UNIVERSITY OF EDINBURGH
Business School

2022-23

CM11450 Industrial Organisation

Individual Assignment

B223193

Word count: 9 pages (4531 words)

1a.

(a)

```
. reg lquant lprice north unemp rainfall , cluster(oilcompany)   // #cluster standard errors

Linear regression                               Number of obs     =     10,342
                                                F(4, 6)           =    1854.42
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1065
                                                Root MSE          =     .96748

                                  (Std. err. adjusted for 7 clusters in oilcompany)
------------------------------------------------------------------------------
             |               Robust
      lquant | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      lprice |  -10.82759   .1808012   -59.89   0.000    -11.26999   -10.38518
       north |  -.0344648   .0319435    -1.08   0.322    -.1126278    .0436982
       unemp |  -.0442028    .152875    -0.29   0.782    -.4182744    .3298687
    rainfall |  -.0001654   .0025273    -0.07   0.950    -.0063495    .0060187
       _cons |   11.94387   .1375842    86.81   0.000     11.60721    12.28052
------------------------------------------------------------------------------
```

Figure 1: OLS Results

The results in Figure 1 show the estimated coefficients, their standard errors (adjusted for clustering at the oil company level), t-statistics, p-values, and 95% confidence intervals.

- The coefficient for lprice is statistically significant and negative, which suggests that gasoline sales are inversely related to the price charged by the gas pump. The magnitude of the coefficient indicates that a 1% increase in the price charged by the gas pump is associated with a 10.83% decrease in the quantity sold, holding other variables constant. This coefficient is statistically significant at the 1% level.
- The coefficient for north is -0.0344648, which means that, on average, gas pumps located in Northern Italy sell 3.45% less gasoline than those located in other regions, holding all other variables constant. However, this coefficient is not statistically significant at the 10% level, which suggests that the effect of the location on the quantity sold may not be statistically significant.
- The coefficient for unemp is -0.0442028, which means that a 1% increase in the unemployment rate in the nearby area of a gas pump leads to a 0.04% decrease in the quantity sold by that pump, holding all other variables constant. However, this coefficient is not statistically significant at the 10% level, which suggests that the effect of the unemployment rate on the quantity sold may not be statistically significant.
- The coefficient for rainfall is -0.0001654, which means that a 1% increase in the average monthly rainfall in the area of a gas pump leads to a 0.02% decrease in the quantity sold by that pump, holding all other variables constant. However, this coefficient is not statistically significant at the 10% level, which suggests that the effect of the rainfall on the quantity sold may not be statistically significant.
- The intercept (constant) is 11.94387, which means that, when all other variables are equal to zero, the natural logarithm of the quantity sold is 11.94. This intercept is statistically significant at the 1% level.

Finally, it's important to note that the R-squared value of 0.1065 indicates that the model accounts for only 10.65% of the total variation in the natural logarithm of quantity sold. This suggests that there may be other factors that affect the quantity sold but are not captured in the model. The root mean squared error (RMSE) of 0.96748 indicates the average difference between the predicted and actual values of the natural logarithm of quantity sold.

Overall, the model has a good statistical fit, as indicated by the F-statistic and the low root mean squared error (RMSE). However, the coefficients for north, unemp, and rainfall are not statistically significant, which suggests that these variables do not have a significant effect on gasoline sales, holding other variables constant. It's worth noting

that the insignificant coefficients could be due to low statistical power or the fact that these variables are not actually related to gasoline sales. Therefore, further investigation may be necessary to determine the importance of these variables, and caution should be taken when interpreting the results and making any policy or business decisions based on them.

From the analytical perspective, we could not argue that log p_i is an endogenous regressor in the model without other sufficient information.

In practice, whether log p_i is endogenous or exogenous depends on the specific context and data. We need to carefully consider the economic theory and any potential sources of omitted variable bias or reverse causality that may affect the relationship between log p_i and qi. If log p_i is endogenous, we may need to use IV or other methods to obtain consistent estimates of its coefficient.

Endogeneity occurs when one or more explanatory variables are correlated with the error term in the regression equation. In this case, if there are other factors that affect both the average price charged and the quantity sold, then the estimated coefficient of log pi may be biased and inconsistent.

