MSIN0094 Individual Coursework 2

by Marfa Popova

Submission date: 01-Dec-2021 09:59PM (UTC+0000)
Submission ID: 164611592
File name: MSIN0094_Individual_Coursework_2_3112889_462880875.pdf (254.24K)
Word count: 2711
Character count: 14156

MSIN0094 Second Assignment

Due Friday December 3rd, 10am London Time for SORA students

Dataset Manipluation (30pts)

1.

a.

```
## write your code of joining below
data_full <- data_product %>%
  left_join(data_sales, by = "product_id")
```

4

b.

```
# write you code below to generate final_price
data_full <- data_full %>%
  mutate(final_price = RRP * (1 - discount))
```

2

c.

```
# write you code below
mean_fp = mean(data_full $ final_price)
mean_fp # where mean_fp = mean of final_price
```

```
## [1] 2026.096
```

4

```
sd_fp = sd(data_full $ final_price)
sd_fp # where sd_fp = standard deviation of final_price
```

```
## [1] 1211.942
```

2.

a.

```
# join the dataset
data_full <- data_full %>%
  left_join(data_marketing, by = c("week_id","brand"))
```

4

b.

```r
# Compute the mean and standard deviation of marketing_expense below
mean_me = mean(data_full $ marketing_expense)
mean_me # where mean_me = mean of marketing_expense
```

```
## [1] 131.874
```

```r
sd_me = sd(data_full $ marketing_expense)
sd_me # where sd_me = standard deviation of marketing_expense
```

```
## [1] 52.05325
```

```r
summary(data_full$marketing_expense)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   16.15  114.07  129.55  131.87  161.31  239.02
```

3.

a.

```r
cor(data_full $ marketing_expense, data_full $ RRP)
```

```
## [1] 0.4035218
```

b.

```r
cor.test(data_full$marketing_expense, data_full$RRP)
```

```
##
##  Pearson's product-moment correlation
##
## data:  data_full$marketing_expense and data_full$RRP
## t = 68.943, df = 24438, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3929727 0.4139648
## sample estimates:
##       cor
## 0.4035218
```

```r
#or:
#cor.test(formula = ~ marketing_expense + RRP, data = data_full)
```

The smaller the p-value, the stronger the evidence that we should reject the null hypothesis, as usually, a p-value <= 0.05 is statistically significant, as there is less than a 5% probability the null is correct (and the results are random). In our case, we conducted the Pearson's product-moment correlation which revealed a positive correlation between marketing_expense and RRP to be approximately 0.4. The correlation's strength is moderate, and it is statistically significant, as p-value < 2.2e-16, which is very small.

c. Recommended price set by retailers (RRP) dictates the final price, which ultimately, affects the marketing expense based on the revenue prediction. The correlation of 0.4 between RRP and marketing expense explains this positive linear relationship: the higher the RRP, the higher the marketing expense is feasible, as larger revenues are expected - however, higher marketing expenses might not increase profit margins as much, as sunken costs would be higher.

Preliminary Customer Analysis (28 pts)

4.

```
pacman::p_load(dplyr)
# Number of DISTINCT required TV models
n_distinct(data_full %>% filter(brand == "LG", resolution == "1080p", support_HDR == 1, technology =="L
%>% distinct(product_id))
```

```
## [1] 16
```

```
#Code below calculates overall number of TVs with the required parameters
nrow(data_full %>% filter(brand == "LG", resolution == "1080p", support_HDR == 1, technology == "LCD"))
```

```
## [1] 832
```

5.

```
data_product %>% group_by(brand) %>% filter(resolution == "4k", technology == "OLED") %>%
  summarize(number = n()) %>% arrange(number) %>% ungroup()
```

```
## # A tibble: 2 x 2
##   brand number
##   <chr>  <int>
## 1 LG        32
## 2 Sony      48
```

Samsung and Philips are excluded from the table as they were filtered out by the "resolution" and "technology" parameters. The results above are in the ascending order, with LG having 32 units, and Sony 48.

6.

a.

```
data_product %>% filter(resolution == "4k", technology == "OLED" | technology == "LCD") %>%
  group_by(technology) %>% summarize(mean = mean(RRP)) %>% ungroup()
```

```
## # A tibble: 2 x 2
##   technology  mean
##   <chr>      <dbl>
## 1 LCD        1572.
## 2 OLED       3508.
```

b.

```
t.test(formula = RRP ~ technology,
       data = data_product %>%
         filter(resolution == "4k", technology == "OLED" | technology == "LCD"))
```

```
##
##  Welch Two Sample t-test
##
## data:  RRP by technology
## t = -14.301, df = 122.51, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group LCD and group OLED is not equal to 0
## 95 percent confidence interval:
##  -2203.752 -1667.847
## sample estimates:
##  mean in group LCD mean in group OLED
##           1571.919           3507.718
```

Q6b 3

I chose to do a t-test, which gives a p-value < 2.2e-16. Thus, the mean RRP across the above 2 groups are statistically different.

c.

