ARE 212 Problem Set 3

Eleanor Adachi, Karla Neri, Anna Cheyette, Stephen Stack, Aline Abayo

2024-02-14

```
options(scipen = 999)

# Load packages
library(pacman)
p_load(tidyverse, haven, knitr, psych, stats4, stargazer, magrittr,
       qwraps2, Jmisc, fastDummies)

# get directory of current file
current_directory <-
  dirname(dirname(rstudioapi::getSourceEditorContext()$path))

# read in data
my_data <- read_dta(file.path(current_directory, "data", "pset3_2024.dta"))
```

Question 1

```
my_data %>%
  # sum the number of NA values across all variables
  summarise(across(everything(), ~sum(is.na(.)))) %>%
  glimpse()
```

```
## Rows: 1
## Columns: 28
## $ year        <int> 0
## $ country     <int> 0
## $ co          <int> 0
## $ type        <int> 0
## $ segment     <int> 0
## $ domestic    <int> 0
## $ firm        <int> 0
## $ brand       <int> 0
## $ loc         <int> 0
## $ qu          <int> 0
## $ pr          <int> 0
## $ princ       <int> 0
## $ price       <int> 0
## $ horsepower  <int> 0
## $ fuel        <int> 0
## $ width       <int> 0
## $ height      <int> 0
## $ weight      <int> 0
```

```
## $ pop          <int> 0
## $ ngdp         <int> 0
## $ ngdpe        <int> 0
## $ country1     <int> 0
## $ country2     <int> 0
## $ country3     <int> 0
## $ country4     <int> 0
## $ country5     <int> 0
## $ yearsquared <int> 0
## $ luxury       <int> 0
```

The table above, which sums the number of NA values across all columns in the data, shows that we do not have any missing values - all sum to 0.

Question 2

```
price_summary <-
  list("Price" =
         list("min" = ~ min(my_data$price),
              "max" = ~ max(my_data$price),
              "mean (sd)" = ~ qwraps2::mean_sd(my_data$price)))

whole <- summary_table(my_data, price_summary)
whole
```

	my_data (N = 57)
Price	
min	8.13041400909424
max	48.0754890441895
mean (sd)	20.77 ± 8.54

```
# get n
n <- nrow(my_data)
# df = n - 1 = 56
(df <- n - 1)
```

```
## [1] 56
```

```
# get sample mean
(pbar <- mean(my_data$price))
```

```
## [1] 20.77293
```

```
# Compute squared differences from the mean
squared_diffs <- (my_data$price - pbar)^2

# Sum the squared differences
sum_squared_diffs <- sum(squared_diffs)

# Divide by the number of observations minus one to get the variance
variance <- sum_squared_diffs / (n - 1)

# Take the square root of the variance to get the standard deviation
s <- sqrt(variance)

# Calculate standard error of the mean
```

```r
(se <- s / sqrt(n))
```

```
## [1] 1.131724
```

```r
# Find the t critical value for a 99% CI with df degrees of freedom
# 0.995 because it's one-tailed (99% CI = 0.5 + 0.99/2)
(t_critical <- qt(0.995, df, lower.tail = T))
```

```
## [1] 2.666512
```

```r
# Calculate the margin of error
(me <- t_critical * se)
```

```
## [1] 3.017757
```

```r
# Construct the confidence interval
lower_bound <- pbar - me
upper_bound <- pbar + me

# Display the confidence interval
c(lower_bound, upper_bound)
```

