Predictive Analysis

Problem Set 3 Question 3

Atrijo Roy

2026-02-26

Problem to demonstrate the role of qualitative (ordinal) predictors in addition to quantitative predictors in multiple linear regression

Consider "diamonds" data set in R. It is in the ggplot2 package. Make a list of all the ordinal categorical variables. Identify the response.

```
rm(list=ls())
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.5.1

attach(diamonds)
head(diamonds) #Overview of the data

## # A tibble: 6 × 10
##   carat cut       color clarity depth table price     x     y     z
##   <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
## 1  0.23 Ideal     E     SI2      61.5    55   326  3.95  3.98  2.43
## 2  0.21 Premium   E     SI1      59.8    61   326  3.89  3.84  2.31
## 3  0.23 Good      E     VS1      56.9    65   327  4.05  4.07  2.31
## 4  0.29 Premium   I     VS2      62.4    58   334  4.2   4.23  2.63
## 5  0.31 Good      J     SI2      63.3    58   335  4.34  4.35  2.75
## 6  0.24 Very Good J     VVS2     62.8    57   336  3.94  3.96  2.48
```

- **Data in Hand:** A dataset containing the price and nine other attributes of n = 53940 diamonds.
- **Ordinal Categorical Variables:**

1. *cut*: quality of the cut (Fair, Good, Very Good, Premium, Ideal)
2. *color*: diamond colour, from D (best) to J (worst)
3. *clarity*: a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))

- **Response:** price: price in US dollars ($326–$18,823)

(a) Run a linear regression of the response on the quality of cut. Write the fitted regression model.

Model:

We take Ideal Cut as the baseline and hence the model is given by,

$$Y_i = \beta_0 + \beta_1 cut_{Fair,i} + \beta_2 cut_{Good,i} + \beta_3 cut_{VeryGood,i} + \beta_4 cut_{Premium,i} + \varepsilon_i, \quad i = 1,2,\dots,n$$

where

$$\varepsilon_i \sim N(0,\sigma^2), \quad i = 1,2,\dots,n$$

β_0,β_1,…,β_4 are unknown parameters which are to be estimated by the method of least squares.

And the dummy variables are defined as:

$$cut_{Good,i} = \begin{cases} 1 & \text{if ith cut = Good} \\ 0 & \text{otherwise} \end{cases}$$

$$cut_{VeryGood,i} = \begin{cases} 1 & \text{if ith cut = Very Good} \\ 0 & \text{otherwise} \end{cases}$$

$$cut_{Premium,i} = \begin{cases} 1 & \text{if ith cut = Premium} \\ 0 & \text{otherwise} \end{cases}$$

$$cut_{Fair,i} = \begin{cases} 1 & \text{if ith cut = Ideal} \\ 0 & \text{otherwise} \end{cases}$$

$$i = 1,2,\dots,n$$

Now we fit the model:

```
diamonds$cut=as.character(diamonds$cut)
fit1=lm(price~relevel(factor(cut),ref="Ideal"),data=diamonds)
summary(fit1)
```

```
##
## Call:
## lm(formula = price ~ relevel(factor(cut), ref = "Ideal"), data = diamonds)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -4258  -2741  -1494   1360  15348
##
## Coefficients:
##                                            Estimate Std. Error t value
## (Intercept)                                 3457.54      27.00 128.051
## relevel(factor(cut), ref = "Ideal")Fair      901.22     102.41   8.800
## relevel(factor(cut), ref = "Ideal")Good      471.32      62.70   7.517
## relevel(factor(cut), ref = "Ideal")Premium  1126.72      43.22  26.067
## relevel(factor(cut), ref = "Ideal")Very Good  524.22      45.05  11.636
##                                            Pr(>|t|)
```

```
## (Intercept)                                    < 2e-16 ***
## relevel(factor(cut), ref = "Ideal")Fair        < 2e-16 ***
## relevel(factor(cut), ref = "Ideal")Good        5.7e-14 ***
## relevel(factor(cut), ref = "Ideal")Premium     < 2e-16 ***
## relevel(factor(cut), ref = "Ideal")Very Good   < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3964 on 53935 degrees of freedom
## Multiple R-squared:  0.01286,    Adjusted R-squared:  0.01279
## F-statistic: 175.7 on 4 and 53935 DF,  p-value: < 2.2e-16
```

Fitted Model:

Since Ideal cut is the baseline, the fitted model is given by,

$$\widehat{Price}_i = 3457.54 + 901.22cut_{Fair,i} + 471.32cut_{Good,i} + 1126.72cut_{Premium,i} + 524.22cut_{VeryGood,i} \quad i = 1,2,\dots,n$$

(b) Test whether the expected price of diamond with premium cut is significantly different from that of the ideal cut.

Since Ideal is taken as the baseline category, the coefficient of Premium represents

$$\beta_{Premium} = \mu_{Premium} - \mu_{Ideal}.$$

We test the hypotheses:

$$H_0: \beta_{Premium} = 0$$

$$H_1: \beta_{Premium} \neq 0$$

From the fitted regression output:

- $\hat{\beta}_{Premium} = 1126.72$
- $SE(\hat{\beta}_{Premium}) = 43.22$

```
#Test statistic
t=1126.72/43.22
t

## [1] 26.06941
```

The test statistic is given by

$$t_{obs} = \frac{\hat{\beta}_{Premium}}{SE(\hat{\beta}_{Premium})} = \frac{1126.72}{43.22} = 26.06941.$$

Under H_0, The degrees of freedom are

$$df = n - 5 = 53935$$

since there are 5 parameters in the model (intercept + 4 dummy variables).