OLED TVs are more expensive than LCD TVs because they have better picture quality. Firstly, "when it comes to black levels, OLED reigns as the undisputed champion", compared to LED TVs that rely on LED backlights shining behind an LCD panel. (Bizzaco, Cohen and Lacoma, 2019) These create so-called "light bleeds", where lighter sections of the screen create a haze or bloom in adjacent darker areas. Meanwhile, "if an OLED pixel isn't getting electricity, it doesn't produce any light and is, therefore, totally black." (ibid.) Additionally, "OLED is lighter and thinner, uses less energy, offers the best viewing angle by far, and, though still a little more expensive, has come down in price considerably." (ibid.) 2

7.

```
data_full %>% group_by(brand) %>% summarise(total_sales = sum(sales)) %>%
 arrange(- total_sales) %>% ungroup()
```

```
## # A tibble: 4 x 2
##   brand   total_sales
##   <chr>         <int>
## 1 Sony          76175
## 2 Samsung       73458
## 3 LG            56564
## 4 Philips        4380
```

6

Based on the table above, the 4 TV brands over all weeks' sales are sorted in the descending order: Sony = 76175, Samsung = 73458, LG = 56564, Philips = 4380.

8.

```
head(data_full %>% filter(brand == "Samsung", technology == "QLED") %>%
       group_by(week_id) %>% summarize(total_weekly_sales = sum(sales)) %>%
       arrange(-total_weekly_sales)) %>% ungroup()
```

4

```
## # A tibble: 6 x 2
##   week_id total_weekly_sales
##     <int>              <int>
## 1      17               1388
```

```
## 2       35              1367
## 3       19              1290
## 4       28              1260
## 5        2              1238
## 6        5              1230
```

The 17th week had the highest number of sales for Samsung's "QLED" TVs.

Simple Linear Regression (28 pts)

9.

a.

```
# write your codes for the regression below
q9a <- lm(data = data_full, formula = sales ~ final_price)
summary(q9a)
```

```
##
## Call:
## lm(formula = sales ~ final_price, data = data_full)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.094  -4.575  -2.137   1.551 170.476
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.201e+00  1.124e-01   19.59   <2e-16 ***
## final_price 3.166e-03  4.759e-05   66.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.016 on 24438 degrees of freedom
## Multiple R-squared:  0.1534, Adjusted R-squared:  0.1533
## F-statistic:  4426 on 1 and 24438 DF,  p-value: < 2.2e-16
```

b. The estimated coefficient b means that when the final price increases by 1 unit, the sales increase by 3.166 * 10-3 units. Since p-value <2e-16, it is a statistically significant result, which has a significant code of 0. The R-squared = 0.1534, meaning that 15.34% of the sales variation (i.e. its change) can be explained by the final price.

10. Run a univariate regression specified as follows (**7pts** in total)

a.

```
# write your codes for the regression below
q10a = lm(data = data_full, formula = log(sales+0.01) ~ log(final_price))
summary(q10a)
```

```
##
## Call:
## lm(formula = log(sales + 0.01) ~ log(final_price), data = data_full)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4419 -0.5094  0.0457  0.6268  3.5366
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -5.157582   0.062519   -82.5   <2e-16 ***
## log(final_price)  0.915931   0.008426   108.7   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.028 on 24438 degrees of freedom
## Multiple R-squared:  0.3259, Adjusted R-squared:  0.3259
## F-statistic: 1.182e+04 on 1 and 24438 DF,  p-value: < 2.2e-16
```

b. The estimated coefficient b means that when the final price increases by 1%, the sales increase by 0.92% (2 d.p.). Hence, the relationship between *log(sales+0.1)* and *log(final_price)* seems almost perfectly proportional, 1:1. Since p-value <2e-16, it is a statistically significant result which has a significant code of 0. R-squared = 0.3259, meaning that 32.59% of the statistical variation (i.e. its change) can be explained by the final price.

c. R-squared in q10a is over two times bigger than in q9a. Hence,it would be useful to learn more about the relationship between the explanatory variable and the outcome variable. Therefore, we should take the log-transformation, as it would better explain the linear relationship between the control- and outcome variables.

