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
```

```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
## -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
```

```r
# 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.

```r
# 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
```

```r
# 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.

```r
# 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
# 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.

Problem 5

a.

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, where the true regression coefficients are $\beta_0 = 2$, $\beta_1 = 2$, $\beta_2 = 0.3$, and the standard deviation of errors is 1 since $\epsilon \sim N(0, 1)$.

```
set.seed(1)
x1 <- runif(100) # part a
x2 <- 0.5*x1 + rnorm(100)/10
y <- 2 + 2*x1 + .3*x2 + rnorm(100)

cor(x1,x2)
```

```
## [1] 0.8351212
```

b. The correlation between x1 and x2 is 0.8351212.

```
model <- lm(y ~ x1 + x2)
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8311 -0.7273 -0.0537  0.6338  2.3359
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1305     0.2319   9.188 7.61e-15 ***
## x1            1.4396     0.7212   1.996   0.0487 *
## x2            1.0097     1.1337   0.891   0.3754
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.056 on 97 degrees of freedom
## Multiple R-squared:  0.2088, Adjusted R-squared:  0.1925
## F-statistic:  12.8 on 2 and 97 DF,  p-value: 1.164e-05
```

```
confint(model)
```

```
##                    2.5 %   97.5 %
## (Intercept)  1.670278673 2.590721
## x1           0.008213776 2.870897
## x2          -1.240451256 3.259800
```

c.

The estimated intercept 2.1305 is very close to the true intercept of 2. The estimated coefficient for x_1 is 1.4396, which is lower than the true value of 2. The estimated coefficient for x_2 is 1.0097, which is higher than the true value of 0.3. The regression provides parameter estimates that are in the same general direction as the true parameters, but with discrepancies due to statistical noise and potential multicollinearity.

The coefficient of x_1 is 1.4396, with a t-statistics of 1.996 and p-value is equal to 0.0487. If we do a hypothesis test with $H_0 : B_1 = 0$ and $H_1 : B_1 \neq 0$, since 0.0487 < 0.05, we reject the null hypothesis, meaning that the coefficient of x is significant at the 0.05 level. The coefficient of x_2 is 1.0097, with a t-statistics of 0.891 and p-value is equal to 0.3754. If we do a hypothesis test with $H_0 : B_2 = 0$ and $H_1 : B_2 \neq 0$, since 0.3754 > 0.05, we fail reject the null hypothesis, meaning that the coefficient of x is not significant at the 0.05 level. Therefore, x1 is significantly different from 0.

Based on the data, the true coefficients are both covered for x_1 and x_2 in the CI at 95 confidence level.

```
modeld <- lm(y ~ x1)
summary(modeld)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89495 -0.66874 -0.07785  0.59221  2.45560
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1124     0.2307   9.155 8.27e-15 ***
## x1            1.9759     0.3963   4.986 2.66e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.055 on 98 degrees of freedom
## Multiple R-squared:  0.2024, Adjusted R-squared:  0.1942
## F-statistic: 24.86 on 1 and 98 DF,  p-value: 2.661e-06
```

```
confint(modeld)
```

```
##                2.5 %   97.5 %
## (Intercept) 1.654488 2.570299
## x1          1.189529 2.762329
```

d.

The coefficient of x_1 in this model is 1.9759, with a t-statistics of 4.986 and p-value is equal to 2.66e-06. If we do a hypothesis test with $H_0 : B_1 = 0$ and $H_1 : B_1 \neq 0$, since 2.66e-06 < 0.05, we reject the null hypothesis, meaning that the coefficient of x is significant at the 0.05 level. The estimate of coefficient is very close to 2, and with relatively large standard error.

Therefore, β_1 is significantly different from 0. The true β_1 is included in the 95% CI.

```
modeld <- lm(y ~ x2)
summary(modeld)
```

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.62687 -0.75156 -0.03598  0.72383  2.44890
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.3899     0.1949   12.26  < 2e-16 ***
## x2            2.8996     0.6330    4.58 1.37e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.072 on 98 degrees of freedom
## Multiple R-squared:  0.1763, Adjusted R-squared:  0.1679
## F-statistic: 20.98 on 1 and 98 DF,  p-value: 1.366e-05
```

```
confint(modeld)
```

```
##                2.5 %   97.5 %
## (Intercept) 2.003116 2.776783
## x2          1.643324 4.155846
```

e.

The coefficient of x_2 in this model is 2.8996, with a t-statistics of 4.58 and p-value is equal to 1.37e-05. If we do a hypothesis test with $H_0 : B_1 = 0$ and $H_1 : B_1 \neq 0$, since 1.37e-05 < 0.05, we reject the null hypothesis, meaning that the coefficient of x is significant at the 0.05 level. The estimate of coefficient is very close to 3, whereas our true coefficient is 0.3 for x_2.

Therefore, β_2 is significantly different from 0. From data, we saw that both the true β_2 and intercept parameter are not included in the 95% CI.