Problem Set 2
Predictive Analytics for Business Strategy

1. We are interested in estimating demand for coffee sold in retail shops around the U.S.

a. Would regressing QuantitySold on Price give you a good estimate of the causal impact of Price on QuantitySold? Explain why or why not.

Regressing QuantitySold on Price alone might not give a reliable estimate of the causal impact of Price on QuantitySold due to the potential endogeneity of the Price variable and it being a naïve model. Price could be endogenous if it's correlated with other unobserved factors that also influence QuantitySold, leading to biased and inconsistent estimates. For example, factors such as FuelCost and LandCost might be correlated with both Price and QuantitySold, creating an endogeneity problem.

For instance, we see that a $1 increase in price has a positive impact on the quantity sold of 19 units. This does not make intuitive or economic sense. Other things will influence price, such as local competitiveness and local income.

To address this issue, we can implement a two-stage least squares (2SLS) regression and use FuelCost and LandCost as instrumental variables. These variables are suitable as instruments under the assumption that they are correlated with Price (relevance condition), but not correlated with the error term U in the QuantitySold equation (exogeneity condition). This way, the variation in Price that is correlated with these instruments is 'purged' of endogeneity, leading to more reliable estimates of the causal effect of Price on QuantitySold.

```
. reg QuantitySold Price
```

Source	SS	df	MS	Number of obs	=	488
				F(1, 486)	=	537.18
Model	74479.5718	1	74479.5718	Prob > F	=	0.0000
Residual	67383.7982	486	138.649791	R-squared	=	0.5250
				Adj R-squared	=	0.5240
Total	141863.37	487	291.300554	Root MSE	=	11.775

QuantitySold	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
Price	19.1963	.8282444	23.18	0.000	17.56892	20.82368
_cons	224.2483	3.936793	56.96	0.000	216.5131	231.9836

b. Classify the type of variable Income is and explain why it falls into that category.

Income is an exogenous variable because it is not correlated with U. This designation means that it's a variable whose value is determined outside the model and is independent from the model's error term. It is generally assumed that income is determined by factors not captured in this model, such

as employment, career choices, education, and so on, and that it doesn't depend on the quantity of the goods sold.

FuelCost and LandCost not associated with Income: The variables "FuelCost" and "LandCost" represent expenses incurred by the coffee shops and, presumably, affect the pricing decisions made by those shops. On the other hand, "Income" represents the earnings of consumers. While there could be broader macroeconomic factors that influence all of these variables, in the context of this model, it is generally assumed that individual or household income levels do not directly influence the fuel or land costs for businesses. Likewise, the costs a business incurs for fuel or land do not directly determine the income levels of consumers. Thus, we can assume that "FuelCost" and "LandCost" are not associated with "Income".

Exogeneity is an essential assumption in many econometric models because it assures us that there's no correlation between the independent variables and the error term, which is a requirement for unbiased and consistent estimates in Ordinary Least Squares (OLS) regression.

c. Classify the type of variable FuelCost is and explain why it falls into that category.

FuelCost is an instrumental variable because it satisfies the conditions of instrument validity:

1. **Relevance condition: FuelCost is correlated with Price (e.g., through a cost-pass-through mechanism where rising fuel costs may increase the price of goods like coffee).**
2. **Exogeneity condition: It does not directly impact QuantitySold, affecting it only through Price. Since FuelCost does not appear in the error term of the QuantitySold equation, it does not introduce endogeneity.**

d. Classify the type of variable LandCost is and explain why it falls into that category.

Like FuelCost, LandCost can also be thought of as an exogenous variable. It may be correlated with the Price variable (relevance condition) as changes in land costs can impact the prices set by coffee shops. This correlation can introduce endogeneity into the model. Yet, we posit that LandCost is uncorrelated with the error term (U)(exogeneity condition), making it another potential instrumental variable. This implies that LandCost does not have a direct impact on QuantitySold, apart from its influence on Price. By using LandCost as an instrumental variable, we can address the endogeneity problem associated with Price, leading to more consistent and unbiased estimates of the Price effect on QuantitySold.

e. Run an instrumental regression and do the following:

```
. reg Price FuelCost LandCost Income

      Source |       SS           df       MS      Number of obs   =       488
-------------+----------------------------------   F(3, 484)       =     68.53
       Model |  60.2568409         3  20.0856136   Prob > F        =    0.0000
    Residual |  141.859861       484  .293098887   R-squared       =    0.2981
-------------+----------------------------------   Adj R-squared   =    0.2938
       Total |  202.116702       487  .415024029   Root MSE        =    .54139

------------------------------------------------------------------------------
       Price | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    FuelCost |    1.039208   .081633    12.73   0.000     .8788095    1.199607
    LandCost |    .2269668   .0814607    2.79   0.006     .0669065    .3870271
      Income |    .0226918   .0040531    5.60   0.000     .0147279    .0306558
       _cons |    .6492615   .3279984    1.98   0.048     .0047848    1.293738
------------------------------------------------------------------------------
```

i. Provide evidence that any instrument you use satisfies the requirements of a valid instrument.