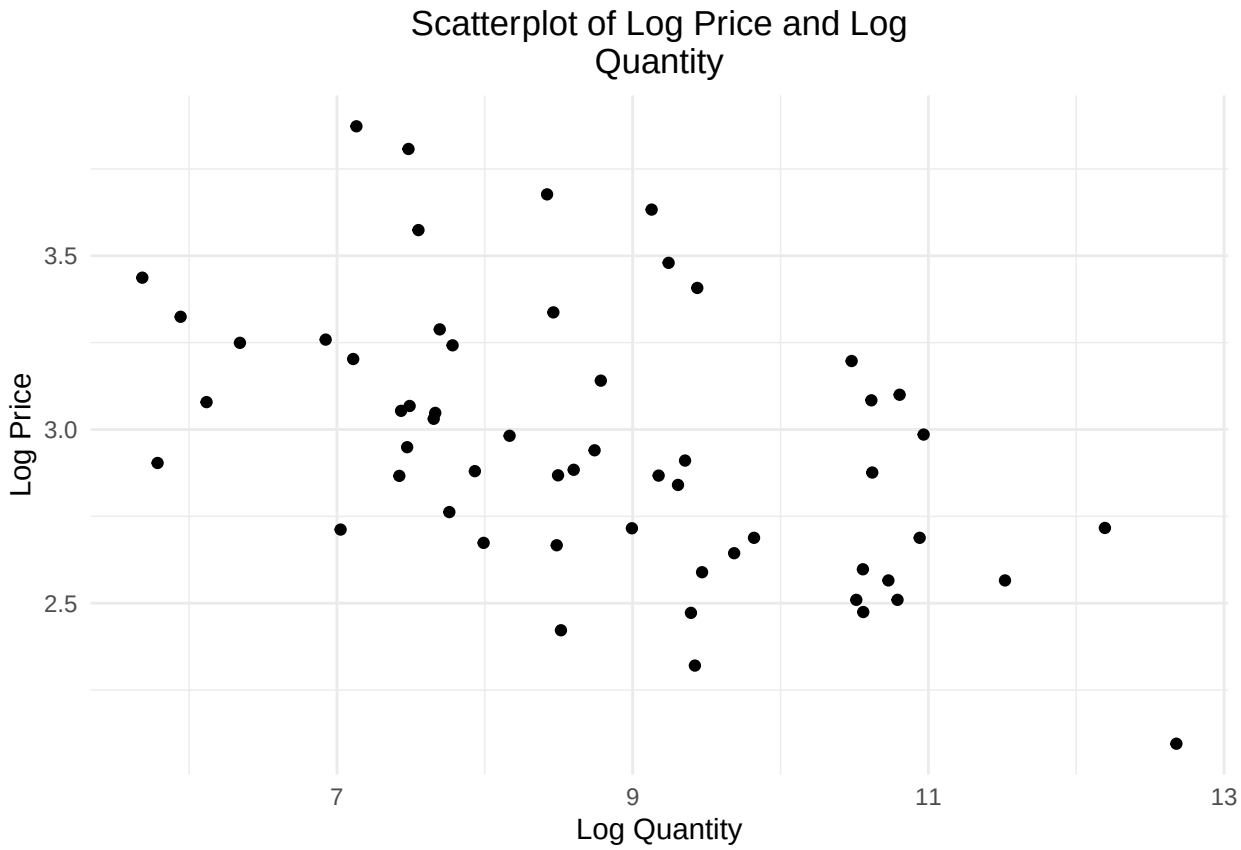
```
## [1] 17.75517 23.79069
```

Question 3

```r
# Create two new variables log of price and log of quantity, lprice and lqu
# Create the scatter plot of the two variables lqu and lprice
# What is the estimated OLS linear model slope associated with this scatter
# plot? Estimate a regression to answer this.

# create log price and log quantity
my_data <-
  my_data %>%
  mutate(lprice = log(price),
         lqu = log(qu))

my_data %>%
  ggplot(aes(x=lqu, y=lprice)) +
  geom_point() +
  theme_minimal() +
  labs(x = "Log Quantity",
       y = "Log Price",
       title =
         str_wrap("Scatterplot of Log Price and Log Quantity", 40)) +
  theme(plot.title = element_text(hjust = 0.5))
```




```
y_3 <- my_data$lprice

X_3 <- cbind(1, my_data$lqu)

# find coefficient
b_3 <- solve(t(X_3) %*% X_3) %*% t(X_3) %*% y_3
b_3
```

```
##             [,1]
## [1,]  4.0055928
## [2,] -0.1190505
```

```
# do we need to do more than this?? this gives the slope
```

Question 4

```
# create y for log quantity
y_4_qu <- my_data$lqu

# create X = fuel, luxury, domestic, and a constant
X_4 <- cbind(1, my_data$fuel, my_data$luxury, my_data$domestic)

# find coefficients for log quantity
(b_4_qu <- solve(t(X_4) %*% X_4) %*% t(X_4) %*% y_4_qu)
```

```
##             [,1]
## [1,] 10.9239821
```

```
## [2,] -0.3049930
## [3,] -0.2616135
## [4,]  2.4008093
```

```r
# projection matrix, P
P <- X_4 %*% solve(t(X_4) %*% X_4) %*% t(X_4)

# residual maker, M = I - P
M <- diag(nrow(my_data)) - P

# calculate residuals for log quantity
elqu <- M %*% y_4_qu

# create y for log price
y_4_price <- my_data$lprice

# find coefficients for log price
(b_4_pr <- solve(t(X_4) %*% X_4) %*% t(X_4) %*% y_4_price)
```

```
##             [,1]
## [1,]  1.82649309
## [2,]  0.12496107
## [3,]  0.50002347
## [4,] -0.02752284
```

```r
# calculate residuals
elprice <- M %*% y_4_price

# Scatter plot the residuals elqu on vertical axis and elprice
# on horizontal axis
ggplot() +
  geom_point(aes(x = elprice, y = elqu)) +
  theme_minimal() +
  labs(x = "Residuals - Log Price",
       y = "Residuals - Log Quantity",
       title =
         str_wrap("Scatterplot of Residuals - Log Price and Log Quantity", 40)) +
  theme(plot.title = element_text(hjust = 0.5))
```


```
# What is the estimated OLS slope associated with this scatter plot? Estimate a
# regression (no constant) to
# answer this and explain what theorem underlies the fact that this slope is
# the marginal effect of lprice on lqu
# in a regression that also features fuel,luxury, domestic, and a constant.

# find coefficient
(b_4_e <- solve(t(elprice) %*% elprice) %*% t(elprice) %*% elqu)
```

```
##           [,1]
## [1,] -3.335674
```

```
# create X and y
X_4_e <- elprice
y_4_e <- elqu
# projection matrix, P
P <- X_4_e %*% solve(t(X_4_e) %*% X_4_e) %*% t(X_4_e)
# residual maker, M = I - P
M <- diag(nrow(X_4_e)) - P

# calculate residuals of regression of residuals of lprice and lquantity
residual_residuals <- M %*% y_4_e
```

