PRESCRIPTIVE MODELS AND DATA ANALYTICS

Problem Set #2

Arnav Garg (906310841)

1 Hospital admission & quality of service

Download health data.csv and load it into python. 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.

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

In [ ]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import os
import sys
import statsmodels.api as sm
from statsmodels.formula.api import ols
import warnings
warnings.filterwarnings('ignore')
```

In [ ]:
```
# Load dataset
health_data = pd.read_csv('health_data.csv')

# Print the number of rows and columns
print(health_data.shape)

# Print the first few rows
health_data.head()
```

```
(24480, 6)
```

Out[ ]:

	patient_id	hospital_id	admin_year	patient_died_dummy	startage	female_dummy
0	1	D	2003	0	81	0
1	2	H	2003	1	67	0
2	3	A	2003	0	54	0
3	4	E	2003	0	81	0
4	5	H	2003	0	69	0

In [ ]:
```
model = ols('patient_died_dummy ~ hospital_id', data = health_data).fit()
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     patient_died_dummy   R-squared:                       0.042
Model:                            OLS   Adj. R-squared:                  0.042
Method:                 Least Squares   F-statistic:                     119.3
Date:                Sun, 11 Feb 2024   Prob (F-statistic):         1.75e-220
Time:                        18:47:49   Log-Likelihood:                -7416.5
No. Observations:               24480   AIC:                         1.485e+04
Df Residuals:                   24470   BIC:                         1.493e+04
Df Model:                           9
Covariance Type:            nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
Intercept            0.0970      0.006     16.355      0.000       0.085       0.109
hospital_id[T.B]     0.0072      0.008      0.890      0.373      -0.009       0.023
hospital_id[T.C]    -0.0483      0.010     -4.776      0.000      -0.068      -0.028
hospital_id[T.D]     0.1882      0.008     23.188      0.000       0.172       0.204
hospital_id[T.E]    -0.0531      0.011     -4.771      0.000      -0.075      -0.031
hospital_id[T.F]     0.0003      0.009      0.030      0.976      -0.017       0.017
hospital_id[T.G]     0.0441      0.008      5.273      0.000       0.028       0.061
hospital_id[T.H]     0.0038      0.009      0.414      0.679      -0.014       0.022
hospital_id[T.I]     0.0318      0.009      3.480      0.001       0.014       0.050
hospital_id[T.J]     0.0112      0.011      1.028      0.304      -0.010       0.032
==============================================================================
Omnibus:                     9228.833   Durbin-Watson:                   2.025
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            25954.225
Skew:                           2.096   Prob(JB):                         0.00
Kurtosis:                       5.807   Cond. No.                         10.0
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

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

The coefficient on the constant term (which is essentially a dummy for hospital A) is ~0.097 and the coefficient on the dummy for hospital D is ~0.188. This means that the probability of death for patients who went to hospital D is 18.8% higher than the patients who went to hospital A. The probability of death for patients who went to hospital A is 9.7%

In [ ]:
```
model.params[0], model.params[3]
```
Out[ ]:
```
(0.09701737135364538, 0.18824729245181226)
```

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

The difference between mortality rates at hospital D and E is ~0.2414.

In [ ]:
```
model.params[3] - model.params[4]
```
Out[ ]:
```
0.24139049162003756
```

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.

In [ ]:
```
## data for patients that visited either hospital A or B.
q2 = health_data[health_data['hospital_id'].isin(['A', 'B'])]
## dummy for whether the patient visited hospital B.
q2['hospital_B_dummy'] = 0
q2.loc[q2.hospital_id == 'B', 'hospital_B_dummy'] = 1
```

In [ ]:
```
model = ols('patient_died_dummy ~ hospital_B_dummy', data = q2).fit()
print(model.summary())
```

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:     patient_died_dummy   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                 -0.000
Method:                 Least Squares   F-statistic:                    0.9377
Date:                Sun, 11 Feb 2024   Prob (F-statistic):              0.333
Time:                        18:47:49   Log-Likelihood:                -1446.8
No. Observations:                6611   AIC:                             2898.
Df Residuals:                    6609   BIC:                             2911.
Df Model:                           1                                         
Covariance Type:            nonrobust                                         
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
Intercept            0.0970      0.005     17.791      0.000       0.086       0.108
hospital_B_dummy     0.0072      0.007      0.968      0.333      -0.007       0.022
==============================================================================
Omnibus:                     3377.258   Durbin-Watson:                   2.031
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            14683.814
Skew:                           2.650   Prob(JB):                         0.00
Kurtosis:                       8.022   Cond. No.                         2.72
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

(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).

The difference in mortality rate implied by this regression cannot be interpreted as the causal effect of visiting a different hospital because the regression does not account for the fact that patients who visit hospital B may be different from those who visit hospital A in ways that are correlated with the outcome. For example, patients who visit hospital B may be sicker than those who visit hospital A, and this difference in patient health may be correlated with the outcome. If this is the case, the difference in mortality rate between the two hospitals may be due to differences in patient health rather than differences in the quality of care provided by the hospitals.

(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.