11.

a.

```
#design your regression below
q11a = lm(data = data_full, formula = log(sales+0.01) ~ factor(brand) + factor(technology) + factor(res
summary(q11a)
```

```
##
## Call:
## lm(formula = log(sales + 0.01) ~ factor(brand) + factor(technology) +
##     factor(resolution) + factor(energy_class) + support_HDR +
##     factor(refresh_rate) + factor(screensize) + log(final_price),
##     data = data_full)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3208 -0.3336  0.0390  0.4027  2.7265
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     15.56412    0.32917  47.283  < 2e-16 ***
## factor(brand)Philips            -3.18772    0.04942 -64.499  < 2e-16 ***
```

```
## factor(brand)Samsung                  0.57416    0.01794  32.007  < 2e-16 ***
## factor(brand)Sony                    -0.13490    0.01352  -9.978  < 2e-16 ***
## factor(technology)OLED                2.66545    0.03492  76.323  < 2e-16 ***
## factor(technology)QLED                2.65479    0.03872  68.569  < 2e-16 ***
## factor(resolution)4k                  1.31955    0.01749  75.445  < 2e-16 ***
## factor(resolution)720p               -1.12353    0.02895 -38.809  < 2e-16 ***
## factor(energy_class)A+               -0.01191    0.01077  -1.106  0.26856
## factor(energy_class)B                -0.04263    0.01598  -2.668  0.00765 **
## factor(energy_class)C                -0.04035    0.01598  -2.525  0.01156 *
## support_HDR                           0.29152    0.01649  17.677  < 2e-16 ***
## factor(refresh_rate)up to 60 hz      -0.63258    0.02389 -26.477  < 2e-16 ***
## factor(screensize)40-49 inch          0.23173    0.01919  12.078  < 2e-16 ***
## factor(screensize)50-59 inch          0.96596    0.02704  35.724  < 2e-16 ***
## factor(screensize)60 inch and above   0.76405    0.03087  24.753  < 2e-16 ***
## factor(screensize)up to 29 inch      -0.33522    0.03082 -10.876  < 2e-16 ***
## log(final_price)                     -2.13291    0.04788 -44.550  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7048 on 24422 degrees of freedom
## Multiple R-squared:  0.6832, Adjusted R-squared:  0.683
## F-statistic:  3098 on 17 and 24422 DF,  p-value: < 2.2e-16
```

b. All other factors being equal (ceteris paribus), Philips' sales are 3.19 (2 d.p.) units < LG's sales, Samsung's sales are 0.57 (2 d.p.) units > LG's sales, Sony's sales are 0.13 (2 d.p.) units < LG's sales.

c. Samsung has the highest brand equity perceived by customers, as its coefficient is the biggest in our regression model, meaning, it must have the highest utility factor for customers, making them derive the most satisfaction from Samsung purchase.

12.

a.

```
# design your regression below
q12a = lm(data = data_full, formula = log(sales+0.01) ~ factor(brand) + factor(technology) + factor(res
summary(q12a)
```

```
##
## Call:
## lm(formula = log(sales + 0.01) ~ factor(brand) + factor(technology) +
##     factor(resolution) + factor(energy_class) + support_HDR +
##     factor(refresh_rate) + factor(screensize) + log(final_price),
##     data = data_full)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3208 -0.3336  0.0390  0.4027  2.7265
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       15.56412    0.32917  47.283  < 2e-16 ***
## factor(brand)Philips              -3.18772    0.04942 -64.499  < 2e-16 ***
```

```
## factor(brand)Samsung                 0.57416    0.01794  32.007  < 2e-16 ***
## factor(brand)Sony                   -0.13490    0.01352  -9.978  < 2e-16 ***
## factor(technology)OLED               2.66545    0.03492  76.323  < 2e-16 ***
## factor(technology)QLED               2.65479    0.03872  68.569  < 2e-16 ***
## factor(resolution)4k                 1.31955    0.01749  75.445  < 2e-16 ***
## factor(resolution)720p              -1.12353    0.02895 -38.809  < 2e-16 ***
## factor(energy_class)A+              -0.01191    0.01077  -1.106  0.26856
## factor(energy_class)B               -0.04263    0.01598  -2.668  0.00765 **
## factor(energy_class)C               -0.04035    0.01598  -2.525  0.01156 *
## support_HDR                          0.29152    0.01649  17.677  < 2e-16 ***
## factor(refresh_rate)up to 60 hz     -0.63258    0.02389 -26.477  < 2e-16 ***
## factor(screensize)40-49 inch         0.23173    0.01919  12.078  < 2e-16 ***
## factor(screensize)50-59 inch         0.96596    0.02704  35.724  < 2e-16 ***
## factor(screensize)60 inch and above  0.76405    0.03087  24.753  < 2e-16 ***
## factor(screensize)up to 29 inch     -0.33522    0.03082 -10.876  < 2e-16 ***
## log(final_price)                    -2.13291    0.04788 -44.550  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7048 on 24422 degrees of freedom
## Multiple R-squared:  0.6832, Adjusted R-squared:  0.683
## F-statistic:  3098 on 17 and 24422 DF,  p-value: < 2.2e-16
```

