# PMDA: Homework 2 Template (Week 4)

## 1. Hospital admission & 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 died 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.

```
df = read.csv("health_data.csv", header = TRUE)
```

QUESTION 1: Start by regressing the patient-died dummy variable on a set of hospital dummies (Note: use the `factor(var_name)` syntax when including dummies in the 'lm' regression command).

```
reg1 = lm(patient_died_dummy~factor(hospital_id), data = df)
summary(reg1)
```

```
##
## Call:
## lm(formula = patient_died_dummy ~ factor(hospital_id), data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28526 -0.12883 -0.10084 -0.09702  0.95613
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            0.0970174  0.0059321  16.355  < 2e-16 ***
## factor(hospital_id)B   0.0071961  0.0080838   0.890 0.373371
## factor(hospital_id)C -0.0482978  0.0101118  -4.776 1.80e-06 ***
## factor(hospital_id)D   0.1882473  0.0081184  23.188  < 2e-16 ***
## factor(hospital_id)E -0.0531432  0.0111385  -4.771 1.84e-06 ***
## factor(hospital_id)F   0.0002589  0.0085538   0.030 0.975855
## factor(hospital_id)G   0.0441247  0.0083681   5.273 1.35e-07 ***
## factor(hospital_id)H   0.0038230  0.0092364   0.414 0.678951
## factor(hospital_id)I   0.0318115  0.0091406   3.480 0.000502 ***
## factor(hospital_id)J   0.0111743  0.0108701   1.028 0.303967
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3277 on 24470 degrees of freedom
## Multiple R-squared:  0.04203,    Adjusted R-squared:  0.04168
## F-statistic: 119.3 on 9 and 24470 DF,  p-value: < 2.2e-16
```

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

ANSWER: 9.7% of death rate in hospital A. 18.8 percentage points higher probability of death after the surgery in hospital D (i.e. death rate is 28.5% at hospital D)

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

ANSWER: 18.8% - (-5.3%) = 24.1% 24.1 percentage points higher probability of death after the surgery in hospital D compared to hospital E.

## Causal interpretation (or lack thereof)

QUESTION 2: 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.

```
ab = df[df$hospital_id == "A" | df$hospital_id == "B",]
reg2 = lm(patient_died_dummy~factor(hospital_id), data = ab)
summary(reg2)
```

```
##
## Call:
## lm(formula = patient_died_dummy ~ factor(hospital_id), data = ab)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.10421 -0.10421 -0.09702 -0.09702  0.90298
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.097017   0.005453  17.791   <2e-16 ***
## factor(hospital_id)B 0.007196   0.007431   0.968    0.333
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3012 on 6609 degrees of freedom
## Multiple R-squared:  0.0001419,  Adjusted R-squared:  -9.42e-06
## F-statistic: 0.9377 on 1 and 6609 DF,  p-value: 0.3329
```

(a) Explain why the difference in mortality rate implied by this regression cannot be interpreted as the causal effect of visiting a different hospital (i.e., the change in risk of dying when moving a patient from hospital A to B cannot be inferred from this regression).

ANSWER: Because whether to visit hospital A or B isn't a random assignment. Additionally, other factors that might contribute to the death rate are not accounted. Therefore, no casual effect can be interpreted.

(b) Do you think difference in mortality between hospitals are over- or under-estimated? Think about what type of patients go to which type of hospital.

ANSWER: I think it's underestimated since the regression doesn't consider other factors that affect mortality rate. The omitted variables could be positively or negatively correlated with hospital B. For example, under the circumstance that there are many male or young patients are present (i.e. female dummy have a negative coefficient) at B, the coefficient of hospital_id B would be underestimated. Other independent variables would also affect the mortality and it depends by case.

