Problem Set 3

Predictive Analytics for Business Strategy

Spring 2023

McDermott

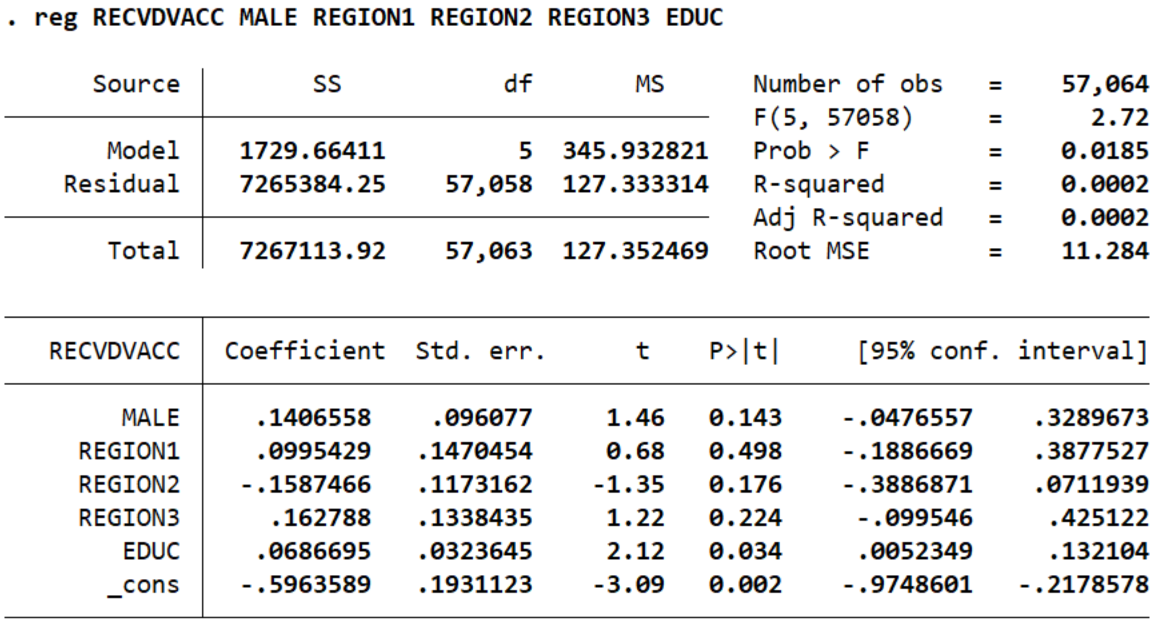
Team Members :

- **Jimmy Yang**
- **Youngsun Song**
- **Gege Susilohadi**
- **Arun Thomas**
- **Yashada Nikam**

1. **In this problem, you will use the Pulse data in Canvas to build a model for passive prediction. You will passively predict whether someone has gotten a Covid 19 Vaccine.**
   a. **What would you include and why? (include at minimum 3 independent variables)**
      i. I would include **Region**, **Educ,** and **Genid_birth**. I picked these three because we're trying to perform a passive prediction for COVID-19 vaccination, and those three are all variables that can only be passively observed. Furthermore, none of those three are strongly correlated with each other, in theory. While there could be some weak correlation between **Region** and **Educ**, it shouldn't really be strong enough to cause a multicollinearity issue.

   b. **How can you tell if you did a good job.**
      i. High R^2 and adjusted R^2
      ii. All independent variables are statistically significant for the desired critical value.
      iii. The "eyeball" test. Plot the model and see how well it fits a scatterplot of the data.

c. **Give an example of a passive prediction. Show and explain the results (both the regression model estimates and how you calculate what you passively predict for Y or for the change in Y).**



```
. reg RECVDVACC MALE REGION1 REGION2 REGION3 EDUC

      Source |       SS           df       MS      Number of obs   =    57,064
-------------+----------------------------------   F(5, 57058)     =      2.72
       Model |  1729.66411         5  345.932821   Prob > F        =    0.0185
    Residual |  7265384.25    57,058  127.333314   R-squared       =    0.0002
-------------+----------------------------------   Adj R-squared   =    0.0002
       Total |  7267113.92    57,063  127.352469   Root MSE        =    11.284

------------------------------------------------------------------------------
   RECVDVACC | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        MALE |   .1406558    .096077     1.46   0.143    -.0476557    .3289673
     REGION1 |   .0995429   .1470454     0.68   0.498    -.1886669    .3877527
     REGION2 |  -.1587466   .1173162    -1.35   0.176    -.3886871    .0711939
     REGION3 |    .162788   .1338435     1.22   0.224     -.099546     .425122
        EDUC |   .0686695   .0323645     2.12   0.034     .0052349     .132104
       _cons |  -.5963589   .1931123    -3.09   0.002    -.9748601   -.2178578
------------------------------------------------------------------------------
```