In this case, log p_i could be endogenous if there are unobserved factors that affect both the average price charged by gas pumps and the quantity sold, but are not included in the model. For example, if some gas pumps have more loyal customers than others, they may be able to charge a higher price and sell more gasoline than their competitors. Gas pumps may also charge higher prices in areas with higher demand for gasoline, which would also lead to higher sales. Another possible reason for endogeneity is reverse causality. For example, higher gasoline sales at a particular gas pump may lead to an increase in the price charged, rather than the other way around. In this case, the relationship between the price and quantity sold would be bi-directional, making log p_i endogenous and making it difficult to distinguish the true causal effect of the price on the quantity.

Therefore, in this situation, we cannot argue that log p_i is an endogenous regressor without the identification of valid instrumental variables for log p_i.

Nevertheless, from a theoretical perspective, even if we don't know the instrumental variables for log pi, we could argue that it is an endogenous regressor based on economic reasoning and prior knowledge of the market.

Endogeneity arises when a regressor is correlated with the error term in the regression equation, which violates the assumption of exogeneity. In the case of log p_i, there could be several reasons why it might be endogenous:

1. Reverse causality: Gasoline prices could be determined by the quantity sold rather than the other way around. If the gas pumps adjust their prices to clear the market, then the quantity sold could be driving the prices, rather than the other way around.
2. Omitted variable bias: There could be other factors that affect both the quantity sold and the gasoline prices, and if those factors are not included in the regression, the estimated coefficients on the included variables may be biased.
3. Measurement error: If there is a measurement error in the quantity sold or gasoline prices, this could create a correlation between the error term and the regressors.

In sum, even if we don't have a clear instrumental variable to use in our regression analysis, there may be good reasons to suspect that log p_i is endogenous, which would require us to take steps to address endogeneity in our analysis.

(b)

In addition to log p_i, the other control variables in the model (Northi, Unempi, and Raini) may also be endogenous. Endogeneity arises when a control variable is correlated with the error term in the regression equation, violating the assumption of exogeneity.

1. Northi may be endogenous if there are unobserved factors that affect both the location of gas pumps and the demand for gasoline. For example, gas pumps in the North may be located in areas with higher population density, more traffic, or more commercial activity, which could increase the demand for gasoline. If these factors are not accounted for in the

model, Northi may be correlated with the error term and biased. To address this, we may need to find suitable instruments for Northi, such as geographic or demographic characteristics of the area where the gas pump is located.

2. Unempi may also be endogenous if there are unobserved factors that affect both the unemployment rate and the demand for gasoline. For example, if areas with high unemployment also have lower income levels, this could reduce the demand for gasoline. Alternatively, if areas with high unemployment also have fewer alternative modes of transportation (such as public transit), this could increase the demand for gasoline. Again, we may need to find suitable instruments for Unempi, such as measures of economic activity or transportation infrastructure.
3. Raini may be endogenous if there are unobserved factors that affect both the amount of rainfall and the demand for gasoline. For example, if areas with higher rainfall also have more outdoor recreational activities, this could increase the demand for gasoline. Alternatively, if areas with higher rainfall also have more frequent traffic accidents or road closures, this could reduce demand for gasoline. We may need to find suitable instruments for Raini, such as measures of climate or topography.

It is important to note that the potential endogeneity of the control variables depends on the specific context and data. We need to carefully consider the economic theory and any potential sources of omitted variable bias or reverse causality that may affect the relationship between the control variables and qi. If any of the control variables are found to be endogenous, we need to find valid instrumental variables that are correlated with the endogenous variable but uncorrelated with the error term.

1b.

(a)

In order for OilQuali to be a valid instrument for log pi, it must satisfy the two instrumental variable assumptions:

1. Relevance: OilQuali must be correlated with log pi.
2. Exogeneity: OilQuali must be uncorrelated with the error term in the regression equation.

The validity of OilQuali as a possible instrument for log pi depends on whether it satisfies these assumptions. Here are some arguments to consider:

- Relevance: OilQuali may be relevant for log pi if there is a plausible causal mechanism linking the sulfur content of the crude oil input quality of an oil company and the distance of a gas pump from its nearest refinery to the price charged by the gas pump. For example, if oil companies with higher sulfur content in their crude oil input quality have to pay higher transportation costs to deliver their products to refineries farther away, they may pass on these costs to consumers in the form of higher prices. Similarly, if gas pumps located farther away from refineries have to pay higher transportation costs to receive their supplies, they may charge higher prices to cover these costs. If these mechanisms hold, then OilQuali may be a relevant instrument for log pi.
- Exogeneity: OilQuali may not be exogenous if there are other unobserved factors that affect both the sulfur content of the crude oil input quality of an oil company and the distance of a gas pump from its nearest refinery as well as the price charged by the gas pump. For example, if oil companies with lower sulfur content in their crude oil input quality also have more modern refineries that are located closer to gas pumps, then both the sulfur content of the crude oil input quality and the distance to refineries could be correlated with unobserved factors that affect gas pump prices. In this case, OilQuali would not be a valid instrument for log pi.

