Advanced Stats HW 2 (Corrected)

Joel Vasama — 6012872

2022-06-13

Notes on corrections

Errors in R lm() for 1. b) & f), and 2. were fixed by specifying group variables to R as factors (errors due to R interpreting them as continuous).

DFs in 1. e), and 2. were corrected and new F- and p-values were computed.

The design matrix in 2. b) should be accurate after correcting mistake in recode().

Plots added at end.

Setup

```
f_test <- function(n, p, r, SSE_h0, SSE){
  np_diff <- n - p
  SSE_diff <- SSE_h0 - SSE
  f_val <- (np_diff/r) * (SSE_diff/SSE)
  return(f_val)
}

rVsHand <- function(betaHand, betaR, modelHand, modelR){
  cat("Beta's by hand:", betaHand, "\n")
  cat("Beta's by R:", betaR, "\n")
  cat("\nF by hand:", modelHand, "\n")
  cat("F by R:", modelR, "\n")
}

setwd("/Users/joelvasama/Desktop/NB MSc/Advanced Stats/Exercises — Advanced Stats/
Previous/16 HW2")
table1 <- read.csv("Table1.csv")

n <- nrow(table1)
y <- table1$y
intercept <- rep(1, n)
x <- table1$cov
dummies <- recode(table1$group, "1" = 0, "2" = 1)
```

Question 1 — Regression

a. Dummy coding — Placebo vs. Drug

Design matrix

```
X_d <- matrix(c(intercept, dummies), ncol = 2)
```

```
##
##  Column rank = 2
```

Beta, y_hat, error, and residual sum of squares (SSE)

```
Beta_d <- solve(t(X_d) %*% X_d) %*% t(X_d) %*% y
y_hat_d <- X_d %*% Beta_d
error_d <- y - y_hat_d

SSE_drug <- t(error_d) %*% error_d
```

Null hypothesis

```
X_d_H0 <- matrix(c(intercept), ncol = 1)
```

```
## Column rank = 1
```

```
Beta_d_H0 <- solve(t(X_d_H0) %*% X_d_H0) %*% t(X_d_H0) %*% y
y_hat_d_H0 <- X_d_H0 %*% Beta_d_H0
error_d_H0 <- y - y_hat_d_H0

SSE_drug_H0 <- t(error_d_H0) %*% error_d_H0
```

F-test

```
f_val_drug <- f_test(n = nrow(table1), p = 2, r = 1, SSE_drug_H0, SSE_drug)
```

```
## With F-statistic of 7.427115 and df 1, 58, p < 0.01
```

b. R Analysis

```
drug_reg <- lm(y ~ as.factor(group), data = table1)
summary(drug_reg)
```

```
##
## Call:
## lm(formula = y ~ as.factor(group), data = table1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.1406  -8.1072   0.1354   6.8367  21.5754
##
## Coefficients:
##                   Estimate Std. Error t value             Pr(>|t|)
## (Intercept)         49.506      1.864  26.554 < 0.0000000000000002 ***
## as.factor(group)2   -7.185      2.637  -2.725              0.00848 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.21 on 58 degrees of freedom
## Multiple R-squared:  0.1135, Adjusted R-squared:  0.09823
## F-statistic: 7.427 on 1 and 58 DF,  p-value: 0.00848
```

```
## Beta's by hand: 49.50572 -7.185448
## Beta's by R: 49.50572 -7.185448
##
## F by hand: 7.427115
## F by R: 7.427115
```

Is there a significant effect of the drug?

- There is a significant effect of drug, with drug delivery resulting in lower running times

c. Running Speed

Design Matrix

```
X_s <- matrix(c(intercept, x), ncol = 2)
```

```
## Column rank = 2
```

Beta, y_hat, error, and residual sum of squares (SSE)

```
Beta_s <- solve(t(X_s) %*% X_s) %*% t(X_s) %*% y
y_hat_s <- X_s %*% Beta_s
error_s <- y - y_hat_s

SSE_speed <- t(error_s) %*% error_s
```

Null hypothesis

```
X_s_H0 <- matrix(c(intercept), ncol = 1)
```

```
## Column rank = 1
```

```
Beta_s_H0 <- solve(t(X_s_H0) %*% X_s_H0) %*% t(X_s_H0) %*% y
y_hat_s_H0 <- X_s_H0 %*% Beta_s_H0
error_s_H0 <- y - y_hat_s_H0

SSE_speed_H0 <- t(error_s_H0) %*% error_s_H0
```