For this model, we picked three independent variables that can only be passively observed: **GENID_BIRTH**, **EDUC**, and **REGION**. Since **GENID_BIRTH** and **REGION** were both categorical variables, we created a few dummy variables. **MALE** is a dummy variable that has a value of 1 if **GENID_BIRTH** =1, **REGION1** is a dummy variable that has a value of 1 where **REGION** has a value of 1, **REGION2** and **REGION3** are similar but with 2 and 3 respectively. This means that a female from **REGION4** would be the baseline comparison.

Unfortunately, it looks like neither **REGION** nor **GENID_BIRTH** was statistically significant. Each **REGION** variable had a P value larger than 0.05, and **MALE** had a P value of 0.143. **EDUC** was statistically significant.

Since **RECDVACC** is a binary variable that can be either 0 or 1, observed changes correlated with our independent variables can be interpreted as percentage changes. For instance, **EDUC**'s coefficient is **0.0686695**. This can be interpreted as "each additional year of education is correlated with a **6.87%** increase in the chance that a particular person will receive the COVID vaccine, holding all else constant."

d. **Explain whether any of the variables you included in your model are control variables and why/why not.**

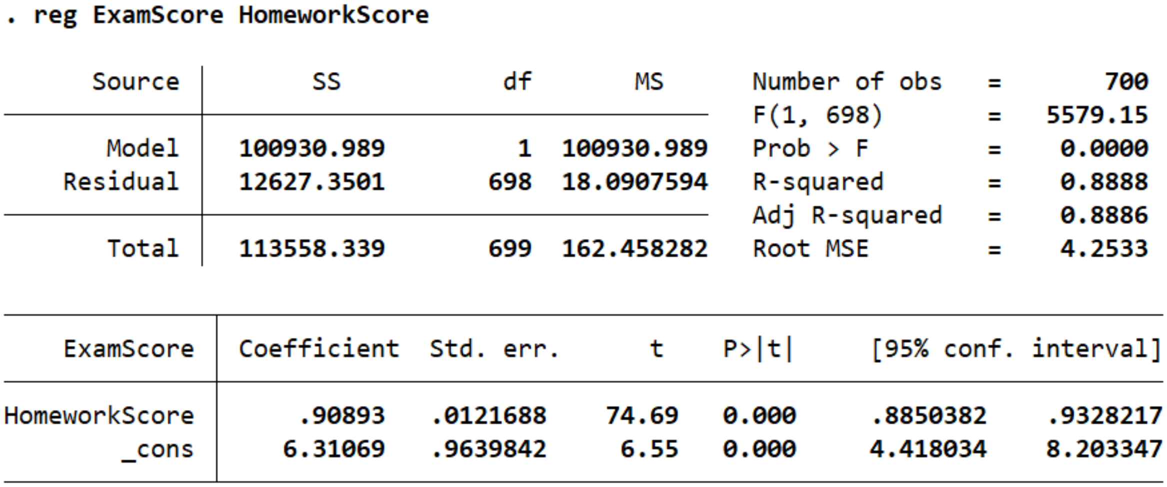
We have been handling a passive prediction, so we haven't included any control variables to alleviate an endogeneity problem. In other words, we didn't include any endogenous variables in our model and thus do not need control variables to account for endogeneity.

e. **Does your model have a variable of interest? Explain.**

No, this is meant to be a model of passive prediction, which does not have a variable of interest.

2. **Suppose you have observational data on student exam scores in worksheet 1 of PS3.xlsx.**

a. **Report your results of this simple regression with a screenshot.**



```
. reg ExamScore HomeworkScore

      Source |       SS           df       MS      Number of obs   =       700
-------------+----------------------------------   F(1, 698)       =   5579.15
       Model |  100930.989         1  100930.989   Prob > F        =    0.0000
    Residual |  12627.3501       698  18.0907594   R-squared       =    0.8888
-------------+----------------------------------   Adj R-squared   =    0.8886
       Total |  113558.339       699  162.458282   Root MSE        =    4.2533

------------------------------------------------------------------------------
   ExamScore | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
HomeworkScore |     .90893   .0121688    74.69   0.000     .8850382    .9328217
       _cons |    6.31069   .9639842     6.55   0.000     4.418034    8.203347
------------------------------------------------------------------------------
```

b. **Discuss the results.**

This simple model has a decently high R^2 and adjusted R^2, this means that this model is good at explaining the variation in the dependent variable. HomeworkScore is a statistically significant variable, as shown by the very low P value. It has a coefficient of **0.90893**, this means that each additional point earned on the homework assignment is correlated with an average increase of **0.90893** points on the exam, holding all other factors constant.

c. **Discuss possible confounding factors.**

Since an exam score is only representative of one instance of a student's performance, there are many possible confounding factors at play.