b. I included the above-mentioned variables for the regression, as the combination resulted in the biggest adjusted R-squared of 0.683. Ceteris paribus, the sales of screen size:

- 0-29 inch: 0.33522 < sales of 30-39 inch;
- 40-49 inch: 0.23173 > sales of 30-39 inch;
- 50-59 inch: 0.96596 > sales of 30-39 inch;
- 60 inch+: 0.76405 > sales of 30-39 inch.

c. There is no monotonic relationship, as the coefficients do not increase to the same extent as the screensize. However, we see a decrease in coefficients from "50-59 inch" to "60 inch and above". Hence, the price decreases when the screensize becomes more than 60 inches, meaning, that the curve for the relationship may look like a downward parabola, peaking at 60 inches on the y-axis.

Endogeneity and Instrumental Variables (14pts)

13. Most often, as price levels increase, quantity sold reduces, due to the reverse relationship between the supply and demand. However, our first two regressions reveal the opposite. The primary reason for it is endogeneity, defined by a high correlation between the explanatory variable and the error term. Other factor variables (brand, technology, resolution, screensize) were omitted from the regressions. Since these variables also affect the outcome variable and are correlated with our explanatory variable "final price", we conclude that there is an omitted variable bias in the final results derived. Secondly, simultaneity may exist, i.e. the final price and sales can mutually affect each other in the same direction, triggering reverse causality. For example, customers may associate higher price with higher quality of TVs, making sales of the pricier brands to increase. Or vice versa, the increasing quantity of sales made brands to increase their price in order to exploit the customers' high demand, and meet the price-quantity equilibrium. Finally, there is always a possibility of incorrect records of sales, inconsistent units, and missed values, leading to the true mathematical model being skewed. However, for such pronounce counter-intuitive results, that would have to be a systematic issue, which is unlikely in the data frame we have.

14. *discount is not a valid instrument, because it does not satisfy the exogeneity requirement. Please refer to the mark scheme for valid instruments discussed in class.*

a. An instrumental variables is an observable variable that, in our case, should be correlated with the explanatory variable "final price" but uncorrelated with the explanatory variable "sales", and the error term. The 2 things considered, discount can be an instrumental variable because of its direct effect on the final price but merely any effect on sales, at the same time being uncorrelated with the error term. The reason for it is the endogenous nature of TV prices on sales - as stated earlier, the counter-intuitive logic of the model means that lower price does not necessarily mean higher quantity of sales, as usually anticipated in economic theory.

1 Q14a

b.

```
pacman::p_load(AER)
# regress X on Z
stage_1 <- lm(data = data_full,
formula = log(final_price) ~ discount)
# predict X hat
data_full <- data_full %>% mutate(predicted = predict(stage_1))
# second stage
stage_2 <- lm(data = data_full, formula = log(sales+0.01) ~ predicted)
# summary(second_stage)
summary(stage_2)
```

2.5 Q14b

```
##
## Call:
## lm(formula = log(sales + 0.01) ~ predicted, data = data_full)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3814 -0.6742  0.0459  0.8333  3.5703
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.9787     1.0456  10.500   <2e-16 ***
## predicted    -1.2711     0.1417  -8.969   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.25 on 24438 degrees of freedom
## Multiple R-squared:  0.003281,   Adjusted R-squared:  0.00324
## F-statistic: 80.45 on 1 and 24438 DF,  p-value: < 2.2e-16
```

You're welcome, Tom. Your consultancy fee quote is £250.

References: [1] Bizzaco, M., Cohen, S. and Lacoma, T. (2019). "OLED or LED? We Pick the Winner in the Battle of Competing TV Tech." [Online] Digital Trends. Available at: https://www.digitaltrends.com/home-theater/oled-vs-led/ [Accessed 1 Dec. 2021].

