
## Multiple R-squared:  4.197e-07,  Adjusted R-squared:  -0.002008
## F-statistic: 0.000209 on 1 and 498 DF,  p-value: 0.9885
```

```
# Extracting 95% confidence interval for the coefficient of x
confint(model1, "x", level = 0.95)
```

```
##        2.5 %    97.5 %
## x -0.1339717 0.1320145
```

c. The large p-value of 0.988 means we fail to reject the null hypothesis that the coefficient of x is 0, which suggests that the coefficient of x is not statistically significant at the 0.05 significance level. The 95% CI (-0.1339717, 0.1320145) does not cover the true slope of 2 for x.

```
# Linear regression of y on x and w
model2 <- lm(y ~ x + w, data = data)

# Summary of the model to check coefficients
summary(model2)
```

```
##
## Call:
## lm(formula = y ~ x + w, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5871 -0.7032 -0.0118  0.6028  3.1817
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.03394    0.09146   44.11   <2e-16 ***
## x            1.98951    0.04532   43.90   <2e-16 ***
## w           -2.99402    0.05582  -53.63   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.018 on 497 degrees of freedom
## Multiple R-squared:  0.8527, Adjusted R-squared:  0.8521
## F-statistic:  1438 on 2 and 497 DF,  p-value: < 2.2e-16
```

```
# 95% confidence interval for the coefficient of x in the second model
confint(model2, "x", level = 0.95)
```

```
##      2.5 %   97.5 %
## x 1.900471 2.078544
```

d. The small p-value that is < 2.2e-16 means we can reject the null hypothesis that the coefficient of x is 0, which suggests that the coefficient of x is significant at the 0.05 significance level. The 95% CI (1.900471, 2.078544) cover the true slope of 2 for x.

```
# Calculate VIF
vif_values <- vif(model2)
vif_values
```

```
##        x        w
## 3.036348 3.036348
```

e. The VIF for x is 3.036348 and the VIF for w is also 3.036348.

Problem 3

a. In this case, w depends on y, which depends on x. This is a **collider** structure since y is an effect of both x and w.

```r
# Set seed for reproducibility
set.seed(123)

# Generate the data
n <- 500
x <- runif(n, min = 0, max = 5)
delta <- rnorm(n, mean = 0, sd = 1)
y <- x + delta
epsilon <- rnorm(n, mean = 0, sd = 1)
w <- 4 + 2 * x + 3 * y + epsilon

# Correlation matrix
data <- data.frame(x = x, y = y, w = w)
cor_matrix <- cor(data)

# Basic descriptive statistics
summary_stats <- summary(data)

list(correlation_matrix = cor_matrix, summary_statistics = summary_stats)
```

```
## $correlation_matrix
##           x         y         w
## x 1.0000000 0.8189365 0.9138879
## y 0.8189365 1.0000000 0.9691315
## w 0.9138879 0.9691315 1.0000000
##
## $summary_statistics
##        x                  y                w
##  Min.   :0.002327   Min.   :-1.546   Min.   :-0.4431
##  1st Qu.:1.229984   1st Qu.: 1.138   1st Qu.: 9.8428
##  Median :2.382781   Median : 2.472   Median :16.2572
##  Mean   :2.476418   Mean   : 2.498   Mean   :16.4698
##  3rd Qu.:3.664487   3rd Qu.: 3.908   3rd Qu.:22.7563
##  Max.   :4.997023   Max.   : 6.943   Max.   :34.1036
```

```r
# Linear regression of y on x
model1 <- lm(y ~ x, data = data)

# Summary of the model to check coefficients
summary(model1)
```

```
##
## Call:
## lm(formula = y ~ x, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.82796 -0.61831  0.03553  0.69367  2.68062
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.552e-05  9.044e-02    0.00        1
## x            1.009e+00  3.168e-02   31.84   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.006 on 498 degrees of freedom
## Multiple R-squared:  0.6707, Adjusted R-squared:   0.67
## F-statistic:  1014 on 1 and 498 DF,  p-value: < 2.2e-16
```

```
# Extracting 95% confidence interval for the coefficient of x
confint(model1, "x", level = 0.95)
```

```
##       2.5 %   97.5 %
## x 0.9465387 1.071016
```

c. The small p-value that is < 2.2e-16 means we can reject the null hypothesis that the coefficient of x is 0, which suggests that the coefficient of x is significant at the 0.05 significance level. The 95% CI (0.9465387, 1.071016) cover the true slope of 1 for x.

```
# Linear regression of y on x and w
model2 <- lm(y ~ x + w, data = data)

# Summary of the model to check coefficients
summary(model2)
```

```
##
## Call:
## lm(formula = y ~ x + w, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.92548 -0.20067 -0.00282  0.22810  0.88489
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.211061   0.034308  -35.30   <2e-16 ***
## x           -0.498835   0.025007  -19.95   <2e-16 ***
## w            0.300218   0.004551   65.97   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3225 on 497 degrees of freedom
## Multiple R-squared:  0.9662, Adjusted R-squared:  0.9661
## F-statistic:  7113 on 2 and 497 DF,  p-value: < 2.2e-16
```

```
# 95% confidence interval for the coefficient of x in the second model
confint(model2, "x", level = 0.95)
```

```
##       2.5 %     97.5 %
## x -0.547967 -0.4497033
```

d. The small p-value that is < 2.2e-16 means we can reject the null hypothesis that the coefficient of x is 0, which suggests that the coefficient of x is significant at the 0.05 significance level. The 95% CI (-0.547967, -0.4497033) does not cover the true slope of 1 for x.