To show that FuelCost and LandCost are valid instruments:

- **Relevance:**
  - **FuelCost is correlated with Price (coefficient: 1.039208, p-value: 0.000).**
  - **LandCost is correlated with Price (coefficient: 0.22696, p-value: 0.006). Both are statistically significant, confirming relevance.**
- **Exogeneity:**
  - **The assumption is that neither FuelCost nor LandCost directly affects QuantitySold. There is no theoretical reason to believe that fuel or land costs would influence coffee demand apart from their indirect effect via Price.**

ii. Provide the results of your 2sls regression and discuss what you can say from that regression and why (or what you can't say and why)?

```
. ivreg QuantitySold Income (Price = FuelCost LandCost)

Instrumental variables 2SLS regression

      Source |       SS           df       MS      Number of obs   =       488
-------------+----------------------------------   F(2, 485)       =     34.18
       Model |  -47380.6086         2 -23690.3043   Prob > F        =    0.0000
    Residual |  189243.979        485   390.19377   R-squared       =         .
-------------+----------------------------------   Adj R-squared   =         .
       Total |   141863.37        487  291.300554   Root MSE        =    19.753

------------------------------------------------------------------------------
QuantitySold | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       Price |   -8.341385    2.76249    -3.02   0.003    -13.76931    -2.91346
      Income |     1.32532   .1603586     8.26   0.000     1.010237    1.640404
       _cons |    301.0441   12.09144    24.90   0.000      277.286    324.8022
------------------------------------------------------------------------------
Instrumented: Price
 Instruments: Income FuelCost LandCost
```

The 2SLS regression model provides an estimate of the causal effect of the endogenous variable (Price) on the dependent variable (Quantity Sold), purged of the endogeneity bias that could be present in a

single-equation OLS regression. In this case, the coefficient on Price hat (Price predicted from the first stage regression) is -8.341385 with a p-value of 0.003.

Interpretation of the Price coefficient: The coefficient of -8.341385 for Price suggests that a $1 increase in the price of coffee decreases the quantity sold by approximately 8.34 units, holding all other factors constant. The negative sign indicates an inverse relationship between Price and Quantity Sold, which is consistent with the law of demand. The p-value of 0.003 is less than 0.05, which suggests that this relationship is statistically significant at conventional levels.

What you can't directly interpret from these results:

Causality: While IV methods like 2SLS can help establish causal relationships, they don't prove causality on their own. The assumptions behind the model need to hold true, and even then, the results only show a correlation from which we infer a causal relationship based on those assumptions.

2. Consider now a Yelp example. The data from the Yelp spreadsheet in the PS2 data file contain information on Sales, average Yelp rating, and Yelp stars for a number of comparable restaurants (say, mid-level-priced, American food).

a. When we use a regression discontinuity model, we exploit cases where there is an arbitrary jump/change in treatment status at a specific point and compare those just above and just below that cutoff. Describe, in words, what that jump looks like, what is jumping, how the treatment status is changing, and what we are trying to learn. You may wish to sort one variable at a time (either in Excel or after importing into Stata) if you are not sure where it might be.

In a regression discontinuity (RD) model, a jump refers to an abrupt change in the outcome variable at a specific cutoff point in a forcing (continuous) variable. In this case:

- **The continuous forcing variable is average Yelp rating.**
- **The treatment status changes when restaurants cross from one Yelp star category to another (e.g., from 1 star to 2 stars).**

This jump is used to identify the causal effect of changes in treatment status (e.g., star rating) on the outcome (sales), while observations close to the cutoff serve as comparable groups.

b. In the data, what is the:

i. Outcome?

Sales

ii. Forcing (continuous) variable?

Average Yelp rating

iii. Treatment (categorical) variable?

Yelp stars for number of comparable restaurants

c. Execute regression discontinuity to determine whether a change from 1 star to 2 stars impacts sales. What do you conclude?

```
. rdrobust Sales Rating, c(1.495)
Mass points detected in the running variable.

Sharp RD estimates using local polynomial regression.

  Cutoff c = 1.495 | Left of c  Right of c           Number of obs =       1483
                   |                                 BW type       =      mserd
     Number of obs |       174        1309           Kernel        = Triangular
Eff. Number of obs |       143         168           VCE method    =         NN
    Order est. (p) |         1           1
    Order bias (q) |         2           2
       BW est. (h) |     0.412       0.412
       BW bias (b) |     0.538       0.538
         rho (h/b) |     0.765       0.765
       Unique obs  |        49         342

Outcome: Sales. Running variable: Rating.

------------------------------------------------------------------------------
      Method |     Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Conventional |    2517.6     1522.3    1.6539   0.098    -465.987      5501.22
      Robust |         -           -    1.6903   0.091    -514.566      6965.68
------------------------------------------------------------------------------

Estimates adjusted for mass points in the running variable.
```

There tends (P = 0.098) to be a difference in sales between restaurants earning a 1-star rating and those earning a two-star rating, but those differences do not reach the 95% confidence level. Still, if I am the restaurant owner, I would strive to earn two stars instead of one star because 2-star rated restaurants have sales that are $2,517.60 greater than 1-star rated restaurants.

d. If you widen the bandwidth of your regression discontinuity estimator, what are the consequences? Give an example of a bandwidth that would be unreasonably large and explain why it would be too large. (The bandwidth value is applied to each side of the cutoff)

Narrowing the bandwidth decreases the bias. Conversely, increasing the bandwidth increases the bias. That is because as the bandwidth increases more potential confounding factors are introduced. In other words, as you move farther from the cutoff, the more different observations are likely to be for reasons unrelated to the treatment.