F-test

```
f_val_speed <- f_test(n = nrow(table1), p = 2, r = 1, SSE_speed_H0, SSE_speed)
```

```
## With F-statistic of 21.52721 and df 1, 58, p < 0.01
```

d. R Analysis

```
speed_reg <- lm(y ~ cov, data = table1)
summary(speed_reg)
```

```
##
## Call:
## lm(formula = y ~ cov, data = table1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.435  -6.584   0.299   5.562  19.930
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  63.2221     3.9176   16.14 < 0.0000000000000002 ***
## cov          -1.3034     0.2809   -4.64            0.0000204 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.262 on 58 degrees of freedom
## Multiple R-squared:  0.2707, Adjusted R-squared:  0.2581
## F-statistic: 21.53 on 1 and 58 DF,  p-value: 0.00002036
```

```
## Beta's by hand: 63.22215 -1.303392
## Beta's by R: 63.22215 -1.303392
##
## F by hand: 21.52721
## F by R: 21.52721
```

Does running speed significantly influence total running time?

- Running speed has a significant influence on total running time, with higher speed resulting in lower running time

e. Group as effect and Running Speed as covariate

Design matrix

```
X_sd <- matrix(c(intercept, x, dummies), ncol = 3)
```

```
## Column rank = 3
```

Beta, y_hat, error, and residual sum of squares (SSE)

```
Beta_sd <- solve(t(X_sd) %*% X_sd) %*% t(X_sd) %*% y
y_hat_sd <- X_sd %*% Beta_sd
error_sd <- y - y_hat_sd

SSE_sd <- t(error_sd) %*% error_sd
```

Null hypotheses can be used from above

F-test

```
f_val_sd1 <- f_test(n = nrow(table1), p = 3, r = 1, SSE_speed, SSE_sd)
```

```
## With F-statistic of 0.597761 and df 1, 57, p > 0.1
```

```
f_val_sd2 <- f_test(n = nrow(table1), p = 3, r = 1, SSE_drug, SSE_sd)
```

```
## With F-statistic of 13.01056 and df 1, 57, p < 0.01
```

f. R Analysis

```
sd_reg <- lm(y ~ cov + as.factor(group), data = table1)
summary(sd_reg)
```

```
##
## Call:
## lm(formula = y ~ cov + as.factor(group), data = table1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.0984  -5.8376   0.4796   5.8303  19.6629
##
## Coefficients:
##                   Estimate Std. Error t value             Pr(>|t|)
## (Intercept)        65.1546     4.6585  13.986 < 0.0000000000000002 ***
## cov                -1.5558     0.4313  -3.607             0.000653 ***
## as.factor(group)2   2.8389     3.6719   0.773             0.442628
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.294 on 57 degrees of freedom
## Multiple R-squared:  0.2783, Adjusted R-squared:  0.2529
## F-statistic: 10.99 on 2 and 57 DF,  p-value: 0.00009202
```

```
## Beta's by hand: 65.15458 -1.555792 2.838904
## Beta's by R: 65.15458 -1.555792 2.838904
##
## F by hand: 13.01056
## F by R: 10.98784
```

Does the drug have a significant effect? Is there a significant effect of running speed? How do you interpret your findings?

- There is a significant effect of only running speed, but not drug, with higher running speed decreasing running time. One interpretation might be a faster consumption of stamina at higher running speeds, resulting in shorter running times.

Question 2 — Regression interactions

```
setwd("/Users/joelvasama/Desktop/NB MSc/Advanced Stats/Exercises — Advanced Stats/
Previous/16 HW2")
table2 <- read.csv("Table2.csv")

n <- nrow(table2)
y <- table2$y
intercept <- rep(1, n)
treatment1 <- recode(table2$A, "1" = 1, "2" = 0, "3" = -1)
treatment2 <- recode(table2$A, "1" = 0, "2" = 1, "3" = -1)
gender <- recode(table2$B, "1" = -1, "2" = 1)
```

a. Treatment effect

Design matrix

```
X_t3 <- matrix(c(intercept, treatment1, treatment2), ncol = 3)
```

```
## Column rank = 3
```

Beta, y_hat, error, and residual sum of squares (SSE)

```
Beta_t3 <- solve(t(X_t3) %*% X_t3) %*% t(X_t3) %*% y
y_hat_t3 <- X_t3 %*% Beta_t3
error_t3 <- y - y_hat_t3

SSE_t3 <- t(error_t3) %*% error_t3
```