```
# Calculate VIF
vif_values <- vif(model2)
vif_values
```

```
##        x        w
## 6.067631 6.067631
```

e. The VIF for x is 6.067631 and the VIF for w is also 6.067631.

```
# R-squared and residual standard error for both models
r_squared_model1 <- summary(model1)$r.squared
se_model1 <- summary(model1)$sigma

r_squared_model2 <- summary(model2)$r.squared
se_model2 <- summary(model2)$sigma

list(
  model1 = list(R_squared = r_squared_model1, Residual_SE = se_model1),
  model2 = list(R_squared = r_squared_model2, Residual_SE = se_model2)
)
```

```
## $model1
## $model1$R_squared
## [1] 0.670657
##
## $model1$Residual_SE
## [1] 1.006324
##
##
## $model2
## $model2$R_squared
## [1] 0.9662433
##
## $model2$Residual_SE
## [1] 0.3225004
```

f. Model 2 is the better model as it explains more of the variability in y and has lower prediction error on average. Since the VIF values from Model 2 are around 3.04 for both predictors, multicollinearity is present but moderate. This suggests some dependence between predictors, which could affect coefficient stability and interpretability, although it isn't excessively high.

Problem 4

a. In this case, w is influenced by x, and y depends on w. This configuration resembles a **pipe**, where the relationship flows from x to w and then to y.

```r
# Set seed for reproducibility
set.seed(123)

# Generate the data
n <- 500
x <- runif(n, min = 0, max = 5)
delta <- rnorm(n, mean = 0, sd = 1)
w <- x + delta
epsilon <- rnorm(n, mean = 0, sd = 1)
y <- 2 * w + epsilon

# Correlation matrix
data <- data.frame(x = x, w = w, y = y)
cor_matrix <- cor(data)

# Basic descriptive statistics
summary_stats <- summary(data)

list(correlation_matrix = cor_matrix, summary_statistics = summary_stats)
```

```
## $correlation_matrix
##           x         w         y
## x 1.0000000 0.8189365 0.7871243
## w 0.8189365 1.0000000 0.9602132
## y 0.7871243 0.9602132 1.0000000
##
## $summary_statistics
##        x                  w                y
##  Min.   :0.002327   Min.   :-1.546   Min.   :-4.744
##  1st Qu.:1.229984   1st Qu.: 1.138   1st Qu.: 2.116
##  Median :2.382781   Median : 2.472   Median : 4.848
##  Mean   :2.476418   Mean   : 2.498   Mean   : 5.019
##  3rd Qu.:3.664487   3rd Qu.: 3.908   3rd Qu.: 7.667
##  Max.   :4.997023   Max.   : 6.943   Max.   :14.767
```

```r
# Linear regression of y on x
model1 <- lm(y ~ x, data = data)

# Summary of the model to check coefficients
summary(model1)
```

```
##
## Call:
## lm(formula = y ~ x, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.0346 -1.2424 -0.0477  1.3816  6.7231
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.03391    0.20180   0.168    0.867
## x            2.01295    0.07068  28.479   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.245 on 498 degrees of freedom
## Multiple R-squared:  0.6196, Adjusted R-squared:  0.6188
## F-statistic:   811 on 1 and 498 DF,  p-value: < 2.2e-16
```

```
# Extracting 95% confidence interval for the coefficient of x
confint(model1, "x", level = 0.95)
```

```
##      2.5 %   97.5 %
## x 1.874081 2.151829
```

c. The small p-value that is < 2.2e-16 means we can reject the null hypothesis that the coefficient of x is 0, which suggests that the coefficient of x is significant at the 0.05 significance level.

```
# Linear regression of y on x and w
model2 <- lm(y ~ x + w, data = data)

# Summary of the model to check coefficients
summary(model2)
```

```
##
## Call:
## lm(formula = y ~ x + w, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5871 -0.7032 -0.0118  0.6028  3.1817
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.033938   0.091461   0.371    0.711
## x           0.005985   0.055822   0.107    0.915
## w           1.989508   0.045317  43.902   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.018 on 497 degrees of freedom
## Multiple R-squared:  0.922,  Adjusted R-squared:  0.9217
## F-statistic:  2938 on 2 and 497 DF,  p-value: < 2.2e-16
```

d. The large p-value of 0.915 means we cannot reject the null hypothesis that the coefficient of x is 0, which suggests that the coefficient of x is not significant at the 0.05 significance level.

e. The R squared of the first model is 0.6196, and the R squared of the second model is 0.922. Therefore, Model 2 is the better model as it explains more of the variability in y on average.