MSIN0094 Individual Coursework 2

ORIGINALITY REPORT

66%	20%	16%	65%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to University College London Student Paper	63%
2	easternfasr628.weebly.com Internet Source	1%
3	Submitted to University of York Student Paper	1%
4	koreascience.or.kr Internet Source	1%
5	www.tandfonline.com Internet Source	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		

MSIN0094 Individual Coursework 2

GRADEMARK REPORT

FINAL GRADE

85/100

GENERAL COMMENTS

Instructor

PAGE 1

QM 4

QM 2

QM 4

QM 4

PAGE 2

QM 4

QM 3

QM Q3a

Good attempt!

You will have to include the following to gain full points:

correct code to check the correlation coefficient (2pts)

can use cor.test or cor

The correlation coefficient is 0.403 (1pts)
There is a positive correlation between the RRP and Marketing expense (1pts)

QM **0**

Additional Comment you need to state it that there is a positive correlation between the RRP and Marketing expense

QM **3**

QM **Q3b**

Good attempt!

You will have to include the following to gain full points:

correct code to conduct correlation test (2pts)
Yes, the correlation is statistically significant (1pt)
Because the p-value is smaller than 0.01, so the correlation coefficient is significant at the 1% level.

(5% or 10% significant are inaccurate)(1pt)

QM **0**

Additional Comment the correct reason is because the p-value is smaller than 0.01, so the correlation coefficient is significant at the 1% level. 5% or 10% significant are inaccurate

QM **3**

QM **Q3c**

Good attempt!

You will have to include the following to gain full points:

A high quality product is likely to have a high RRP; because high quality product is like to be a flagship product by the company, so the company is willing to spend more on ads to promote the product. (2pts)

When setting the RRP, companies often have anticipated or made marketing expense budget for the next year, so the RRP has partially reflected the future marketing expense. (2pts)

QM **4**

QM **6**

QM **2**

PAGE 4

QM

Q6b

Good attempt!

You will have to include the following to gain full points:

correct code using t.test (1pts) correct t.test result (1pts)

check the t-value to be 14.3
Yes, the mean difference is statistically different (1pts)
Since the p-value is less than 1%, so the mean is different at the 1% level (1pts)

QM

3

QM

2

QM

6

QM

4

PAGE 5

QM

2

QM

Q9b

Good attempt!

You will have to include the following to gain full points:

Interpretation: If the price increases by 1 unit, then the sales increase by 0.00316 units. (1pts)
The coefficients of price is statistically significant at the 1% level. (1pts)
The R-squared is 0.15, which means 15% of the variation in sales can be explained by price (1pts)

QM

3

PAGE 6

QM

2

QM

1

QM

Q10b

Good attempt!

You will have to include the following to gain full points:

Interpretation: If the price increases by 1 unit, then the sales increase by 149%. (1pts)
The coefficients of price is statistically significant at the 1% level. (1pts)
The R-squared is 0.3259, which means 32.59% of the variation in log(sales) can be explained by log(price) (1pts): must be log(sales) and log(price)

QM

2

QM

2

QM

Q11a

Good attempt!

You will have to include the following to gain full points:

log(price) and log(sales)(1pts; must be log price and sales; because from the previous question we know log gives better R2)

technology (0.5pts; only Samsung has QLED,)

brand(0.5pts)

marketing_expense (1pts), other controls (1pt)

justify the specification(1pt)

QM

2

QM

2

QM

2

QM

Q12a

Good attempt!

You will have to include the following to gain full points:

factor(screensize) (1pts)

Regression design: price(0.5pts), marketing_expense(0.5pts), two controls(1pts)

students have to justify the specifications (1pts)

PAGE 8

QM **2**

QM **2**

QM **4**

QM **Q13**

Good attempt!

You will have to include the following to gain full points:

omitted variable (1pts) and explanation (1pts)

reverse causality (1pts) and explanation (1pts)

measurement error (1pts) and explanation (1pts)

PAGE 9

Text Comment. discount is not a valid instrument, because it does not satisfy the exogeneity requirement. Please refer to the mark scheme for valid instruments discussed in class.

QM **1**

QM **Q14a**

Good attempt!

You will have to include the following to gain full points:

cost shifters such as manufacturing costs; wholesale prices; BLP instruments (2pts)

you must justify why the instruments are good candidates:

relevance: instruments are correlated with price (1pts)

exogeneity: instruments do not directly affect the sales (1pts)

QM

2.5

QM

Q14b

Good attempt!

You will have to include the following to gain full points and need to discuss each points fully with out errors:

Step 1:
regress the log(price) on Z (1pts)
predict the log(price) using Z (1pts)
check the F-statistics of the first stage regression to make sure that the IV is not a weak IV (1pts)

Step 2:
replace the predicted log(price) back into the original regression and run the second stage regression (1pts)