Did the student eat breakfast that day, Yes or No. It's a well-known fact that hungry students do not perform as well as well-fed students. Students that are not eating well are generally not doing as well on their homework.

How many hours of sleep did a student get last night? Sleepy and tired students would not perform as well on an exam compared to a student who received plenty of sleep. Chronically tired students do not do as well on their homework, so this U factor is also correlated with X.

Household income. Does the student live in a high- or low-income household? High-income households can afford to hire tutors, enroll a student in academic extracurriculars and can afford more general study time for a student. Where as a student from a low-income household may have to work to support their family instead of studying for an exam or doing their homework. So this is correlated with X and has an effect on Y.

d. Which confounding factors are you likely to be able to use as control variables (*meaning they wouldn't actually be CFs*) and which ones are you likely not able to (*assume you aren't necessarily limited to the data for the latter*).

We can most likely use "**hours of sleep**" and "**household income**" as control variables. You can use some pretty simple categorical variables (*i.e. "StudentTired: Y/N, Income: High/Low"*) as control variables.

On the other hand, whether or not a student ate breakfast is probably harder to control for, since that itself has so many variables (*caloric size of breakfast, quality of breakfast, nutritional variety, etc*.) and could arguably be said to be correlated with household income.

e. Let's say determination is a confounding factor (you won't get credit for citing it above). Propose a proxy variable and discuss how it satisfies all 3 assumptions/requirements of a good proxy variable.

Proposed proxy variable: average GPA in high school. Assuming this is an undergrad/graduate level exam.

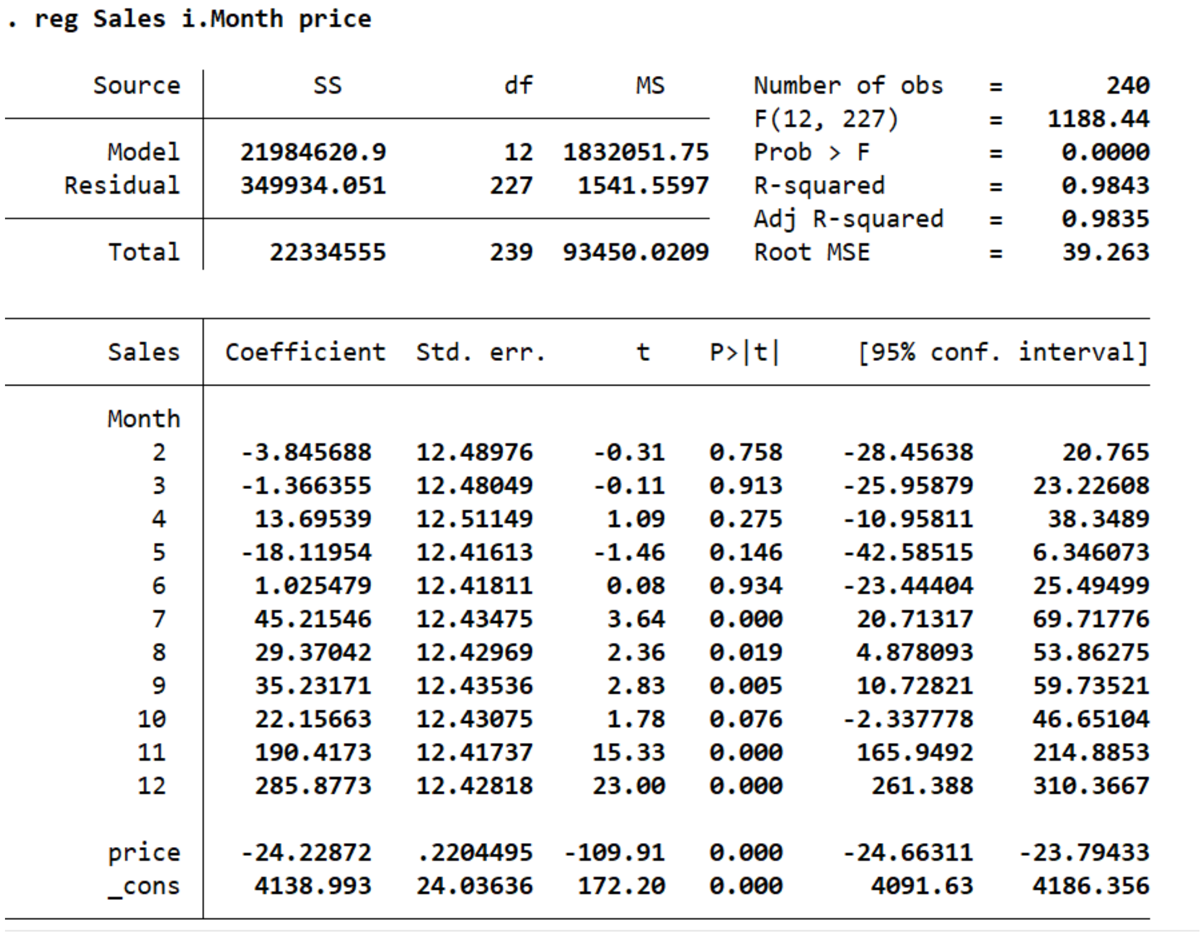
A student's overall GPA in high school should be correlated with determination. More determined students are more likely to maintain a good GPA throughout their academic career, thus satisfying assumption 1.