Overall, whether OilQuali is a valid instrument for log pi depends on the specific context and the plausibility of the causal mechanisms that link OilQuali to log pi. A thorough analysis of the data and the underlying economic factors would be needed to make a definitive judgment on the validity of this instrument.

Exogeneity is a crucial assumption for an instrument to be considered valid. Exogeneity implies that the instrument is not correlated with the error term in the equation for the endogenous variable, which means that it should not be affected by any unobservable factors that also affect the outcome variable.

In the case of OilQuali, it is constructed from two observable variables: crude oil input quality and distance from the nearest refinery. These variables are unlikely to be affected by any unobservable factors that also affect the price of gasoline. Therefore, it is reasonable to assume that OilQuali is exogenous.

However, there could be other factors that are not observed in the data, but still affect the price of gasoline and are correlated with OilQuali. For instance, local taxes, government subsidies, or market regulations could affect the price of gasoline and also vary with the quality of crude oil and distance from the refinery. In such cases, OilQuali would not be exogenous.

Moreover, if there is reverse causality between OilQuali and the outcome variable (log pi), then OilQuali would not be a valid instrument. For example, if gas stations with higher prices for gasoline choose to source higher-quality crude oil, then OilQuali would be correlated with the error term, and its exogeneity assumption would be violated.

Therefore, while OilQuali seems to be exogenous based on observable factors, further analysis is required to test its exogeneity and confirm its validity as an instrument.

(b)

```
. ivreg2 lquant (lprice=oilqual) north unemp rainfall, cluster(oilcompany)

IV (2SLS) estimation
--------------------

Estimates efficient for homoskedasticity only
Statistics robust to heteroskedasticity and clustering on oilcompany

Number of clusters (oilcompany) =      7              Number of obs =    10342
                                                      F(  4,     6) =     2.37
                                                      Prob > F      =   0.1657
Total (centered) SS     =  10829.27702                Centered R2   =   0.0057
Total (uncentered) SS   =  816451.5524                Uncentered R2 =   0.9868
Residual SS             =  10767.54261                Root MSE      =     1.02

------------------------------------------------------------------------------
             |               Robust
      lquant |  Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
      lprice |   -.2824487   .2162161    -1.31   0.191    -.7062245    .1413271
       north |   -.0337186   .0262448    -1.28   0.199    -.0851575    .0177204
       unemp |   -.0817203   .1697931    -0.48   0.630    -.4145085     .251068
    rainfall |    -.001007   .0026804    -0.38   0.707    -.0062605    .0042465
       _cons |    9.012289   .1382488    65.19   0.000     8.741326    9.283251
------------------------------------------------------------------------------
Underidentification test (Kleibergen-Paap rk LM statistic):             6.183
                                                   Chi-sq(1) P-val =    0.0129
------------------------------------------------------------------------------
Weak identification test (Cragg-Donald Wald F statistic):               8.1e+04
                         (Kleibergen-Paap rk Wald F statistic):         2.6e+05
Stock-Yogo weak ID test critical values: 10% maximal IV size             16.38
                                         15% maximal IV size              8.96
                                         20% maximal IV size              6.66
                                         25% maximal IV size              5.53
Source: Stock-Yogo (2005).  Reproduced by permission.
NB: Critical values are for Cragg-Donald F statistic and i.i.d. errors.
------------------------------------------------------------------------------
Hansen J statistic (overidentification test of all instruments):         0.000
                                                 (equation exactly identified)
------------------------------------------------------------------------------
Instrumented:         lprice
Included instruments: north unemp rainfall
Excluded instruments: oilqual
------------------------------------------------------------------------------
```

Figure 2.1: IV Results (OilQuali)

```
.  estat endog

  Tests of endogeneity
  H0: Variables are exogenous

  Robust regression F(1,6)        =  72952.6  (p = 0.0000)
    (Adjusted for 7 clusters in oilcompany)
```

Figure 2.2: Test of Endogeneity (OilQuali)

To estimate the coefficient β1 using OilQuali as an instrument for log-prices, we can use the two-stage least squares (2SLS) estimator. The first stage involves regressing the endogenous variable (log-prices) on the instrument variable (OilQuali) to obtain predicted

values of log-prices. The second stage involves regressing the dependent variable (log-quantity) on the predicted values of log-prices and the other exogenous variables.