We reject H_0 at 5% level of significance iff $|t|_{obs} > t_{0.025,53935}$.

```
qt(0.975,53935)
```

```
## [1] 1.960008
```

Therefore, we reject H_0 at 5% level of significance.

Conclusion:
In light of the given data, there is strong statistical evidence that the expected price of diamonds with Premium cut is significantly different from that of Ideal cut, on an average.

(c) What is the expected price of a diamond of ideal cut?

For an Ideal cut diamond, all dummy variables are 0. Therefore,

$$\mathbb{E}(Price \mid \text{Ideal}) = \beta_0.$$

Using the estimated intercept,

$$\widehat{\mathbb{E}}(Price \mid \text{Ideal}) = 3457.54.$$

Hence, the expected price of a diamond with Ideal cut is 3457.54 US dollars.

(d) Modify the regression model in (a) by incorporating the predictor "table". Write the fitted regression model

Model:

$$Y_i = \beta_0 + \beta_1 cut_{Fair,i} + \beta_2 cut_{Good,i} + \beta_3 cut_{Premium,i} + \beta_4 cut_{VeryGood,i} + \beta_5 Table_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2). \quad i = 1,2,\dots,n$$

```
fit2=lm(price~relevel(factor(as.character(cut)),ref="Ideal")+table,data=diamonds)
summary(fit2)
```

```
##
## Call:
## lm(formula = price ~ relevel(factor(as.character(cut)), ref = "Ideal") +
##     table, data = diamonds)
##
## Residuals:
##    Min     1Q Median     3Q    Max
```

```
##  -5630  -2694  -1458   1346  15690
##
## Coefficients:
##                                                              Estimate Std.
Error
## (Intercept)                                                 -6563.672
517.450
## relevel(factor(as.character(cut)), ref = "Ideal")Fair        345.611
106.002
## relevel(factor(as.character(cut)), ref = "Ideal")Good        -19.957
67.426
## relevel(factor(as.character(cut)), ref = "Ideal")Premium     626.220
50.215
## relevel(factor(as.character(cut)), ref = "Ideal")Very Good   165.206
48.562
## table                                                        179.105
9.236
##                                                             t value
Pr(>|t|)
## (Intercept)                                                 -12.685  < 2e-
16 ***
## relevel(factor(as.character(cut)), ref = "Ideal")Fair        3.260
0.001113 **
## relevel(factor(as.character(cut)), ref = "Ideal")Good       -0.296
0.767246
## relevel(factor(as.character(cut)), ref = "Ideal")Premium    12.471  < 2e-
16 ***
## relevel(factor(as.character(cut)), ref = "Ideal")Very Good   3.402
0.000669 ***
## table                                                       19.393  < 2e-
16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3950 on 53934 degrees of freedom
## Multiple R-squared:  0.0197, Adjusted R-squared:  0.01961
## F-statistic: 216.7 on 5 and 53934 DF,  p-value: < 2.2e-16
```

Fitted Model:

$$\widehat{Price}_i = -6563.672 + 345.611 cut_{Fair,i} - 19.957 cut_{Good,i} + 626.220 cut_{Premium,i} + 165.206 cut_{VeryGood,i} + 179.105\, Table_i. \quad i = 1,2,\dots,n$$

(e) Test for the significance of "table" in predicting the price of diamond

To test,

$$H_0: \beta_{table} = 0$$

$$H_1: \beta_{table} \neq 0$$

From the fitted regression output:

- $\hat{\beta}_{table} = 179.105$
- $SE(\hat{\beta}_{table}) = 9.236$

```
#Test statistic
t2=179.105/9.236
t2

## [1] 19.39205
```

The test statistic is

$$t = \frac{\hat{\beta}_{table}}{SE(\hat{\beta}_{table})} = \frac{179.105}{9.236} \approx 19.39.$$

Under H_0, degrees of freedom:

$$df = n - 6 = 53934,$$

We reject H_0 at 5% level of significance iff $|t|_{obs} > t_{0.025,53934}$.

```
qt(0.975,53934)

## [1] 1.960008
```

Therefore, we reject H_0 at 5% level of significance.

Conclusion:
In light of the given data, there is strong statistical evidence that table is a significant predictor of diamond price.

(f) Find the average estimated price of a diamond with an average table value and which is of fair cut

For a Fair cut diamond:

$$cut_{Fair} = 1, \quad cut_{Good} = 0, \quad cut_{Premium} = 0, \quad cut_{VeryGood} = 0.$$

```
#Average table value
tab.avg=mean(diamonds$table)
tab.avg
```

```
## [1] 57.45718
```

And the average table value is 57.45718.

Thus,

$$\widehat{Price} = -6563.672 + 345.611 + 179.105 \times 57.45718.$$

```
#Average estimated price

price.req=-6563.672+345.611+179.105*57.45718
price.req

## [1] 4072.807
```

Hence,

$$\widehat{Price} = 4072.807.$$

Therefore, the average estimated price of a Fair cut diamond with average table value is approximately 4072.807 US Dollars.