Null hypothesis

```
X_t3_H0 <- matrix(c(intercept), ncol = 1)
```

```
## Column rank = 1
```

```
Beta_t3_H0 <- solve(t(X_t3_H0) %*% X_t3_H0) %*% t(X_t3_H0) %*% y
y_hat_t3_H0 <- X_t3_H0 %*% Beta_t3_H0
error_t3_H0 <- y - y_hat_t3_H0

SSE_t3_H0 <- t(error_t3_H0) %*% error_t3_H0
```

F-test

```
f_val_t3 <- f_test(n = nrow(table2), p = 3, r = 2, SSE_t3_H0, SSE_t3)
```

```
## With F-statistic of 3.457554 and df 2, 117, p < 0.05
```

```
t3_reg <- lm(y ~ as.factor(A), data = table2)
summary(t3_reg)
```

```
##
## Call:
## lm(formula = y ~ as.factor(A), data = table2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.117 -11.200   1.194   9.239  23.913
##
## Coefficients:
##              Estimate Std. Error t value            Pr(>|t|)
## (Intercept)   99.78496    2.23354  44.676 <0.0000000000000002 ***
## as.factor(A)2  7.15833    3.15870   2.266              0.0253 *
## as.factor(A)3 -0.06979    3.15870  -0.022              0.9824
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.13 on 117 degrees of freedom
## Multiple R-squared:  0.05581,    Adjusted R-squared:  0.03967
## F-statistic: 3.458 on 2 and 117 DF,  p-value: 0.03476
```

```
rVsHand(Beta_t3, t3_reg$coefficients, f_val_t3, summary(t3_reg)$fstatistic[1])
```

```
## Beta's by hand: 102.1478 -2.362848 4.795481
## Beta's by R: 99.78496 7.158328 -0.0697852
##
## F by hand: 3.457554
## F by R: 3.457554
```

Is there a significant effect of the factor drug?

- Yes, drug does seem to have an influence as seen by the significant intercept (drug 1) and drug 2, but not placebo. Specifically, from looking at the Beta's by hand (due to R including drug 1 in the intercept, while by hand placebo is included in the intercept), drug 1 seems to decrease memory, while drug 2 increases memory.