The results of Figure 2.1 show the coefficients, standard errors, z-statistics, and p-values for the endogenous variable lprice (log-prices) and exogenous variables (north, unemp, and rainfall). The coefficient for lprice is -0.282, which means that a 1% increase in log-prices leads to a 0.282% decrease in the quantity of gas sold, holding other variables constant. However, the p-value for the coefficient is 0.191, which indicates that it is not statistically significant at the 5% level of significance.

The coefficients for north, unemp, and rainfall are -0.033, -0.082, and -0.001, respectively, and none of them are statistically significant at the 5% level of significance.

The results also provide some diagnostic tests for the validity of the instrument. The underidentification test (Kleibergen-Paap rk LM statistic) has a value of 6.183 with a p-value of 0.0129, which means that the model is not underidentified.

Weak instruments test whether the instrument has a low correlation with the endogenous explanatory variable. If it is not significant, it implies that the IV in the model might be too weak. The weak identification test (Cragg-Donald Wald F statistic) has a very high value of 8.1e+04 which is higher than any value of the Stock-Yogo test, then we can reject that our instrument is weak. However, the Hansen J statistic, which tests for overidentification of all instruments, has a value of 0.000, suggesting that the model is exactly identified.

Here in Figure 2.2, we are testing whether using OLS is appropriate. The null hypothesis of Sargan-Hansen test is that the instrument (OilQuali) is exogenous, meaning it is not correlated with the error term in the main equation. The output shows that the p-value for the robust regression F-test of endogeneity is 0.0000, which is less than the significance level of 0.05. Therefore, we reject the null hypothesis and conclude that there is no evidence of endogeneity and OLS is not an appropriate way to go. This suggests that there may be unobserved factors that affect both the instrument and the endogenous variable, which could bias the results of the 2SLS estimation. Further diagnostics or alternative instruments may be needed to address this issue. Note that the results of the Sargan-Hansen test or the 2SLS estimation do not prove endogeneity definitively, but they provide evidence for or against it.

Overall, the model has a weak instrument, and the coefficients for the variables are not statistically significant at the 5% level. Therefore, the model does not provide strong evidence for the effect of log-prices on the quantity of gas sold.

(c)

```
. reg lquant lprice north unemp rainfall , cluster(oilcompany)

Linear regression                               Number of obs     =     10,342
                                                F(4, 6)           =    1854.42
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1065
                                                Root MSE          =     .96748

                            (Std. err. adjusted for 7 clusters in oilcompany)
------------------------------------------------------------------------------
             |               Robust
      lquant |   Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      lprice |   -10.82759   .1808012   -59.89   0.000    -11.26999   -10.38518
       north |   -.0344648   .0319435    -1.08   0.322    -.1126278    .0436982
       unemp |   -.0442028    .152875    -0.29   0.782    -.4182744    .3298687
    rainfall |   -.0001654   .0025273    -0.07   0.950    -.0063495    .0060187
       _cons |    11.94387   .1375842    86.81   0.000     11.60721    12.28052
------------------------------------------------------------------------------

. ivregress 2sls lquant (lprice=oilqual) north unemp rainfall , cluster(oilcompany)

Instrumental variables 2SLS regression          Number of obs     =     10,342
                                                Wald chi2(4)      =     11.04
                                                Prob > chi2       =     0.0261
                                                R-squared         =     0.0057
                                                Root MSE          =     1.0204

                            (Std. err. adjusted for 7 clusters in oilcompany)
------------------------------------------------------------------------------
             |               Robust
      lquant |   Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
      lprice |   -.2824487   .2162161    -1.31   0.191    -.7062245    .1413271
       north |   -.0337186   .0262448    -1.28   0.199    -.0851575    .0177204
       unemp |   -.0817203   .1697931    -0.48   0.630    -.4145085     .251068
    rainfall |    -.001007   .0026804    -0.38   0.707    -.0062605    .0042465
       _cons |    9.012289   .1382488    65.19   0.000     8.741326    9.283251
------------------------------------------------------------------------------
Instrumented: lprice
 Instruments: north unemp rainfall oilqual
```

Figure 3: Comparison of IV and OLS Results (OilQuali)

Comparing the OLS and IV results, we can see in Figure 3 that the coefficient estimate for lprice is -10.82759 in the OLS regression, while it is -0.2824487 in the IV regression. This is a substantial difference, suggesting that the endogeneity of lprice may be biasing the OLS estimate. The coefficient estimates for the control variables (north, unemp, and rainfall) are similar in both regressions.

