Velo.com Revisited Case

Ali Abbas Ladha

2024-10-18

Q1

```
library(tidyverse)
library(gt)


velo <- read_csv('/Users/aliladha/Documents/Files/School Work/College/Graduate/IS 6489-001 Stats & Pred Analytics/velo.csv')

velo <-velo %>% mutate(checkout_system = factor(checkout_system, levels = c('old', 'new')))

glimpse(velo)
```

```
## Rows: 3,483
## Columns: 7
## $ customer_id     <dbl> 8968, 36687, 42232, 82931, 7010, 83252, 51631, 38122, …
## $ checkout_system <fct> old, new, new, old, old, new, new, old, old, old, old,…
## $ device          <chr> "computer", "mobile", "computer", "mobile", "computer"…
## $ country         <chr> "ESP", "GBR", "DEU", "MEX", "USA", "GBR", "FRA", "ESP"…
## $ gender          <chr> "F", "F", "F", "M", "M", "F", "F", "M", "M", "M", "M",…
## $ purchases       <dbl> 3, 4, 4, 4, 4, 4, 4, 4, 5, 2, 3, 4, 2, 3, 3, 3, 4, 4, …
## $ spent           <dbl> 2217.78, 4304.56, 396.59, 360.66, 2597.86, 1458.12, 17…
```

```
##Based on the model summary, answer the following questions:

#1. What is average spending for customers using 1) the new system and 2) the old system?
This information can be extracted directly from the linear model output.

model_velo <- lm(spent ~ checkout_system, data = velo)
model_velo
```

```
##
## Call:
## lm(formula = spent ~ checkout_system, data = velo)
##
## Coefficients:
```

```
##        (Intercept)  checkout_systemnew
##            2217.15               62.74
```

```
summary(model_velo)
```

```
##
## Call:
## lm(formula = spent ~ checkout_system, data = velo)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2256.0  -986.2  -156.5   791.8  6541.5
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2217.15      31.90  69.511   <2e-16 ***
## checkout_systemnew    62.74      44.03   1.425    0.154
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1298 on 3481 degrees of freedom
## Multiple R-squared:  0.000583,   Adjusted R-squared:  0.0002959
## F-statistic: 2.031 on 1 and 3481 DF,  p-value: 0.1542
```

```
# Average spending for the new checkout system

checkout_system_new <- 1
(avg_spent_new <- 2217.148 +  62.7413*(checkout_system_new))
```

```
## [1] 2279.889
```

```
# Average spending for the old checkout system

checkout_system_old <- 0
(avg_spent_old <- 2217.148 +  62.7413*(checkout_system_old))
```

```
## [1] 2217.148
```

```
#2. What is the difference in average customer spending between the new and old systems, a
nd is the difference statistically significant at the p < .05 level?

# Difference in average customer spending between the new & old systems
(avg_spent_difference <- avg_spent_new - avg_spent_old)
```

```
## [1] 62.7413
```

```
# Determining p level significance
summary(model_velo)$coefficients[2,4]
```

```
## [1] 0.1542364
```

The p value is approximetely 0.15 which is greater than 0.05. Therefore, the difference in average customer spending between the new and old systems is Not statistiscally significant. So we fail the reject the null hypothesis.

```
#3. Compare these results to those you obtained using a t-test in the last module (see the above output).

t.test(filter(velo, checkout_system == "new")$spent,
       filter(velo, checkout_system == "old")$spent)
```

```
##
##  Welch Two Sample t-test
##
## data:  filter(velo, checkout_system == "new")$spent and filter(velo, checkout_system == "old")$spent
## t = 1.4272, df = 3464.4, p-value = 0.1536
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -23.45215 148.93475
## sample estimates:
## mean of x mean of y
##  2279.890  2217.148
```

Using the t test, the p value is also approximetely 0.154 which leads to the same conclusion as the linear regression model in which the differences between the new and old checkout systems are not statistically significant. Therefore we fail to reject the null hypothesis since the p value of 0.154 is greater than 0.05.