(c) What are potential control variables that you might want to include in the regression, in order to obtain a causal estimate (or at least get closer to a causal estimate)? Run such a regression with suitable controls and interpret the change in the coefficient on the hospital B dummy. Explain why you included the specific set of variables.

```
reg3 = lm(patient_died_dummy ~ factor(hospital_id) +startage +female_dummy, data = ab)
summary(reg3)
```

```
##
## Call:
## lm(formula = patient_died_dummy ~ factor(hospital_id) + startage +
##     female_dummy, data = ab)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28318 -0.07401 -0.06172 -0.05223  0.95533
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            0.1165413  0.0268842   4.335 1.48e-05 ***
## factor(hospital_id)B   0.0113760  0.0072061   1.579    0.114
## startage              -0.0009457  0.0004029  -2.347    0.019 *
## female_dummy           0.1836355  0.0087384  21.015  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2917 on 6607 degrees of freedom
## Multiple R-squared:  0.0628, Adjusted R-squared:  0.06238
## F-statistic: 147.6 on 3 and 6607 DF,  p-value: < 2.2e-16
```

ANSWER: First of all, different hospitals have different mortality rate so I want to include hospital_id. The age of the patients and how long they have been staying in the hospital(how serious their conditions are) can also be a good indicator of mortality rate. Whether the patient is a female can also affect death rate. I included all of the independent variables in the dataset and deleted "admin_year" as it's not statistically significant in the full model.

0.0113760 - 0.007196 = 0.00418 The mortality rate for hospital B over A increases by 0.418 percentage points when accounting for the additional control variables.

## 2. Demand estimation

The dataset demand_data.csv contains data on sales and prices at a set of ice-cream vendors measured over 52 weeks. All ice-cream at a given store is always priced the same, so there is only one price variable. However, different vendors charge different prices and most vendors vary their prices throughout the year.

QUESTION 1: Load `demand_data.csv` into R. For vendor 1, run a regression of sales on price and also a regression of sales on price and a summer dummy (make sure your regression selects only the 52 weeks of data for vendor 1). Use the omitted variable bias formula to explain why the price coefficient changes when the summer dummy is also included in the regression.

```
df1 = read.csv("demand_data.csv")
vendor1 = df1[df1$vendor_id==1,]
```

```
reg4 = lm(sales ~ price, data=vendor1)
summary(reg4)
```

```
##
## 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
```

```
reg5 = lm(sales ~ price + summer_dummy, data=vendor1)
summary(reg5)
```

```
##
## 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
```

ANSWER: Omitted-variable bias occurs when the model leaves out relevant variable(s). The coefficient changed because summer_dummy is also relevant in terms of determining the dependent variable. The effect of summer_dummy was included in price in the first model and after the positive relationship was taken out separately, the coefficient of price decreased.

QUESTION 2: Repeat the two regressions that you just ran in question 1, but now use data only for vendor 2. In the case of the regression with the summer dummy, you should find that price or the summer dummy are reported with a coefficient of NA. This means that R dropped the variable from the regression. Why

does this happen? (Hint: look at the correlation between price and summer dummy for vendor 2 using the "cor" syntax, where, for example, the price for vendor 2 is obtained as price[frame_name$vendor_id==2]).

```
vendor2 = df1[df1$vendor_id==2,]
reg6 = lm(sales ~ price, data=vendor2)
summary(reg6)
```

```
##
## 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
```

```
reg7 = lm(sales ~ price + summer_dummy, data=vendor2)
summary(reg7)
```

```
##
## 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
```

```
print(paste("Correlation between price and summer dummer is:" ,cor(vendor2$price, vendor2$summer_dummy)))
```

```
## [1] "Correlation between price and summer dummer is: 1"
```

ANSWER: Since price and summer_dummy have a perfect positive linear relationship, R Studio automatically dropped summer_dummy for us since accounting this extra variable doesn't affect our model.

QUESTION 3: Suppose that one of the vendors did not systematically charge higher or lower prices in summer. If you were to repeat the analysis you just did for vendors 1 and 2, what would you expect to happen to the price coefficient estimate and its precision in the two regressions with and without the summer dummy?

ANSWER: I would expect including summer dummy would produce a similar coefficient and a similar precision compared to the model without summer dummy. Since price and summer dummy are not correlated, taking summer dummy out or not won't really affect the coefficient and precision of price.