Machine Learning and Text Analysis

Problem Set-1

Bikash Panda and Mithilesh Shah

2025-01-22

(1) Sequential Estimation

Derive the expression at the top of slide 18. Start with the

$$y = X_1\hat{\boldsymbol{\beta}}_1 + X_2\hat{(}\boldsymbol{\beta}_2) + \boldsymbol{\varepsilon} \tag{1}$$

and then pre-multiply it by M2, the residual-maker matrix with respect to X2. Hint: you will need to use your understanding of what happens when you project different variables onto X2 (via the M2 / P2 matrices). Also note that M2 * M2 = M2.

Step-by-Step Derivation

1. **Start with the Model**

$$y = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

2. **Multiply Both Sides by M_2**

   Recall that

$$M_2 = I - X_2(X_2^T X_2)^1 X_2^T$$

We get the following:

$$M_2 y = M_2 X_1\boldsymbol{\beta}_1 + M_2 X_2\boldsymbol{\beta}_2 + M_2\boldsymbol{\varepsilon}$$

Since

$$M_2 X_2 = 0$$

We get

$$M_2 y = M_2 X_1\boldsymbol{\beta}_1 + M_2\boldsymbol{\varepsilon}$$

3. **Multiply Both Sides by X_1^T**

$$X_1^T(M_2y) = X_1^T(M_2X_1\boldsymbol{\beta}_1) + X_1^T(M_2\boldsymbol{\varepsilon})$$

Since

$$M_2\boldsymbol{\varepsilon} = 0$$

We get

$$X_1^T M_2 y = X_1^T M_2 X_1 \boldsymbol{\beta}_1$$

Now Solving for $\boldsymbol{\beta}_1$, we get

$$\boldsymbol{\beta}_1 = (X1_T M_2 X_1)^-1 X_1^T M_2 y$$

Thus we get an expression for $\boldsymbol{\beta}_1$

To get the equation on Slide 18, we could left multiply M_2 with M_2^T, since $M_2^T M_2 = M_2$

(2) Demand Estimation

Estimation 'demand_data.csv' and load it into R. These data include sales and pricing information at aset of 100 ice-cream vendors over a 52 week period. All ice-cream flavors at a given store in a given week are always priced the same, so there is only one price value per vendor-week. However, different vendors charge different prices and most vendors vary their prices throughout the year.

```
#Load demand data
library(tidyverse)
demand <- read.csv("data/demand_data.csv")
```

##Single vendor regressions

(a) For vendor 1, run a regression of sales on price as well as a regression of sales on price and a summer dummy. Make sure to select only the 52 weeks of data for vendor 1 when running the regression.Use the omitted variable bias formula to explain why the price coefficient changes when summer dummy is also included in the regression.

```
#
vendor1 <- demand %>%
```

```r
          filter(vendor_id == 1)
vendor1_reg <- lm(sales~ price, data = vendor1)
vendor1_reg_summer <-lm(sales~ price+summer_dummy, data = vendor1)
summary(vendor1_reg)
```

```
Call:
lm(formula = sales ~ price, data = vendor1)

Residuals:
    Min      1Q  Median      3Q     Max 
-616.97 -147.17   15.82  152.59  715.17 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8983.82     145.44   61.77   <2e-16 ***
price         -31.23      54.78   -0.57    0.571    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 252.2 on 50 degrees of freedom
Multiple R-squared:  0.006458,	Adjusted R-squared:  -0.01341 
F-statistic: 0.325 on 1 and 50 DF,  p-value: 0.5712
```

```r
summary(vendor1_reg_summer)
```

```
Call:
lm(formula = sales ~ price + summer_dummy, data = vendor1)

Residuals:
    Min      1Q  Median      3Q     Max 
-535.80 -146.73  -10.55  165.51  492.81 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   9177.55     128.43  71.458  < 2e-16 ***
price         -141.19      51.41  -2.746   0.0084 ** 
summer_dummy   358.50      75.79   4.730 1.94e-05 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 211.1 on 49 degrees of freedom
Multiple R-squared:  0.3179,	Adjusted R-squared:  0.2901 
F-statistic: 11.42 on 2 and 49 DF,  p-value: 8.493e-05
```