Q2

```
# Fit a simple linear regression with spent as the target variable and checkout_system as the predictor, but include only customers who were using a mobile device. (Later you will learn how to fit this sort of regression using all the data, but for now subset the data to include only mobile users.) Answer these questions based on the model summary for this regression:


velo_m <- velo %>% filter(device == 'mobile') %>% group_by(checkout_system)

(model_velo_m <- lm(spent ~ checkout_system, data = velo_m))
```

```
##
## Call:
## lm(formula = spent ~ checkout_system, data = velo_m)
##
## Coefficients:
```

```
##        (Intercept)  checkout_systemnew
##             2174.9               148.1
```

```
summary(model_velo_m)
```

```
##
## Call:
## lm(formula = spent ~ checkout_system, data = velo_m)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2284.4  -976.2  -171.8   803.1  6498.4
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2174.92      46.21  47.066   <2e-16 ***
## checkout_systemnew   148.08      61.98   2.389    0.017 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1305 on 1795 degrees of freedom
## Multiple R-squared:  0.00317,    Adjusted R-squared:  0.002615
## F-statistic: 5.708 on 1 and 1795 DF,  p-value: 0.01699
```

```
# 1. What is the difference in mobile customer spending between the new and old systems?
```

```
# Average spending for the new checkout systems for mobile devices
checkout_system_new_m <- 1
(avg_spent_new_m <- 2174.92 + 148.08*(checkout_system_new_m))
```

```
## [1] 2323
```

```
# Average spending for the old checkout system for mobile devices
checkout_system_old_m <- 0
(avg_spent_old_m <- 2174.92 + 148.08*(0))
```

```
## [1] 2174.92
```

```
# Difference in the average customer spending between the new & old checkout systems for m
obile devices
(avg_spent_difference_m <- avg_spent_new_m - avg_spent_old_m)
```

```
## [1] 148.08
```

```
# 2. Is the difference statistically significant at the p < .05 level?

(summary(model_velo_m)$coefficients[2,4])
```

```
## [1] 0.01698621
```

The p value is approximetely 0.017 which is less than 0.05 which means that there is likely to be a population difference rather than a difference in sampling variation between the two checkout systems for mobile devices. Based on the low p value, we can reject the null hypothesis and state that there is a statistical difference between the new and old checkout systems for mobile devices.

```
# 3. Compare these results to those you obtained using a t-test in the last module (see th
e above output).
# Remember that the syntax for creating a simple linear regression model in R has the for
m: lm(y ~ x, data). This code will produce estimates of the model intercept and coefficien
ts which you can use to assemble the regression equation: y = intercept + coefficient * x.
To get more information about the model (including standard errors and p-values) use the g
eneric summary() function.

t.test(filter(velo, checkout_system == "new" & device == "mobile")$spent,
       filter(velo, checkout_system == "old" & device == "mobile")$spent)
```

```
##
##  Welch Two Sample t-test
##
## data:  filter(velo, checkout_system == "new" & device == "mobile")$spent and filter(vel
o, checkout_system == "old" & device == "mobile")$spent
## t = 2.399, df = 1733.1, p-value = 0.01655
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   27.01302 269.13848
## sample estimates:
## mean of x mean of y
##  2322.996  2174.920
```

Using the t test, the p value is also approximetely 0.017 which leads to the same conclusion as the linear regression model in which the differences between the new and old checkout systems is statistically significant. Therefore we reject the null hypothesis since the p value of 0.017 is less than 0.05.

Q3

```
#Using the summary of the regression model object from Q2, calculate a 95% confidence inte
rval (CI) for the checkout_system coefficient using 2 as the critical value. Interpret the
CI: What does it mean exactly?

(CI <- c(148.08 + 2 * 61.98, 148.08 - 2 * 61.98))
```

```
## [1] 272.04  24.12
```

The CI ranges between: 24.12 and 272.04. This means under repeated sampling/tests, we can expect the coefficient to vary within that range. This means we can expect additional revenue of that amount between that range for each customer on average using the new checkout system. Since the CI does not contain an interval of 0, the difference between the new and old checkout systems for mobile devices is statistically significant.

