MS5318_Assignment4

HO Yin Shan (57487297)

2023-04-04

Question 1

```
convenience <-  read_csv("convenience.csv")
```

(a) The manager of the chain wants to compare the sales performance of the two service stations. Is it appropriate for him/her to conclude based on only a two-sample t-test on the sales? Would such a comparison be confounded by different levels of traffic (as measured by the volume of gas sold)?

It is not appropriate to make conclusion based on a two-sample t-test on the sales. And such a comparison will be confounded by different levels of traffic.

(b) Perform the two-sample t-test to compare the sales of the two service stations. Summarize this analysis, assuming that there are no confounding variables. (You may run a linear regression to perform the two-sample t-test.)

```
S1_data <- convenience %>%
  filter(Site == "Site 1")

S2_data <- convenience %>%
  filter(Site == "Site 2")
```

method 1

```
t.test(S1_data$Sales, S2_data$Sales)
```

```
##
##  Welch Two Sample t-test
##
## data:  S1_data$Sales and S2_data$Sales
## t = 28.717, df = 551.48, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  677.4665 776.9493
## sample estimates:
## mean of x mean of y
##  2218.127  1490.919
```

method 2

```
lm(Sales~Site, data = convenience) %>%
  summary()
```

```
##
## Call:
## lm(formula = Sales ~ Site, data = convenience)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1288.13  -197.17   -36.63   201.12  1260.87
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2218.13      17.93  123.73   <2e-16 ***
## SiteSite 2   -727.21      25.31  -28.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 301.6 on 566 degrees of freedom
## Multiple R-squared:  0.5933, Adjusted R-squared:  0.5926
## F-statistic: 825.6 on 1 and 566 DF,  p-value: < 2.2e-16
```

Based on the t-test and the regression coefficients, it is found that the station in Site 2 has 727.21 dollar sales of gases than Site 1.

(c) Compare the sales of the two stations while including Volume in the analysis. Summarize the comparison of sales based on this analysis. (Assume for the moment that the model meets the conditions for the multiple regression model.)

```
lm(Sales ~ ., data = convenience) %>%
  summary()
```

```
##
## Call:
## lm(formula = Sales ~ ., data = convenience)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -733.75 -164.79  -26.04  146.96 1191.96
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1171.34056   61.89565   18.92   <2e-16 ***
## Volume         0.31366    0.01803   17.39   <2e-16 ***
## SiteSite 2  -520.42454   23.64755  -22.01   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 243.6 on 565 degrees of freedom
## Multiple R-squared:  0.7351, Adjusted R-squared:  0.7342
## F-statistic:   784 on 2 and 565 DF,  p-value: < 2.2e-16
```

Based on the regression result, it is found that the sales of the gases in Site 2 is 520.42 dollar less than Site 1 even when keeping all other variable unchanged. This means that even considering the volume of sales, the sales are still different.

(d) Compare the results from parts (b) and (c). Do they agree? Explain why they agree or differ. You should take into account the precision of the estimates and your answer to part (a).

```
convenience %>%
  group_by(Site) %>%
  summarise(avg_price = mean(Sales/Volume))
```

```
## # A tibble: 2 x 2
##   Site   avg_price
##   <chr>      <dbl>
## 1 Site 1     0.693
## 2 Site 2     0.566
```

Based on the regression result, it is found that the sales of the two Sites are different even considering the volume of Sales. According to the calculation, the reason of the different is probably because of the price of the gases at the two sites are different.

Question 2

(a)

```
univeralbank <- read_csv("UniversalBank.csv")
#change the datatype of Education and Family from to category
univeralbank <- univeralbank %>%
  mutate(Education = as.factor(Education),
         Family = as.factor(Family))
head(univeralbank,3)
```

```
## # A tibble: 3 x 8
##     Age Experience Income Family CCAvg Education Mortgage PersonalLoan
##   <dbl>      <dbl>  <dbl> <fct>  <dbl> <fct>        <dbl>        <dbl>
## 1    25          1     49 4        1.6 1                0            0
## 2    45         19     34 3        1.5 1                0            0
## 3    39         15     11 1        1   1                0            0
```

```r
model <- glm(PersonalLoan~., family="binomial", data = univeralbank)
summary(model)
```

```
##
## Call:
## glm(formula = PersonalLoan ~ ., family = "binomial", data = univeralbank)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1138  -0.2083  -0.0749  -0.0232   4.2304
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.175e+01  1.726e+00  -6.807 9.99e-12 ***
## Age         -5.897e-02  6.454e-02  -0.914   0.3608
## Experience   6.827e-02  6.428e-02   1.062   0.2883
## Income       6.244e-02  2.877e-03  21.703  < 2e-16 ***
## Family2     -3.267e-01  2.205e-01  -1.481   0.1385
## Family3      1.984e+00  2.355e-01   8.425  < 2e-16 ***
## Family4      1.497e+00  2.229e-01   6.713 1.90e-11 ***
## CCAvg        2.100e-01  4.343e-02   4.835 1.33e-06 ***
## Education2   3.951e+00  2.623e-01  15.060  < 2e-16 ***
## Education3   4.078e+00  2.610e-01  15.622  < 2e-16 ***
## Mortgage     1.114e-03  5.747e-04   1.938   0.0526 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3162.0  on 4999  degrees of freedom
## Residual deviance: 1270.9  on 4989  degrees of freedom
## AIC: 1292.9
##
## Number of Fisher Scoring iterations: 8
```

(b)

```r
new_data <- data.frame(Age=33, Experience=8, Income=85, Family=as.factor(3), Education=as.factor(2), Mo

predict(model, new_data, type = "response")
```

```
##         1
## 0.1577678
```

The probability of loan acceptance is 15.78%.