What is the estimated OLS slope associated with this scatter plot? Estimate a regression (no constant) to answer this and explain what theorem underlies the fact that this slope is the marginal effect of lprice on lqu in a regression that also features fuel,luxury, domestic, and a constant.

The theorem underlying the fact that the slope of the scatterplot between $\hat{e}_{lprice}$ and $\hat{e}_{lqu}$ is the marginal effect of *lprice* and *lqu* in a regression that also includes fuel, luxury, domestic, and a constant is the

Frisch-Waugh-Lovell theorem. This theorem states that in a linear regression model in which y (lqu) is regressed on a set of variables X (fuel, luxury, domestic, and a constant) and another variable z, the coefficient on z obtained from this regression is the same as the coefficient obtained from regressing the residuals of y on X on the residuals of z on X without a constant.

Question 5

The slope in question 3 is not the same as the one in question 4 because question 4 regresses the residuals of `lqu` on the residuals of `lprice` after both have been adjusted for other variables (fuel, luxury, domestic). Question 3 does not consider other variables in the regression of `lqu` on `lprice.` The slope from question 3 represents the simple bivariate relationship between `lqu` and `lprice`, not controlling for any other factors.

Theoretically, the slopes from steps 3 and 4 would be equal if the variables fuel, luxury, and domestic had no effect on `lqu` and `lprice`—if these control variables were unrelated to both the dependent variable and the independent variable In this scenario, adjusting for these variables would not change the estimated relationship between `lqu` and `lprice`, because their inclusion in the regression model would not account for any additional variance in `lqu` that is associated with `lprice`.

Question 6

The point estimate from question 4 is -3.3356739 and the sign is negative. This indicates that, holding all else equal, for every 1% increase in price, we expect a 3.3356739% decrease in quantity.

```
# sum of squared residuals
ssr <- as.numeric(t(residual_residuals)%*%residual_residuals)
# Number of observations
n <- nrow(residual_residuals)
# Number of parameters estimated
k <- ncol(X_4_e)
(df <- n-k)
```

```
## [1] 56
```

```
(sigma_squared <- ssr / df)
```

```
## [1] 0.8425636
```

```
(var_cov_matrix <- sigma_squared * solve(t(elprice) %*% elprice))
```

```
##           [,1]
## [1,] 0.3260789
```

```
(se_beta <- sqrt(diag(var_cov_matrix)))
```

```
## [1] 0.5710332
```

```
(t_statistic_6 <- b_4_e / se_beta)
```

```
##           [,1]
## [1,] -5.841471
```

```
# Since it's a two-tailed test, double the one-tailed p-value
(p_value_two_tailed <- 2 * pt(t_statistic_6, df))
```

```
##                 [,1]
## [1,] 0.0000002748858
```

What is the pvalue for the estimated lprice coefficient? Using the t-table, we find the nearest number below our degrees of freedom (which is 40, in this case) and move to the right to find the t-value. Since our test

statistic is |-5.8414711| and the largest t-statistic in the table for 40 df is 3.551, we can conclude that our p-value for a two-sided test is < 0.001.