The OLS results show that only the coefficient for lprice is statistically significant at a 1% level, while the coefficients for north, unemp, and rainfall are not statistically significant. The adjusted R-squared value is 0.1065, indicating that the model explains only a small proportion of the variance in lquant.

The 2SLS results, on the other hand, show that the coefficient for lprice is not statistically significant at a 5% level, suggesting that the OLS estimate may be biased due to endogeneity. The coefficients for north, unemp, and rainfall are also not statistically significant. The R-squared value is lower at 0.0057, indicating that the IV model explains even less of the variance in lquant than the OLS model.

The test for endogeneity (estat endog) suggests that the null hypothesis of exogeneity of the instruments and the endogeneity of the explanatory variable can be rejected at a very low p-value of 0.0000, supporting the use of the 2SLS model.

Overall, the 2SLS model results provide a better estimate of the causal effect of lprice on lquant, taking into account the potential endogeneity issue. However, the weak instrument and the low R-squared value suggest that there may still be other unobserved factors that influence the relationship between lprice and lquant.

(d)

To assess whether lp dieseli is a valid instrument for log pi, we need to check whether it satisfies the two instrumental variable assumptions:

1. Relevance: lp dieseli must be correlated with log pi.
2. Exogeneity: lp dieseli must be uncorrelated with the error term in the regression equation.

Here are some arguments to consider:

- Relevance: lp dieseli may be relevant for log pi if there is a plausible causal mechanism linking diesel fuel prices to gasoline prices. For example, if diesel fuel and gasoline are substitutes, an increase in diesel fuel prices could lead consumers to switch to gasoline, causing an increase in gasoline prices. In this case, lp dieseli would be correlated with log pi, and would be a relevant instrument.
- Exogeneity: lp dieseli may not be exogenous if there are other unobserved factors that affect both diesel and gasoline prices as well as the quantity of gasoline sold. For example, if there are shocks to the global oil market that affect both diesel and gasoline prices, lp dieseli would be correlated with unobserved factors that affect gasoline prices. In this case, lp dieseli would not be a valid instrument for log pi.

Overall, whether lp dieseli is a valid instrument for log pi depends on the specific context and the plausibility of the causal mechanisms that link diesel fuel prices to gasoline prices. It is possible that lp dieseli satisfies the instrumental variable assumptions, but a thorough analysis of the data and the underlying economic factors would be needed to make a definitive judgment on the validity of this instrument.

```
. ivreg2 lquant (lprice=lp_diesel) north unemp rainfall, cluster(oilcompany)

IV (2SLS) estimation
--------------------

Estimates efficient for homoskedasticity only
Statistics robust to heteroskedasticity and clustering on oilcompany

Number of clusters (oilcompany) =      7              Number of obs =     10342
                                                      F(  4,     6) =      1.06
                                                      Prob > F      =    0.4499
Total (centered) SS     =  10829.27702                Centered R2   =    0.0882
Total (uncentered) SS   =  816451.5524                Uncentered R2 =    0.9879
Residual SS             =  9874.159776                Root MSE      =     .9771

------------------------------------------------------------------------------
             |               Robust
      lquant |  Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
      lprice |  -6.330417   3.377649    -1.87   0.061    -12.95049     .2896526
       north |  -.0341466   .0281753    -1.21   0.226    -.0893692     .021076
       unemp |  -.0602028   .1476413    -0.41   0.683    -.3495744     .2291687
    rainfall |  -.0005243   .0024937    -0.21   0.833    -.0054119     .0043633
       _cons |   10.69364   .9544664    11.20   0.000     8.822921     12.56436
------------------------------------------------------------------------------
Underidentification test (Kleibergen-Paap rk LM statistic):                 5.995
                                                      Chi-sq(1) P-val =     0.0143
------------------------------------------------------------------------------
Weak identification test (Cragg-Donald Wald F statistic):                  82.493
                         (Kleibergen-Paap rk Wald F statistic):           118.736
Stock-Yogo weak ID test critical values: 10% maximal IV size               16.38
                                         15% maximal IV size                8.96
                                         20% maximal IV size                6.66
                                         25% maximal IV size                5.53
Source: Stock-Yogo (2005).  Reproduced by permission.
NB: Critical values are for Cragg-Donald F statistic and i.i.d. errors.
------------------------------------------------------------------------------
Hansen J statistic (overidentification test of all instruments):           0.000
                                                      (equation exactly identified)
------------------------------------------------------------------------------
Instrumented:         lprice
Included instruments: north unemp rainfall
Excluded instruments: lp_diesel
------------------------------------------------------------------------------
```