b. Treatment and gender effect

Design matrix

```
int1 <- treatment1 * gender
int2 <- treatment2 * gender
X_t3_g <- matrix(c(intercept, treatment1, treatment2, gender, int1, int2), ncol =
6)

X_t3_g
```

```
##        [,1] [,2] [,3] [,4] [,5] [,6]
##   [1,]    1    1    0   -1   -1    0
##   [2,]    1    1    0   -1   -1    0
##   [3,]    1    1    0   -1   -1    0
##   [4,]    1    1    0   -1   -1    0
##   [5,]    1    1    0   -1   -1    0
##   [6,]    1    1    0   -1   -1    0
##   [7,]    1    1    0   -1   -1    0
##   [8,]    1    1    0   -1   -1    0
##   [9,]    1    1    0   -1   -1    0
##  [10,]    1    1    0   -1   -1    0
##  [11,]    1    1    0   -1   -1    0
##  [12,]    1    1    0   -1   -1    0
##  [13,]    1    1    0   -1   -1    0
##  [14,]    1    1    0   -1   -1    0
##  [15,]    1    1    0   -1   -1    0
##  [16,]    1    1    0   -1   -1    0
##  [17,]    1    1    0   -1   -1    0
##  [18,]    1    1    0   -1   -1    0
##  [19,]    1    1    0   -1   -1    0
##  [20,]    1    1    0   -1   -1    0
##  [21,]    1    0    1   -1    0   -1
##  [22,]    1    0    1   -1    0   -1
##  [23,]    1    0    1   -1    0   -1
##  [24,]    1    0    1   -1    0   -1
##  [25,]    1    0    1   -1    0   -1
##  [26,]    1    0    1   -1    0   -1
##  [27,]    1    0    1   -1    0   -1
##  [28,]    1    0    1   -1    0   -1
##  [29,]    1    0    1   -1    0   -1
##  [30,]    1    0    1   -1    0   -1
##  [31,]    1    0    1   -1    0   -1
##  [32,]    1    0    1   -1    0   -1
##  [33,]    1    0    1   -1    0   -1
##  [34,]    1    0    1   -1    0   -1
##  [35,]    1    0    1   -1    0   -1
##  [36,]    1    0    1   -1    0   -1
##  [37,]    1    0    1   -1    0   -1
##  [38,]    1    0    1   -1    0   -1
##  [39,]    1    0    1   -1    0   -1
##  [40,]    1    0    1   -1    0   -1
##  [41,]    1   -1   -1   -1    1    1
##  [42,]    1   -1   -1   -1    1    1
##  [43,]    1   -1   -1   -1    1    1
##  [44,]    1   -1   -1   -1    1    1
##  [45,]    1   -1   -1   -1    1    1
##  [46,]    1   -1   -1   -1    1    1
##  [47,]    1   -1   -1   -1    1    1
##  [48,]    1   -1   -1   -1    1    1
##  [49,]    1   -1   -1   -1    1    1
##  [50,]    1   -1   -1   -1    1    1
##  [51,]    1   -1   -1   -1    1    1
##  [52,]    1   -1   -1   -1    1    1
```

```
##  [53,]    1   -1   -1   -1    1    1
##  [54,]    1   -1   -1   -1    1    1
##  [55,]    1   -1   -1   -1    1    1
##  [56,]    1   -1   -1   -1    1    1
##  [57,]    1   -1   -1   -1    1    1
##  [58,]    1   -1   -1   -1    1    1
##  [59,]    1   -1   -1   -1    1    1
##  [60,]    1   -1   -1   -1    1    1
##  [61,]    1    1    0    1    1    0
##  [62,]    1    1    0    1    1    0
##  [63,]    1    1    0    1    1    0
##  [64,]    1    1    0    1    1    0
##  [65,]    1    1    0    1    1    0
##  [66,]    1    1    0    1    1    0
##  [67,]    1    1    0    1    1    0
##  [68,]    1    1    0    1    1    0
##  [69,]    1    1    0    1    1    0
##  [70,]    1    1    0    1    1    0
##  [71,]    1    1    0    1    1    0
##  [72,]    1    1    0    1    1    0
##  [73,]    1    1    0    1    1    0
##  [74,]    1    1    0    1    1    0
##  [75,]    1    1    0    1    1    0
##  [76,]    1    1    0    1    1    0
##  [77,]    1    1    0    1    1    0
##  [78,]    1    1    0    1    1    0
##  [79,]    1    1    0    1    1    0
##  [80,]    1    1    0    1    1    0
##  [81,]    1    0    1    1    0    1
##  [82,]    1    0    1    1    0    1
##  [83,]    1    0    1    1    0    1
##  [84,]    1    0    1    1    0    1
##  [85,]    1    0    1    1    0    1
##  [86,]    1    0    1    1    0    1
##  [87,]    1    0    1    1    0    1
##  [88,]    1    0    1    1    0    1
##  [89,]    1    0    1    1    0    1
##  [90,]    1    0    1    1    0    1
##  [91,]    1    0    1    1    0    1
##  [92,]    1    0    1    1    0    1
##  [93,]    1    0    1    1    0    1
##  [94,]    1    0    1    1    0    1
##  [95,]    1    0    1    1    0    1
##  [96,]    1    0    1    1    0    1
##  [97,]    1    0    1    1    0    1
##  [98,]    1    0    1    1    0    1
##  [99,]    1    0    1    1    0    1
## [100,]    1    0    1    1    0    1
## [101,]    1   -1   -1    1   -1   -1
## [102,]    1   -1   -1    1   -1   -1
## [103,]    1   -1   -1    1   -1   -1
## [104,]    1   -1   -1    1   -1   -1
## [105,]    1   -1   -1    1   -1   -1
## [106,]    1   -1   -1    1   -1   -1
```

```
## [107,]    1   -1   -1    1   -1   -1
## [108,]    1   -1   -1    1   -1   -1
## [109,]    1   -1   -1    1   -1   -1
## [110,]    1   -1   -1    1   -1   -1
## [111,]    1   -1   -1    1   -1   -1
## [112,]    1   -1   -1    1   -1   -1
## [113,]    1   -1   -1    1   -1   -1
## [114,]    1   -1   -1    1   -1   -1
## [115,]    1   -1   -1    1   -1   -1
## [116,]    1   -1   -1    1   -1   -1
## [117,]    1   -1   -1    1   -1   -1
## [118,]    1   -1   -1    1   -1   -1
## [119,]    1   -1   -1    1   -1   -1
## [120,]    1   -1   -1    1   -1   -1
```

```
## Column rank = 6
```

Beta, y_hat, error, and residual sum of squares (SSE)

```
Beta_t3_g <- solve(t(X_t3_g) %*% X_t3_g) %*% t(X_t3_g) %*% y
y_hat_t3_g <- X_t3_g %*% Beta_t3_g
error_t3_g <- y - y_hat_t3_g

SSE_t3_g <- t(error_t3_g) %*% error_t3_g
```