Using the class example of comparing driving accidents of people older and younger than 21, a bandwidth of 50 years would be unreasonably wide because this would potentially compare people that are 71 years old for the treatment older than 21 with those that are yet to be conceived for the treatment younger than 21 years old. Obviously, babies do not drive so they would artificially have very low accident rates and would provide incorrect and misleading conclusions.

An example of a bandwidth that would be unreasonably large would be a bandwidth that crosses over to the next jump point. For example, if you are trying the measure the impact of the Sales jump from 1 star to 2 stars and the cutoff is 1.495 average yelp rating, and the Sales jump from 2 stars to 3 stars has a 2.5 average yelp rating at 3 stars (the first 3 star from the 2-3 star cutoff), a bandwidth of 1.005 (2.5-1.495) would be unreasonable because the effect of the Sales jump from 2 stars to 3 stars is included.

174	687	36910	1.49	1
175	475	36985	1.5	2
	728	47459	2.49	2
	1394	42601	2.5	3

e. Execute regression discontinuity for changes from 2 to 3 stars, 3 to 4 stars, and 4 to 5 stars. What are your findings?

When I left the bandwidth at the standard STATA settings, 2-3 stars, 3-4 stars, and 4-5 stars all had PValues >.05 (not statistically significant). When I adjusted the bandwidth to .75, 2-3 stars and 4-5 stars had P-Values <.05(statistically significant), and 3-4 stars had a P-Value of > .05 (not statistically significant). Crossing from 4 stars to 5 stars (bandwidth .75) had the greatest statistically significant impact on sales with a coefficient of 2540.6.

2 stars - 3 stars (Bandwidth .75):

728	47459	2.49	2
1394	42601	2.5	3

```
. rdrobust Sales Rating, c(2.495) h(.75)

Sharp RD estimates using local polynomial regression.

  Cutoff c = 2.495 | Left of c  Right of c           Number of obs =       1483
                   |                                 BW type       =     Manual
       Number of obs |       550        933          Kernel        = Triangular
  Eff. Number of obs |       282        286          VCE method    =         NN
     Order est. (p)  |         1          1
     Order bias (q)  |         2          2
         BW est. (h) |     0.750      0.750
         BW bias (b) |     0.750      0.750
           rho (h/b) |     1.000      1.000

Outcome: Sales. Running variable: Rating.

------------------------------------------------------------------------------
      Method |     Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Conventional |    2166.6     865.62    2.5030   0.012     470.033      3863.22
      Robust |         -          -    1.4624   0.144    -667.705      4593.04
------------------------------------------------------------------------------
```

- **P-Value < .05, therefore is statistically significant.**

- **3 stars - 4 stars (Bandwidth .75):**

243	42397	3.49	3
1331	55008	3.5	4

```
. rdrobust Sales Rating, c(3.495) h(.75)

Sharp RD estimates using local polynomial regression.

  Cutoff c = 3.495 | Left of c  Right of c          Number of obs =       1483
                   |                                BW type       =     Manual
      Number of obs |       925         558         Kernel        = Triangular
Eff. Number of obs |       269         271         VCE method    =         NN
     Order est. (p) |         1           1
     Order bias (q) |         2           2
         BW est. (h) |     0.750       0.750
         BW bias (b) |     0.750       0.750
           rho (h/b) |     1.000       1.000

Outcome: Sales. Running variable: Rating.

------------------------------------------------------------------------------
      Method |     Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Conventional |     596.53     986.84    0.6045   0.546    -1337.65      2530.71
      Robust |          -          -    0.2596   0.795    -2324.49      3034.34
------------------------------------------------------------------------------
```

- **The P-Value is >.05 (95% cutoff level) and not statistically significant.**

- **4 stars – 5 stars (Bandwidth .75):**

617	49115	4.49	4
546	53036	4.5	5

```
. rdrobust Sales Rating, c(4.495) h(.75)

Sharp RD estimates using local polynomial regression.

  Cutoff c = 4.495 | Left of c  Right of c          Number of obs =       1483
                   |                                BW type       =     Manual
      Number of obs |      1282         201         Kernel        = Triangular
Eff. Number of obs |       269         201         VCE method    =         NN
     Order est. (p) |         1           1
     Order bias (q) |         2           2
         BW est. (h) |     0.750       0.750
         BW bias (b) |     0.750       0.750
           rho (h/b) |     1.000       1.000

Outcome: Sales. Running variable: Rating.

------------------------------------------------------------------------------
      Method |     Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Conventional |     2540.6     1036.8    2.4503   0.014     508.392      4572.76
      Robust |          -          -    1.4268   0.154    -806.992      5125.86
------------------------------------------------------------------------------
```

P-Value < .05, therefore is statistically significant.