Practicum 2

Kieran Douglas

2. Examining Waugh's 1927 Asparagus Data

A.

We can see that the greatest difference between coeff estimates and those reported in the question are in the number of stalks and the variation in size (nostalks and disp). These end up being an absolute difference of 0.1767 and 0.0697 respecitvely.

```
# Set up environment and load data
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(readr)
library(fixest)

waugh <- read_table("/Users/kieran/Documents/MASTERS/METRICS/code/metrics/practicum_2_files/w
```

```
-- Column specification --------------------------------------------------------
cols(
```

```
  GREEN = col_double(),
  NOSTALKS = col_double(),
  DISPERSE = col_double(),
  PRICE = col_double()
)

Warning: 200 parsing failures.
row col  expected    actual
  1  -- 4 columns 5 columns '/Users/kieran/Documents/MASTERS/METRICS/code/metrics/practicum_1
  2  -- 4 columns 5 columns '/Users/kieran/Documents/MASTERS/METRICS/code/metrics/practicum_1
  3  -- 4 columns 5 columns '/Users/kieran/Documents/MASTERS/METRICS/code/metrics/practicum_1
  4  -- 4 columns 5 columns '/Users/kieran/Documents/MASTERS/METRICS/code/metrics/practicum_1
  5  -- 4 columns 5 columns '/Users/kieran/Documents/MASTERS/METRICS/code/metrics/practicum_1
... ... ......... ......... ..........................................................
See problems(...) for more details.
```

```
waugh_clean <- waugh |>
  rename(
    green = "GREEN",
    nostalks = "NOSTALKS",
    disp = "DISPERSE",
    price = "PRICE"
  ) |>
   mutate(
      green = as.numeric(green),
      nostalks = as.numeric(nostalks),
      disp = as.numeric(disp),
      price = as.numeric(price)
    )

# Run a MLR
model1 = lm(price ~ green+nostalks+disp, data = waugh_clean)
summary(model1)
```

```
Call:
lm(formula = price ~ green + nostalks + disp, data = waugh_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-56.004  -9.485  -0.122   9.422  49.097
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 40.761264   5.327837   7.651 8.82e-13 ***
green        0.137598   0.007099  19.382  < 2e-16 ***
nostalks    -1.357256   0.150822  -8.999  < 2e-16 ***
disp        -0.345283   0.129656  -2.663  0.00839 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.52 on 196 degrees of freedom
Multiple R-squared:  0.7268,    Adjusted R-squared:  0.7226
F-statistic: 173.8 on 3 and 196 DF,  p-value: < 2.2e-16
```

```r
# compare coefficient estimates
coeffs = coefficients(model1)
original_coeffs = c(green = 0.13826, nostalks = -1.53394, disp = -0.27554)
differences <- coeffs[names(original_coeffs)] - original_coeffs
print(differences)
```

```
        green      nostalks          disp
-0.0006617547  0.1766836397 -0.0697428425
```

```r
# We can see that the greatest difference between coeff estimates and those reported in the
```

B.

It looks like the main issue here is that Waugh has transformed green color on the asparagus variable to inches rather than what I have in the raw data, being hundredths of inches. To fix this I need to rescale the variable green by dividing it by 100. After doing this the difference in means is no longer a problem.

```r
means <- c(price_avg = mean(waugh_clean$price), green_avg = mean(waugh_clean$green), nostalks
print(means)
```

```
   price_avg    green_avg nostalks_avg     disp_avg
      90.095      588.750       19.555       14.875
```

```r
means_reported <- c(price_avg = 90.095, green_avg = 5.8875, nostalks_avg = 19.555, disp_avg =
differences_avg <- means[names(means_reported)] - means_reported
print(differences_avg)
```

```
    price_avg     green_avg  nostalks_avg      disp_avg
 0.000000e+00  5.828625e+02 -3.552714e-15  0.000000e+00
```

```
# It looks like the main issue here is that Waugh has transformed green color on the asparag
waugh_clean <- waugh_clean |>
  mutate(
    green = green/100
  )
# Recheck mean differences
means_recheck <- c(price_avg = mean(waugh_clean$price), green_avg = mean(waugh_clean$green),
print(means_recheck)
```

```
   price_avg    green_avg nostalks_avg     disp_avg
     90.0950       5.8875      19.5550      14.8750
```

C.