Q4

```
# Based on the model output in Q2, along with the 95% CI you calculated in Q3, develop bes
t and worst case revenue projections for the new checkout system. What range of increased
revenue might the company expect using the new checkout system (compared to the old syste
m) and, based on this, does the coefficient estimate for checkout_system have practical si
gnificance in your view? (Hint: multiply the lower and upper bounds of the CI you calculat
ed in Q3 by an arbitrary number of customers. That range, remember, is expressed in terms
of average dollars per customer.)
#
# A 95% CI for a coefficient represents the most likely range of values under repeated sam
pling. Working with a range is advantageous. Velo.com executives might well be interested
in the low end of the range in order to assess the financial risks associated with impleme
nting the new system. Additionally, scaling up the average spending per customer gives you
a way to convey the practical significance of using the new system.
#
# What do I mean by “scaling up”? For communication purposes it is often helpful to multip
ly averages, like the upper and lower bounds of the confidence interval, by some arbitrary
scaling number (in this case, for example, 1000 or 5000 customers or whatever you think ma
kes sense) so that a decision-maker can get a better sense of the overall impact to the bu
siness of a proposed change. Think back to the Conley Fisheries case where you calculated
that the maximum daily loss per boat was $10,000. By itself, that may seem like a sustaina
ble loss. But remember: there are 50 boats. So, really, to convey a complete sense of the
risk, the maximum per boat daily loss must be scaled up — multiplied by — the number of bo
ats in the fleet.
```

The sample sizes between the new checkout system and the old checkout system differs for mobile devices. The sample size of the new checkout system is 999 whereas the sample size of the old checkout system is 798. This may be due to customizers losing their order before completing the checkout system due to the glitch.

```
# Best case scenario using the new checkout system for mobile devices assuming 999 custome
rs
best_case_m <- CI[1] * 2000
best_case_m
```

```
## [1] 544080
```

Under the best case scenario with the new checkout system assuming we have 2000 customers, we can expect an additional revenue of $544,080

```
# Worst case scenario using the new checkout system for mobile devices assuming we have 99
9 customers
worst_case_m <- CI[2] * 2000
worst_case_m
```

```
## [1] 48240
```

Under the worst case scenario with the new checkout system assuming we have 2000 customers, we can expect an additional revenue of $48,240

Q5

```
# What course of action should Sarah recommend to the management at velo.com based on this additional analysis? Please incorporate your analytic results from above in fashioning an answer. In particular, cite the results from the new analysis, the simple linear regression, and discuss your interpretation of the range you developed in Question 4.
```

Overall Sarah should recommend to management at velo to use the new checkout system as it is statistically significant for mobile devices. The pvalue for mobile devices using linear regression and a t test between the checkout systems was 0.017. This pvalue is less than 0.05 which means that there is likely to be a population difference rather than a difference in sampling variation between the two checkout systems for mobile devices. Based on the low p value, we can reject the null hypothesis and state that there is a statistical difference between the checkout systems for mobile devices.

Furthermore, Sarah should advocate the significance of the Confidence interval. The confidence interval ranges between: 24.12 and 272.04. This means under repeated sampling/tests, we can expect the difference in revenue to vary within that range for each customer. But, rather than mentioning the effects of the increased revenue for 1 customer, Sarah should scale it up to a reasonable amount of customers to be more persuasive for her findings.

For example using a customer base of 2000 clients, Sarah should mention: that on a best case scenario, Velo can expect an additional revenue from mobile 2000 customers of $544,080. In the Worst case scenario, Velo can expect an additional revenue of $48,240. Furthermore she should remind Velo that customers under the old checkout system lose their order periodically due to a glitch. This can cause the customer to leave the transaction all together and not fulfill the order which represents a missed opportunity and lost revenue.

Ultimately, by presenting the scaled findings to senior management Sarah will be able to convince Velo to integrate the new checkout system entirely.