High school GPA cannot directly affect a student's college exam score, thus satisfying assumption 2.

There are other factors that can determine the omitted variable. For instance: the number of hours spent volunteering, the number of miles ran in a week or the number of competitions won in high school. All of those are good indicators for determination, but none of them are correlated with high school GPA, which satisfies assumption 3.

3. **Suppose you have monthly observational data on sales of a specific children's toy in worksheet 2 of PS3.xlsx.**

   a. **Report your results of this simple regression with a screenshot.**



```
. reg Sales i.Month price

      Source |       SS           df       MS      Number of obs   =       240
-------------+----------------------------------   F(12, 227)      =   1188.44
       Model |  21984620.9        12  1832051.75   Prob > F        =    0.0000
    Residual |  349934.051       227   1541.5597   R-squared       =    0.9843
-------------+----------------------------------   Adj R-squared   =    0.9835
       Total |    22334555       239  93450.0209   Root MSE        =    39.263

------------------------------------------------------------------------------
       Sales | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       Month |
          2  |  -3.845688   12.48976    -0.31   0.758    -28.45638      20.765
          3  |  -1.366355   12.48049    -0.11   0.913    -25.95879    23.22608
          4  |   13.69539   12.51149     1.09   0.275    -10.95811     38.3489
          5  |  -18.11954   12.41613    -1.46   0.146    -42.58515    6.346073
          6  |   1.025479   12.41811     0.08   0.934    -23.44404    25.49499
          7  |   45.21546   12.43475     3.64   0.000     20.71317    69.71776
          8  |   29.37042   12.42969     2.36   0.019     4.878093    53.86275
          9  |   35.23171   12.43536     2.83   0.005     10.72821    59.73521
         10  |   22.15663   12.43075     1.78   0.076    -2.337778    46.65104
         11  |   190.4173   12.41737    15.33   0.000     165.9492    214.8853
         12  |   285.8773   12.42818    23.00   0.000      261.388    310.3667
             |
       price |  -24.22872   .2204495  -109.91   0.000    -24.66311   -23.79433
       _cons |   4138.993   24.03636   172.20   0.000      4091.63    4186.356
------------------------------------------------------------------------------
```

   b. **Clearly explain and discuss the results (*there are lots of categories so focus on month11, month12, and price*).**
   November sales are, on average, **190.42** units higher than sales in **January**. Similarly, **December** sales are, on average, **285.88** units higher than sales in **January**.
   For price, each additional dollar in price is correlated with an average decrease of **24.23** units in sales.

   c. **Discuss possible confounding factors.**
   Received a bonus at work (yes/no). Bonuses are typically handed out at the end of the year, so a higher than typical salary due to the bonus would cause increased spending on kids' toys. So this is correlated with month and affects sales units.

The number of holidays per month. This is correlated with all of the variables in the model. In the US, most major holidays are in July or later, so this correlates with the month variables. Companies will sometimes do price adjustments during the holidays (i.e. Black Friday), thus price also correlates with the number of holidays per month.

d. **Which confounding factors are you likely to be able to use as control variables (meaning they wouldn't actually be CFs) and which ones are you likely not able to (assume you aren't necessarily limited to the data here).**
The number of holidays per month could be used as a control variable. Something like "Month_fed_holiday Y/N" would allow for us to control this factor.

Since there's no standardized bonus schedule for all companies in the US, it would be hard to use "received a bonus at work" as a control variable.

4. **Can you use estimates from an RCT to make a passive prediction? Explain why or why not.**
No. RCT's generate experimental data, passive predictions require observational data.