Null hypothesis

```
X_t3_g_H0_1 <- matrix(c(intercept, gender, int1, int2), ncol = 4)
X_t3_g_H0_2 <- matrix(c(intercept, treatment1, treatment2, int1, int2), ncol = 5)
X_t3_g_H0_3 <- matrix(c(intercept, treatment1, treatment2, gender), ncol = 4)
```

```
## Column rank = 4
```

```
##
## Column rank = 5
```

```
##
## Column rank = 4
```

Null hypothesis 1

```
Beta_t3_g_H0_1 <- solve(t(X_t3_g_H0_1) %*% X_t3_g_H0_1) %*% t(X_t3_g_H0_1) %*% y
y_hat_t3_g_H0_1 <- X_t3_g_H0_1 %*% Beta_t3_g_H0_1
error_t3_g_H0_1 <- y - y_hat_t3_g_H0_1

SSE_t3_g_H0_1 <- t(error_t3_g_H0_1) %*% error_t3_g_H0_1
```

F-test

```
f_val_t3_g_1 <- f_test(n = nrow(table2), p = 6, r = 2, SSE_t3_g_H0_1, SSE_t3_g)
```

```
## With F-statistic of 3.878847 and df 2, 114, p < 0.025
```

Null hypothesis 2

```
Beta_t3_g_H0_2 <- solve(t(X_t3_g_H0_2) %*% X_t3_g_H0_2) %*% t(X_t3_g_H0_2) %*% y
y_hat_t3_g_H0_2 <- X_t3_g_H0_2 %*% Beta_t3_g_H0_2
error_t3_g_H0_2 <- y - y_hat_t3_g_H0_2

SSE_t3_g_H0_2 <- t(error_t3_g_H0_2) %*% error_t3_g_H0_2
```

F-test

```
f_val_t3_g_2 <- f_test(n = nrow(table2), p = 6, r = 1, SSE_t3_g_H0_2, SSE_t3_g)
```

```
## With F-statistic of 9.638381 and df 1, 114, p < 0.01
```

Null hypothesis 3

```
Beta_t3_g_H0_3 <- solve(t(X_t3_g_H0_3) %*% X_t3_g_H0_3) %*% t(X_t3_g_H0_3) %*% y
y_hat_t3_g_H0_3 <- X_t3_g_H0_3 %*% Beta_t3_g_H0_3
error_t3_g_H0_3 <- y - y_hat_t3_g_H0_3

SSE_t3_g_H0_3 <- t(error_t3_g_H0_3) %*% error_t3_g_H0_3
```

F-test

```
f_val_t3_g_3 <- f_test(n = nrow(table2), p = 6, r = 2, SSE_t3_g_H0_3, SSE_t3_g)
```

```
## With F-statistic of 3.808851 and df 2, 114, p < 0.05
```

```
tableR <- table2
tableR$A <- recode(tableR$A, "1" = "drug1", "2" = "drug2", "3" = "placebo")
tableR$B <- recode(tableR$B, "1" = "male", "2" = "female")
colnames(tableR) <- c("y", "drug", "sex")

t3_g_reg <- lm(y ~ drug * sex, data = tableR)
summary(t3_g_reg)
```

```
##
## Call:
## lm(formula = y ~ drug * sex, data = tableR)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.881  -9.715   1.275  10.165  26.829
##
## Coefficients:
##                    Estimate Std. Error t value            Pr(>|t|)
## (Intercept)         102.701      2.982  34.438 <0.0000000000000002 ***
## drugdrug2            12.501      4.218   2.964              0.0037 **
## drugplacebo          -2.821      4.218  -0.669              0.5049
## sexmale              -5.832      4.218  -1.383              0.1694
## drugdrug2:sexmale   -10.685      5.964  -1.791              0.0759 .
## drugplacebo:sexmale   5.503      5.964   0.923              0.3581
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.34 on 114 degrees of freedom
## Multiple R-squared:  0.1799, Adjusted R-squared:  0.144
## F-statistic: 5.003 on 5 and 114 DF,  p-value: 0.0003518
```

Is there a significant main effect of drug, main effect of sex, and/or an interaction?

- There is a main effect of drug 1 (intercept), drug 2, and sex (female, in intercept), and an almost significant interaction of drug and sex.