Figure 4.1: IV Results (lp dieseli)

```
.  estat endog

  Tests of endogeneity
  H0: Variables are exogenous

  Robust regression F(1,6)        =  1.73659  (p = 0.2356)
    (Adjusted for 7 clusters in oilcompany)
```

Figure 4.2: Test of Endogeneity (lp dieseli)

As shown in Figure 4.1, the coefficient of interest is the effect of lprice on lquant, which is estimated to be -6.3304 with a standard error of 3.3776. This implies that a one-unit increase in the log-price of fuel would lead to a 6.33-unit decrease in the log of quantity sold. However, the coefficient is not statistically significant at conventional levels of significance (p-value of 0.061).

The control variables do not have statistically significant effects on lquant either, as none of their coefficients are statistically significant at conventional levels.

The underidentification test suggests that the model may be underidentified, indicating that there might not be enough exogenous variation in the instrumental variable to estimate the causal effect of interest. The weak identification test also suggests that the instrumental variable may be weak, which can lead to biased and inconsistent estimates.

The Hansen J statistic indicates that the model is exactly identified, meaning that the number of instruments used is sufficient to estimate the causal effect of interest.

Here in Figure 4.2, we are testing whether using OLS is appropriate, hence lprice is actually not exogenous and we don't really need the IV approach. The null hypothesis is that the instrumented variables are exogenous. The robust regression F-test has a p-value of 0.2356, which fails to reject the null hypothesis at conventional significance levels. Therefore, lprice is exogenous and OLS is an appropriate way to go.

Wu-Hausman tests that IV is just as consistent as OLS, and since OLS is more efficient, it would be preferable. The null here is that they are equally consistent; Therefore, if it is not significant, it implies that we should adopt the OLS result rather than the IV result.

A good IV regression normally has *Weak instruments test* significant, *Wu-Hausman test* significant and *Sargan test* insignificant. Both instruments have

significant results for weak instruments test and *Sargan test,* whereas only *Wu-Hausman test* of oilqual is significant while that of lp_diesel is insignificant which suggests that OLS is a more appropriate way for lprice to implement if choosing lp_diesel as the instrument variable.

In conclusion, oilqual should be chosen as the instrumental variable to implement 2SLS regression to resolve the endogenous issue.

2a.

There are several concerns that could cause a biased estimation if we run a simple regression to estimate the effect of flooding on housing prices. These include:

1. Selection bias: The properties located in flood-prone areas might be different from those that are not. For example, houses located in flood-prone areas may be cheaper or smaller than those that are not. Therefore, if we only include properties in flood-prone areas, we may be comparing houses that are inherently different. To avoid this, we need to include non-flood-prone properties as a control group.
2. Endogeneity: There may be other factors that affect both the likelihood of flooding and housing prices, such as location, proximity to amenities, and socio-economic characteristics. These factors may lead to endogeneity, making it difficult to identify the causal effect of flooding on housing prices.
3. Omitted variable bias: There may be other factors that affect housing prices that are not included in the regression model. For example, the quality of the construction or materials used to build the house might affect the level of damage caused by flooding. If we do not control for these factors, we may overestimate or underestimate the effect of flooding on housing prices.

To address these concerns, we can use a difference-in-differences (DiD) regression model. The model would compare the changes in housing prices of properties located in flood-prone areas before and after the flood event with the changes in housing prices of properties located in non-flood-prone areas over the same period. This approach helps to control for selection bias, endogeneity, and omitted variable bias.