I notice relative variance in the relative size of the covariances reported. For example, the covariance between price and green according to Waugh is ~3430 while the one I found was 3448 (larger). On the other hand, Waugh found a covairance of ~-154 for green and disp while I found one of ~-180 (mine was smaller). In other cases though mine was larger like with the cov between green and price. I think that the pattern in differences may relate to the way my price variable is coded relative to Waugh's. This would explain the constant differences in price covariance versus the more consistent findings in ohter categories.

```
# I am now going to create a variance covariance matrix similar to Waugh's`
moments_subset <- waugh_clean[, c("price", "green", "nostalks", "disp")]
# to match the layout of the table provided I will modify green to be back to hundredths of a
moments_subset <- moments_subset |>
  mutate(
    green = green*100
  )
# Create the matric, format it correctly, and print
matrix <- cov(moments_subset)
print(matrix)
```

```
              price       green  nostalks       disp
price     868.73967  3448.18467 -93.38465  -87.43028
green    3448.18467 24439.38442 -17.09171 -180.05653
nostalks  -93.38465   -17.09171  60.73063   24.92399
disp      -87.43028  -180.05653  24.92399   83.48681
```

```
matrix_fmt <- formatC(matrix, format="f", digits=1)
matrix_fmt[lower.tri(matrix_fmt)] <- ""
print(noquote(matrix_fmt))
```

```
         price green   nostalks disp
price    868.7 3448.2  -93.4    -87.4
green          24439.4 -17.1    -180.1
nostalks               60.7     24.9
disp                            83.5
```

D.

It is plausible that the principle findings concerning the effects of covariate variation could be affected in terms of the precision and statistical inference of coefficient estimates if there was underestimation or misreporting of variance and covariance among regressors. Depsite OLS estimates remaining unbiased under GM, Waugh's under or over-estiamted variance and coveriance could lead to underestiated standard errors and possibly inacurate significance. Similarly, cov among the regressors included may inflate the standard errors and lead to reductions in the validity/reliability of estimates. The difference in scaled coefficients between Waugh's and my own analysis do not differ too much, but there are some clear differences. For example, my estimate for the marginal effect of a unit increase in the number of stalks per bunch is about \$0.5 less than Waugh's. I estimate a slightly higher increase in price given a unit increase in green, and a roughly \$0.2 larger decrease in price given a unit increase in size variation. Overall, our estimates are very similar. I find the association between both green and nostalks with price to be highly statistically significant at the $p<0.001$ level and the association between disp and price to be statistically signifiacnt at **$0.001<p<0.01$** level. All of these observations are ceterus parabus.

::: {.cell}

```
 # run model
coef(model1) * 2.782
```

::: {.cell-output .cell-output-stdout}

```
(Intercept)       green    nostalks        disp
113.3978351   0.3827983  -3.7758872  -0.9605769
```

:::

```
waugh_scaled <- original_coeffs*2.782
print(waugh_scaled)
```

::: {.cell-output .cell-output-stdout}

```
     green   nostalks       disp
 0.3846393 -4.2674211 -0.7665523
```

:::

```
# The difference in scaled coefficients between Waugh's and my own analysis do not differ too
```

:::

E.

To solve this I am going to experiment with a matrix.

```
# my matrix solutions
# Extract covariance matrix of regressors (columns and rows 2 to 4)
xx <- matrix[2:4, 2:4]
# Extract covariance vector of PRICE with regressors (row 1, columns 2 to 4)
xy <- matrix[1, 2:4]
# Compute beta_hat = solve(Sigma_XX) %*% Sigma_XY
beta_hat <- solve(xx) %*% xy
print(beta_hat)
```

```
              [,1]
green     0.1375982
nostalks -1.3572564
disp     -0.3452828
```

```
# Waugh's matrix solutions
vc_mat <- matrix(c(
  1063.64, 3430.89, -100.92, -82.35,
  3430.89, 24317.19, -17.01, -154.54,
  -100.92, -17.01, 61.33, 25.51,
  -82.35, -154.54, 25.51, 83.07
), nrow = 4, byrow = TRUE)
```

```r
# names for reference
row_col_names <- c("price","green","nostalks","disp")
dimnames(vc_mat) <- list(row_col_names, row_col_names)
# price-cov covariance vector
Sigma_yX <- vc_mat["price", c("green", "nostalks", "disp")]
# covariate-covariance matrix
Sigma_XX <- vc_mat[c("green", "nostalks", "disp"), c("green", "nostalks", "disp")]
# raw coefficients from matrix
beta_raw <- solve(Sigma_XX, Sigma_yX)
names(beta_raw) <- c("green", "nostalks", "disperse")
# scale
beta_final <- beta_raw * 2.782
beta_final
```

```
     green   nostalks   disperse
 0.3847304 -4.1520749 -0.7670883
```

```r
# Computing the least squares estimates from Waugh's VC matrix table, we can see that his est
```

3. Exploring Relationships among R^2, Coefficients of Determination, and Correlation Coefficients

A.