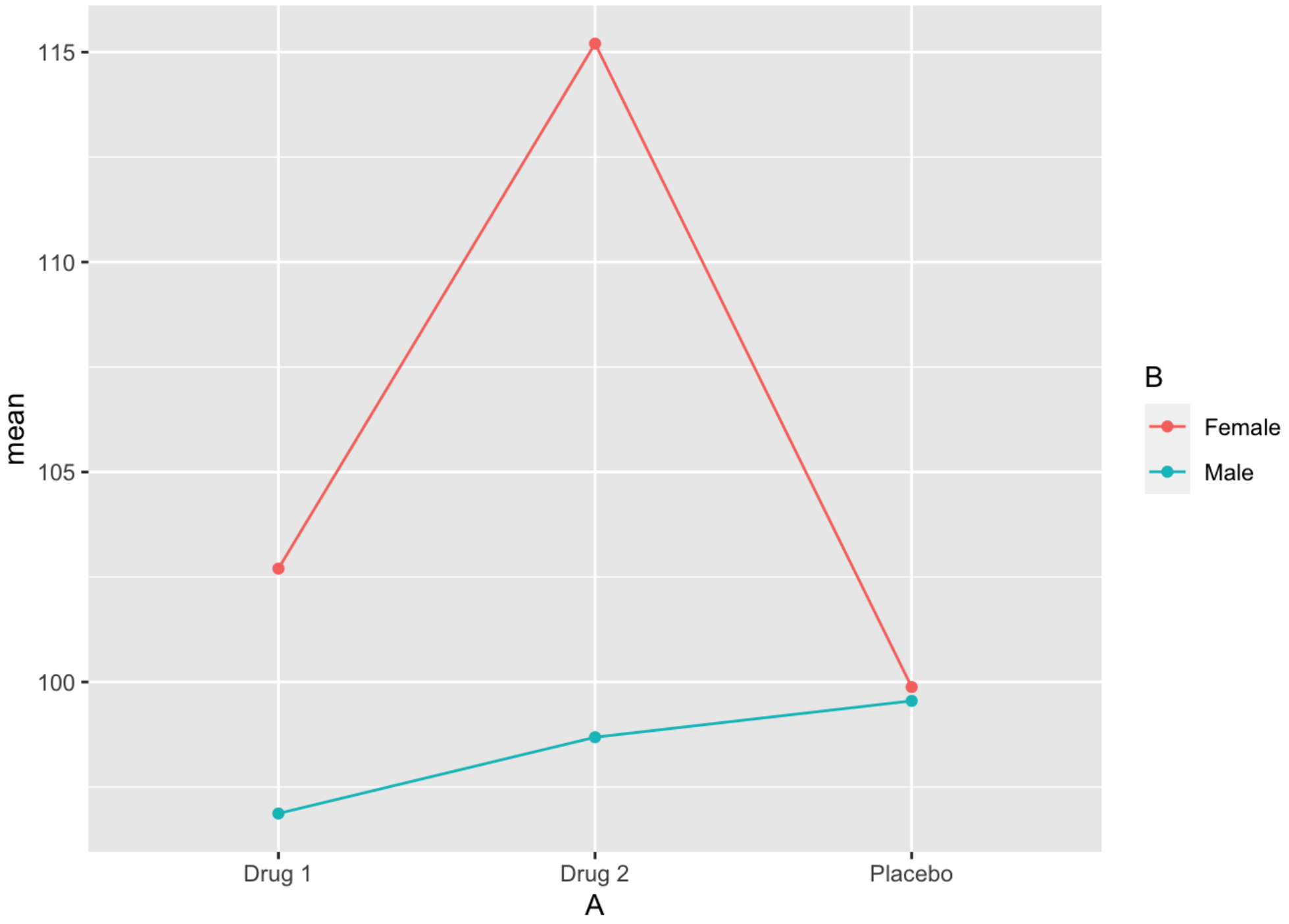
How do you interpret the findings from plots

```
plot_table <- table2 %>% group_by(A, B) %>% summarize(mean = mean(y))
```

```
## `summarise()` has grouped output by 'A'. You can override using the `.groups`
## argument.
```

```
plot_table$A <- recode(plot_table$A, "1" = "Drug 1", "2" = "Drug 2", "3" = "Placeb
o")
plot_table$B <- recode(plot_table$B, "1" = "Male", "2" = "Female")

ggplot(data = plot_table, aes(y = mean, x = A, color = B)) +
  geom_line(aes(group = B)) +
  geom_point()
```



- The influence of drug is increased for females and higher for drug 2