The DiD model can be expressed as follows:

$$\triangle Y\_it = \alpha + \beta Flood\_i + \gamma(Flood\_i \times Post\_t) + \delta Post\_t + \varepsilon\_it$$

Where ΔY_it is the change in the log-housing price of property i at time t, $Flood\_i$ is a dummy variable indicating whether the property is located in a flood-prone area, $Post\_t$ is a dummy variable indicating whether the observation is before or after the flood event, and $Post\_t \times Flood\_i$ is the interaction term. α is the intercept, β is the coefficient of the flood dummy variable, γ is the coefficient of the interaction term, δ is the coefficient of the time dummy variable, and ε_it is the error term.

The coefficient of the interaction term (γ) captures the differential effect of the flood event on the housing prices of properties located in flood-prone areas compared to those located in non-flood-prone areas. If γ is positive and statistically significant, it implies that the flood event had a negative effect on the housing prices of properties located in flood-prone areas.

2b.

If the government wants to implement a policy against flood disasters, we can use a natural experiment approach to estimate the causal effect of floods on housing prices. Specifically, we can use the flood event as a source of exogenous variation that affects some properties but not others.

In this case, a suitable study method would be a regression discontinuity design (RDD) since floods may have a discrete impact on some properties, for example, those located just on the border of the flood-prone area. RDD would allow us to estimate the causal effect of floods on housing prices by comparing the properties located just inside and outside of the flood-prone area.

The RDD model can be expressed as follows:

$$Y\_i = \alpha + \beta(X\_i - X\_0) + \varepsilon\_i$$

Where Y_i is the log-housing price of property i, X_i is the distance of property i from the flood-prone area, X_0 is the cutoff point where the flood-prone area starts, and ε_i is the error term. α is the intercept, and β is the coefficient of the distance variable.

The coefficient of the distance variable (β) captures the effect of being located just inside the flood-prone area compared to being located just outside the flood-prone area. If β is negative and statistically significant, it implies that the flood event had a negative effect on the housing prices of properties located just inside the flood-prone area.

However, there are some limitations to RDD. First, the design assumes that the assignment to treatment (being located in the flood-prone area) is determined solely by the distance from the cutoff point, which may not be entirely accurate. Second, the design assumes that there is no spillover effect from the treatment group to the control group or vice versa, which may not be valid in all cases. Finally, the sample size might be limited, particularly if the cutoff point is narrow. Therefore, it is important to perform a sensitivity analysis and assess the robustness of the results.

2c.

If the government wants to investigate the effect of flooding on different property types, a suitable study method would be a difference-in-differences (DiD) approach. DiD can help us estimate the causal effect of flooding on housing prices of different property types by comparing the change in housing prices of properties affected by flooding (the treatment group) to the change in housing prices of similar properties that were not affected by flooding (the control group).

To proceed with the analysis, we first need to define the treatment and control groups. We can identify the treatment group as properties located in flood-prone areas that were affected by flooding, and the control group as similar properties that were not affected by flooding but located in the same area. We can use matching techniques to create the control group, ensuring that the control group properties are similar in terms of their pre-treatment characteristics to the treatment group properties.

The DiD model can be expressed as follows:

$$Y\_it = \alpha + \beta Treatment\_i + \beta PosTreatment\_i + \beta(Treatment\_i \times PosTreatment\_i) + \varepsilon\_it$$

Where Y_it is the log-housing price of property i at time t, $Treatment\_i$ is a dummy variable indicating whether the property i is in the treatment group (1 if affected by flooding, 0 otherwise), $PosTreatment\_i$ is a dummy variable indicating the property type of i (1 for the type of interest, 0 otherwise), and ε_it is the error term. α is the intercept, $\beta Treatment$ is the effect of flooding on the housing prices of the treatment group, $\beta PosTreatment$ is the effect of property type on housing prices, and β is the difference-in-differences coefficient, which captures the differential effect of flooding on the housing prices of the property type of interest compared to other property types.

By estimating the coefficient β, we can determine the effect of flooding on the housing prices of the property type of interest. It is important to include control variables such as location, size, and other relevant property characteristics to account for differences between the treatment and control groups.

Limitations of DiD include the possibility of unobserved time-varying confounders that could bias the results, as well as the possibility of spill-over effects to adjacent properties that were not directly affected by flooding. Therefore, it is important to perform sensitivity analyses to assess the robustness of the results and to investigate the potential spillover effects.

In summary, a DiD approach can be a useful method to investigate the effect of flooding on the housing prices of different property types. By carefully defining the treatment and control groups and including relevant control variables, we can obtain reliable estimates of the causal effect of flooding on housing prices.