Homework 2

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.

```r
library(readr)
health_data = read.csv('health_data.csv')
```

QUESTION 1: Start by regressing the patient-died dummy variable on a set of hospital dummies.

```r
summary(lm(patient_died_dummy ~ factor(hospital_id), data = health_data))
```

```
##
## Call:
## lm(formula = patient_died_dummy ~ factor(hospital_id), data = health_data)
##
## 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: The constant term pertains to hospital A since that is the dummy variable whose coefficient was excluded from the regression. It means that the probability of a patient dying (i.e. the mortality rate) at hospital A is 0.0970174. The coefficient on hospital D means that, on average, patients there have a mortality rate that is 0.1882473 greater than those at hospital A.

(b) What is the difference between the mortality rates at hospitals D and E?

ANSWER: To compute the difference in mortality rates, we subtract the coefficients: 0.1882473 - (-0.0531432) = 0.2413905. Thus, the mortality rate at hospital D is 0.2413905 higher than the mortality rate at 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.

```
# select only patients from hospitals A and B
health_data_AB =
  health_data[health_data$hospital_id == 'A' | health_data$hospital_id == 'B',]
summary(lm(patient_died_dummy ~ factor(hospital_id), data = health_data_AB))
```

```
##
## Call:
## lm(formula = patient_died_dummy ~ factor(hospital_id), data = health_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: This difference in mortality rate cannot be interpreted as the change in risk of dying when moving a patient from hospital A to B because the regression does not include other variables that may be influencing the coefficient. That is, there are no controls included in the regression, so there may be omitted variable bias since factors like gender and age may be influencing the mortality rates.

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

ANSWER: One factor that would significantly affect the mortality rates at each hospital is age. Thus, we examine the average age at each hospital:

```
# mean age of patients at hospital A
print(mean(health_data_AB[health_data_AB$hospital_id == 'A',]$startage))
```

```
## [1] 65.70501
```

```
# mean age of patients at hospital B
print(mean(health_data_AB[health_data_AB$hospital_id == 'B',]$startage))
```

```
## [1] 64.8823
```

Since the average age at hospital A is higher than the average age of B, this difference in mortality rate may be an overestimate.

It is also the case that males have a lower life expectancy than females, so we examine the proportion of males and females at each hospital:

```
# proportion of females at hospital A
print(mean(health_data_AB[health_data_AB$hospital_id == 'A',]$female_dummy))
```

```
## [1] 0.2320551
```

```
# proportion of females at hospital B
print(mean(health_data_AB[health_data_AB$hospital_id == 'B',]$female_dummy))
```

```
## [1] 0.2050562
```

We can see that there is a higher proportion of females at hospital A, which suggests that this difference in mortality rate is an underestimate.

Overall, since the difference in average ages is very small (less than a year), we place greater emphasis on the differences in gender distributions. Because of this, we hypothesize that the coefficient in the above regression is an underestimate and that including gender and age controls in the regression will lead to a higher coefficient.

(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.

By including age and gender, we can get closer to a causal estimate of the change in risk of dying when moving a patient from hospital A to B:

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

```
##
## Call:
## lm(formula = patient_died_dummy ~ factor(hospital_id) + startage +
##     female_dummy, data = health_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: We can see that the coefficient on hospital B increased relative to the prior regression (difference = final - initial = 0.011376 - 0.007196 = 0.004180) This means that the change in mortality rate that results from moving patients from hospital A to hospital B increases by 0.004180 when gender and age are held constant.

We chose to include age and gender in the regression since these are variables that can greatly affect mortality rate and thus it is important to control for them when evaluating the (potentially) causal effect of moving patients to hospital B on mortality rate.

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.

```r
library(readr)
ice_cream = read.csv('demand_data.csv')
# data frame for vendor 1 only
vendor_1 = ice_cream[ice_cream$vendor_id == 1,]
# univariate regression of sales on price
summary(lm(sales ~ price, data = vendor_1))
```

```
## 
## Call:
## lm(formula = sales ~ price, data = vendor_1)
## 
## 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
# multivariate regression of sales on price and summer_dummy
summary(lm(sales ~ price + summer_dummy, data = vendor_1))
```

```
## 
## Call:
## lm(formula = sales ~ price + summer_dummy, data = vendor_1)
## 
## 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: We have the following formula for omitted variable bias:

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 \frac{Cov(X_1, X_2)}{Var(X_1)}$$

Here, β_2 is the coefficient on the summer dummy and $Cov(X_1, X_2)$ is the covariance between price and the summer dummy. We know that $\beta_2 = 358.50$, and we can calculate $Cov(price, summer\_dummy)$:

```
cov(vendor_1$price, vendor_1$summer_dummy)
```

```
## [1] 0.127451
```

So, $\beta_2 > 0$ and $Cov(price, summer\_dummy) > 0$, so the bias is positive and the estimate of β_1 in the first regression is an overestimate since the summer dummy is omitted. This is why the coefficient on price decreases from -31.23 to -141.19 upon including the summer dummy in the regression.

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).

```
# data frame for vendor 2 only
vendor_2 = ice_cream[ice_cream$vendor_id == 2,]
# univariate regression of sales on price
summary(lm(sales ~ price, data = vendor_2))
```

```
##
## Call:
## lm(formula = sales ~ price, data = vendor_2)
##
## 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
```

```
# multivariate regression of sales on price and summer_dummy
summary(lm(sales ~ price + summer_dummy, data = vendor_2))
```

```
##
## Call:
## lm(formula = sales ~ price + summer_dummy, data = vendor_2)
##
## 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
```

```
# investigating the correlation between price and summer dummy
cor(vendor_2$price, vendor_2$summer_dummy)
```

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

ANSWER: We can see that the correlation between price and the summer dummy in the data for vendor 2 is 1. This means that there is a problem of perfect multicollinearity, which is a violation of the regression assumptions. This is why the coefficient on the summer dummy in the above regression is NA.

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: If prices did not change systematically during the summer, then price would not be correlated with the summer dummy variable. Because of this, the coefficient would likely not change much regardless of whether or not the summer dummy was included. Including the summer dummy, however, would increase the precision of the coefficient.