Theoretically, we do not know whether the difference in mortality between hospitals is over or under estimated. However, by running a regression of mortality on potential control variables, we can figure out practically if the difference is over or under estimated in this dataset. We observe that the average age of patients who go to hospital B (~64.9 years) is slightly less than the average age of patients who go to hospital A (~65.7 years). Also, we observe that hospital A recieves 23.2% females vs hospital B which receives 20.5% females. By running a regression, we find out the following:

1. Difference is very slightly (statistically insignificant) overestimated when only the variable "startage" is used as control variable
2. Difference is underestimated when variable "female_dummy" is used as control variable
3. Difference is overall underestimated when both variables "startage" and "female_dummy" are used as control variables

In [ ]:
```
q2.groupby('hospital_id')['startage'].mean(), q2.groupby('hospital_id')['female_dummy'].value_counts()/q2.groupby('hos
```

Out[ ]:
```
(hospital_id
 A    65.705015
 B    64.882303
 Name: startage, dtype: float64,
 hospital_id  female_dummy
 A            0               76.794494
              1               23.205506
 B            0               79.494382
              1               20.505618
 dtype: float64)
```

(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.

As dicussed in part a, we can use factors such as age, gender, history of diseases, income level, etc as control variables to get closer to a causal estimate. However, we only have age and gender in our dataset so we use those as our control variables. Both hospitals might have a different distribution of age and gender of patients which will affect the mortality rate of such patients.

As mentioned in part b, we observe the following changes:

1. Coefficient of hospital_B_dummy goes down from 0.0072 to 0.0071 when we use "startage" as a control variable, meaning we were slightly overestimating the affect of hospital B on mortality

2. Coefficient of hospital_B_dummy goes up from 0.0072 to 0.0121 when we use "female_dummy" as a control variable, meaning we were highly underestimating the affect of hospital B on mortality
3. Coefficient of hospital_B_dummy goes up overall from 0.0072 to 0.0114 when we use both "startage" and "female_dummy" as a control variables, meaning overall we were underestimating the affect of hospital B on mortality

In [ ]:
```
model = ols('patient_died_dummy ~ hospital_B_dummy + startage + female_dummy', data = q2).fit()
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     patient_died_dummy   R-squared:                       0.063
Model:                            OLS   Adj. R-squared:                  0.062
Method:                 Least Squares   F-statistic:                     147.6
Date:                Sun, 11 Feb 2024   Prob (F-statistic):           1.43e-92
Time:                        18:47:49   Log-Likelihood:                -1232.8
No. Observations:                6611   AIC:                             2474.
Df Residuals:                    6607   BIC:                             2501.
Df Model:                           3
Covariance Type:            nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
Intercept            0.1165      0.027      4.335      0.000       0.064       0.169
hospital_B_dummy     0.0114      0.007      1.579      0.114      -0.003       0.026
startage            -0.0009      0.000     -2.347      0.019      -0.002      -0.000
female_dummy         0.1836      0.009     21.015      0.000       0.167       0.201
==============================================================================
Omnibus:                     3080.506   Durbin-Watson:                   2.037
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            12181.285
Skew:                           2.411   Prob(JB):                         0.00
Kurtosis:                       7.580   Cond. No.                         495.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

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 Python. 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.

The omitted variable bias formula is given by

$$\beta_{price,LR} = \beta_{price,MR} + \beta_{summer,MR} * \frac{Cov(price, summer)}{Var(price)} \tag{1}$$

$$\beta_{price,LR} = -141.2 + 358.50 \times \frac{0.12745}{0.41553} \tag{2}$$

$$\beta_{price,LR} = -31.24(asobserved) \tag{3}$$

In this case, the price coefficient changes when the summer dummy is included in the regression because the summer dummy is correlated with the price variable. This means that the price variable is endogenous. The price coefficient in the first regression is biased because it does not account for the omitted variable, which is the summer dummy. When the summer dummy is included in the regression, the price coefficient changes to account for the omitted variable bias.

In [ ]:

```
# Load dataset
demand_data = pd.read_csv('demand_data.csv')

# Print the number of rows and columns
print(demand_data.shape)

# Print the first few rows
demand_data.head()
```

```
(5200, 5)
```

Out[ ]:

	vendor_id	week	summer_dummy	price	sales
0	1	1	0	2.0	8788.7383
1	1	2	0	3.0	8937.9863
2	1	3	0	3.0	8740.1777
3	1	4	0	3.0	8757.1338
4	1	5	0	3.0	8739.6104

In [ ]:

```
model = ols('sales ~ price', data = demand_data.loc[demand_data.vendor_id == 1]).fit()
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  sales   R-squared:                       0.006
Model:                            OLS   Adj. R-squared:                 -0.013
Method:                 Least Squares   F-statistic:                    0.3250
Date:                Sun, 11 Feb 2024   Prob (F-statistic):              0.571
Time:                        18:47:49   Log-Likelihood:                -360.33
No. Observations:                  52   AIC:                             724.7
Df Residuals:                      50   BIC:                             728.6
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   8983.8227    145.437     61.771      0.000    8691.704    9275.941
price        -31.2310     54.782     -0.570      0.571    -141.264      78.802
==============================================================================
Omnibus:                        3.319   Durbin-Watson:                   1.572
Prob(Omnibus):                  0.190   Jarque-Bera (JB):                2.367
Skew:                           0.346   Prob(JB):                        0.306
Kurtosis:                       3.784   Cond. No.                         12.5
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