Question 7

```
# null hypothesis: marginal effect of lprice on lqu is -4,
# all else equal.
# alternative hypothesis: marginal effect of lprice on lqu
# is not equal to -4.

# null hypothesis
beta_0 <- -4
# Calculate the t-statistic
(t_statistic_7 <- (b_4_e - beta_0) / se_beta)
```

```
##          [,1]
## [1,] 1.163376
```

```
# Calculate the p-value for the two-tailed test
(p_value_two_tailed <- 2 * pt(-abs(t_statistic_7), df))
```

```
##           [,1]
## [1,] 0.2496103
```

```
# Determine the critical value for the two-tailed test at the 5% significance level
alpha <- 0.05
(critical_value <- qt(1 - alpha/2, df))
```

```
## [1] 2.003241
```

Using the t-table for the nearest (smaller) degrees of freedom value to 56 (which is 40), we find the critical value equal to 2.021 for a 95% confidence leve. Our test statistic is 1.1633757, which is less than the critical value. Thus, we cannot reject the null hypothesis that the marginal effect of lprice on lqu is -4, all else equal.

Question 8

```
variables <- my_data[, c("lqu", "lprice", "fuel", "weight", "luxury", "domestic")]

# Calculate the correlation matrix
# ** Are we allowed to use `cor` ?
(correlation_matrix <- cor(variables))
```

```
##                lqu     lprice       fuel      weight      luxury   domestic
## lqu       1.0000000 -0.5121756 -0.3030351 -0.44153421 -0.12611097 0.57723120
## lprice   -0.5121756  1.0000000  0.7739287  0.91246510  0.50805990 0.08792450
## fuel     -0.3030351  0.7739287  1.0000000  0.82759388  0.30973689 0.15311909
## weight   -0.4415342  0.9124651  0.8275939  1.00000000  0.50441087 0.04385066
## luxury   -0.1261110  0.5080599  0.3097369  0.50441087  1.00000000 0.04803253
## domestic  0.5772312  0.0879245  0.1531191  0.04385066  0.04803253 1.00000000
```

The omitted variable from model (8.b) that is included in model (8.a) but could be most concerning for bias is domestic, given its relationship with lqu (0.577) and fuel.

The presence of a positive correlation between fuel and lprice, as well as weight, indicates that any omitted variable related to these aspects (which also influences lqu) could lead to bias in the estimation of the fuel coefficient if not controlled for.

Without controlling for domestic, the estimated coefficient for fuel in model (8.b) might have a positive bias due to the omission of the domestic variable, which is positively correlated with both lqu and fuel to a lesser extent. This suggests that part of the positive effect of domestic cars on lqu could be inaccurately attributed to fuel, leading to an overestimation of the positive impact of fuel efficiency or cost on lqu.

Question 9

Question 10

```
# add advertising variable to my_data
my_data <-
  my_data %>%
  mutate(advertising = 5 * lprice)

# create y equal to log quantity
y_10 <- my_data$lqu

# create X = lprice, fuel, luxury, advertising,
# and a constant
X_10 <- cbind(1, my_data$lprice,
              my_data$fuel,
              my_data$luxury,
              my_data$advertising)

# find coefficients for log quantity
#(b_10_qu <- solve(t(X_10) %*% X_10) %*% t(X_10) %*% y_10)
```

"Error in solve.default(t(X_10) %*% X_10) : system is computationally singular: reciprocal condition number = 5.40872e-18"

We encounter an error when trying to run this regression because it is singular - advertising is perfectly collinear with lprice. **more to say here?**

Question 11

Regression model:

$lqu = \alpha + \beta_1 * lprice + \beta_2 * luxury + \beta_3 * (lprice * luxury) + \epsilon$

$H_0 : \beta_3 = 0$ $H_1 : \beta_3 \neq 0$

```
# test the null hypothesis that the marginal effect# in lprice on lqu does not differ by luxury classif

# Estimate the regression model with interaction term
model <- lm(lqu ~ lprice + luxury + lprice:luxury,
            data = my_data)
summary(model)
```

```
##
## Call:
## lm(formula = lqu ~ lprice + luxury + lprice:luxury, data = my_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0714 -0.9800 -0.1195  0.9132  2.8555
```

```
##
## Coefficients:
##               Estimate Std. Error t value          Pr(>|t|)
## (Intercept)     16.314      1.696   9.620 0.000000000000321 ***
## lprice          -2.568      0.578  -4.443 0.000045581608943 ***
## luxury          23.848     30.386   0.785             0.436
## lprice:luxury   -5.984      8.063  -0.742             0.461
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.41 on 53 degrees of freedom
## Multiple R-squared:  0.2939, Adjusted R-squared:  0.2539
## F-statistic: 7.353 on 3 and 53 DF,  p-value: 0.0003289
```

Based on the summary output, the coefficient of the interaction of $lprice * luxury$ is -5.984, with a t-value of -0.742. The p-value is 0.461, which is greater than 0.05, so we fail to reject the null hypothesis that that the marginal effect of lprice on lqu does not differ between luxury and nonluxury goods.