(b) Repeat the two regressions from part (a), but now use data only for vendor 2. In the case of the multi-variate regression with the summer-dummy control, you should find that price or summer_dummy are reported with a coefficient value of NA. This means that R dropped the variable from the regression. Why does this happen?

```r
vendor2 <- demand %>%
           filter(vendor_id == 2)
vendor2_reg <- lm(sales~ price, data = vendor2)
vendor2_reg_summer <-lm(sales~ price+summer_dummy, data = vendor2)
summary(vendor2_reg)
```

```
Call:
lm(formula = sales ~ price, data = vendor2)

Residuals:
    Min      1Q  Median      3Q     Max 
-601.63 -126.92   13.15   99.71  613.92 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8411.17     219.55  38.312  < 2e-16 ***
price         218.60      78.86   2.772  0.00781 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 246.2 on 50 degrees of freedom
Multiple R-squared:  0.1332,	Adjusted R-squared:  0.1159 
F-statistic: 7.684 on 1 and 50 DF,  p-value: 0.007807
```

```r
summary(vendor2_reg_summer)
```

```
Call:
```

```
lm(formula = sales ~ price + summer_dummy, data = vendor2)

Residuals:
    Min      1Q  Median      3Q     Max 
-601.63 -126.92   13.15   99.71  613.92 

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   8411.17     219.55  38.312  < 2e-16 ***
price          218.60      78.86   2.772  0.00781 ** 
summer_dummy       NA         NA      NA       NA    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 246.2 on 50 degrees of freedom
Multiple R-squared:  0.1332,	Adjusted R-squared:  0.1159 
F-statistic: 7.684 on 1 and 50 DF,  p-value: 0.007807
```

Since price only takes two values—one during the summer and one outside of summer—the summer dummy variable is perfectly correlated with price. Whenever the summer dummy is 1, the price is always the same, and whenever the summer dummy is 0, the price is the other value. This situation creates **perfect collinearity**, which means the regression cannot distinguish the separate impacts of the summer dummy and price on sales.

Because a regression model cannot estimate unique coefficients when two variables are perfectly collinear, R automatically drops one of them. In this case, it drops the summer dummy to ensure the model's parameters can still be estimated.

(c) Repeat the two regressions from part (a), but now use data only for vendor 3. Vendor 3 is different from 1 and 2 in that she did not systematically charge higher or lower prices in summer. Is it still necessary to include the summer_dummy variable in order to avoid omitted variable bias? Are there other benefits from including the summer dummy in the regression?

```r
vendor3 <- demand %>%
           filter(vendor_id == 3)
vendor3_reg <- lm(sales~ price, data = vendor3)
vendor3_reg_summer <-lm(sales~ price+summer_dummy, data = vendor3)
summary(vendor3_reg)
```

```
Call:
lm(formula = sales ~ price, data = vendor3)
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-676.63 -155.07   31.29  166.30  447.73 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4442.31     151.41  29.340  < 2e-16 ***
price        -172.16      64.28  -2.678  0.00999 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 258.5 on 50 degrees of freedom
Multiple R-squared:  0.1255,	Adjusted R-squared:  0.108 
F-statistic: 7.173 on 1 and 50 DF,  p-value: 0.009988
```

```r
summary(vendor3_reg_summer)
```

```
Call:
lm(formula = sales ~ price + summer_dummy, data = vendor3)

Residuals:
    Min      1Q  Median      3Q     Max 
-559.54 -101.95    5.03  128.83  405.04 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   4441.46     121.69  36.497  < 2e-16 ***
price         -210.90      52.17  -4.042 0.000187 ***
summer_dummy   358.08      67.20   5.329 2.48e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 207.8 on 49 degrees of freedom
Multiple R-squared:  0.4463,	Adjusted R-squared:  0.4237 
F-statistic: 19.75 on 2 and 49 DF,  p-value: 5.126e-07
```

Even if price does not systematically change with the season, sales might—due to changes in demand or other seasonal factors (e.g., more ice cream is sold in summer regardless of price). In that situation, including a summer dummy can still help the model by capturing the systematic seasonal fluctuation in sales. Doing so reduces the residual variance and thus allows us to estimate the price coefficient more precisely.

(3) Hospital admissions & quality of service

Download ‘health_data.csv‘ and load it into R. These are data from hospital admissions for coronary artery bypass graft (CABG) in the UK. Among other things, you observe whether the patient passed away after the surgery (coded up as ‘patient_died_dummy‘), which hospital the patient visited (‘hospital_id‘), and a series of patient characteristics such as gender and age.

Dummy variables interpretation

Start by regressing the patient-died dummy variable on a set of hospital dummies (Note: use the ‘lm‘ command and use the ‘factor(var_name)‘ syntax when including dummies).

(a) Based on the regression output, interpret the coefficients on the constant term and the dummy for hospital D.

(b) What is the difference between the mortality rates at hospitals D and E (use the regression output to derive this)?

Long and short regressions

Continue to use the hospital data in this question, but only use data for patients that visited either hospital A or B. Regress mortality on an intercept and a dummy for whether the patient visited hospital B. Also run the same regression, but include age (linearly) and gender as control variables

(c) Explain the difference between the coefficients in the long and short regressions based on the omitted variable bias formula

(d) Assume that you want to capture “true” quality differences between hospitals in terms of mortality rates. Does the short regression over- and under-estimate quality differences between hospitals? Explain your reasoning.

(e) Do you think the long regressions allows you to estimate “true” quality differences. Which additional control variables would you ideally like to include?