In [ ]:

```
model = ols('sales ~ price + summer_dummy', data = demand_data.loc[demand_data.vendor_id == 1]).fit()
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  sales   R-squared:                       0.318
Model:                            OLS   Adj. R-squared:                  0.290
Method:                 Least Squares   F-statistic:                     11.42
Date:                Sun, 11 Feb 2024   Prob (F-statistic):           8.49e-05
Time:                        18:47:49   Log-Likelihood:                -350.56
No. Observations:                  52   AIC:                             707.1
Df Residuals:                      49   BIC:                             713.0
Df Model:                           2
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept     9177.5500    128.432     71.458      0.000    8919.455    9435.644
price         -141.1887     51.407     -2.746      0.008    -244.496     -37.882
summer_dummy   358.5012     75.790      4.730      0.000     206.195     510.807
==============================================================================
Omnibus:                        0.027   Durbin-Watson:                   1.690
Prob(Omnibus):                  0.986   Jarque-Bera (JB):                0.078
Skew:                           0.039   Prob(JB):                        0.962
Kurtosis:                       2.828   Cond. No.                         13.7
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

In [ ]:
```
demand_data.loc[demand_data.vendor_id == 1].cov()
```

Out[ ]:

	vendor_id	week	summer_dummy	price	sales
vendor_id	0.0	0.000000	0.000000	0.000000	0.000000
week	0.0	229.666667	0.382353	0.313725	204.543909
summer_dummy	0.0	0.382353	0.191176	0.127451	50.542365
price	0.0	0.313725	0.127451	0.415535	-12.977568
sales	0.0	204.543909	50.542365	-12.977568	62758.194765

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 there might be multicollinearity

problems. Why does this happen?

In the case of the regression with the summer dummy, there exist multicollinearity problems because vendor 2 systematically prices their products higher during the summer months. This means that there is a perfect correlation between price and summer_dummy which gives rise to multicollinearity.

In [ ]:
```python
model = ols('sales ~ price', data = demand_data.loc[demand_data.vendor_id == 2]).fit()
print(model.summary())
```

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                  sales   R-squared:                       0.133
Model:                            OLS   Adj. R-squared:                  0.116
Method:                 Least Squares   F-statistic:                     7.684
Date:                Sun, 11 Feb 2024   Prob (F-statistic):            0.00781
Time:                        18:47:49   Log-Likelihood:                -359.10
No. Observations:                  52   AIC:                             722.2
Df Residuals:                      50   BIC:                             726.1
Df Model:                           1                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   8411.1748    219.545     38.312      0.000    7970.205    8852.145
price        218.6028     78.863      2.772      0.008      60.202     377.004
==============================================================================
Omnibus:                        1.154   Durbin-Watson:                   2.369
Prob(Omnibus):                  0.562   Jarque-Bera (JB):                0.467
Skew:                           0.114   Prob(JB):                        0.792
Kurtosis:                       3.404   Cond. No.                         20.2
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

In [ ]:
```python
model = ols('sales ~ price + summer_dummy', data = demand_data.loc[demand_data.vendor_id == 2]).fit()
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  sales   R-squared:                       0.133
Model:                            OLS   Adj. R-squared:                  0.116
Method:                 Least Squares   F-statistic:                     7.684
Date:                Sun, 11 Feb 2024   Prob (F-statistic):            0.00781
Time:                        18:47:49   Log-Likelihood:                -359.10
No. Observations:                  52   AIC:                             722.2
Df Residuals:                      50   BIC:                             726.1
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept     2105.3159     29.848     70.534      0.000    2045.364    2165.268
price         2740.9463     10.951    250.283      0.000    2718.950    2762.943
summer_dummy -2522.3436     75.986    -33.195      0.000   -2674.966   -2369.721
==============================================================================
Omnibus:                        1.154   Durbin-Watson:                   2.369
Prob(Omnibus):                  0.562   Jarque-Bera (JB):                0.467
Skew:                           0.114   Prob(JB):                        0.792
Kurtosis:                       3.404   Cond. No.                     8.19e+15
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 6.84e-30. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

In [ ]:

```python
demand_data.loc[demand_data.vendor_id == 2].corr()
```

Out[ ]:

	vendor_id	week	summer_dummy	price	sales
vendor_id	NaN	NaN	NaN	NaN	NaN
week	NaN	1.000000	0.057703	0.057703	0.152865
summer_dummy	NaN	0.057703	1.000000	1.000000	0.364969
price	NaN	0.057703	1.000000	1.000000	0.364969
sales	NaN	0.152865	0.364969	0.364969	1.000000

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?

The price coefficient estimate would be the same in both regressions because price and summer_dummy are uncorrelated in this case. Hence, the bias would be zero. However, precision would be higher in the regression with the summer dummy because variance is lower when more variables are added to the regression.