LPM Regression Analysis

John Morris

2025-10-09

Contents

#first thing I have done is load a saved workspace environment that I created at the end of my last session.

```r
library(ggplot2)
library(dplyr)
library(stargazer)
```

```r
cd<-read.csv("~/Pitt/MQE/2024/capstone/cd.csv")
```

#double check column names

```r
colnames(cd)
```

```
##  [1] "X"                     "P_ID"                  "AR_ID"
##  [4] "YEAR_ID"               "PRINCIPAL"             "PAID"
##  [7] "DISCOUNTED"            "BASE_VALUE"            "BASE_EXEMPT_VALUE"
## [10] "NET_BASE_VALUE"        "RATE_VALUE"            "VA_LAND_VALUE"
## [13] "VA_IMPROVEMENT"        "TOTAL_TAXES"           "TRUE_TAX"
## [16] "CLASS"                 "ROLL_SECTION"          "DELINQUENT"
## [19] "MUNICIPALITY_CODE"     "SCHOOL_ZONE"           "TAX_MAP_UFMT"
## [22] "LONG_DESC"             "MUNICIPALITY_MILLAGE" "SCHOOL_NAME"
## [25] "SCHOOL_MILLAGE"        "LAG1"                  "TOTAL_MILLAGE"
```

Ok lets start some basic analysis. I want to observe the relationship between net base value and and delinquency. I can see this going either way. If the property has a higher value, property owners would pay more tax. if the tax is higher the incidence of delinquency could be higher. Conversely, if the property value is high, this could indicate that property owners have higher income. If property owners have higher income they can more readily pay their property tax.

I will using a Linear Probability Model since my dependent variable (delinquent) is binary.

```r
m1 <- lm(DELINQUENT ~  NET_BASE_VALUE, data=cd)
summary(m1)
```

```
##
## Call:
## lm(formula = DELINQUENT ~ NET_BASE_VALUE, data = cd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.09851 -0.09793 -0.09736 -0.09625  2.70049
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     9.851e-02  1.240e-04  794.25   <2e-16 ***
## NET_BASE_VALUE -1.138e-08  1.288e-10  -88.36   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2955 on 5813761 degrees of freedom
##   (370851 observations deleted due to missingness)
## Multiple R-squared:  0.001341,   Adjusted R-squared:  0.001341
## F-statistic:  7808 on 1 and 5813761 DF,  p-value: < 2.2e-16
```

Model 1 suggests that For ever dollar increase in net base value the probability of delinquency decreases by 0.000000011748. Lets transform net base value for better interpretability.

```r
cd <- cd %>% mutate(NBV_10k = NET_BASE_VALUE/10000 )
```

#now Net Base Value can be interpreted as a $10,000 change

#redo LPM

```r
m1 <- lm(DELINQUENT ~ NBV_10k, data = cd)
summary(m1)
```

```
##
## Call:
## lm(formula = DELINQUENT ~ NBV_10k, data = cd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.09851 -0.09793 -0.09736 -0.09625  2.70049
##
```

```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.851e-02  1.240e-04  794.25   <2e-16 ***
## NBV_10k     -1.138e-04  1.288e-06  -88.36   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2955 on 5813761 degrees of freedom
##   (370851 observations deleted due to missingness)
## Multiple R-squared:  0.001341,   Adjusted R-squared:  0.001341
## F-statistic:  7808 on 1 and 5813761 DF,  p-value: < 2.2e-16
```

Holding all else constant, a $10,000 increase in net base value decrease the probability of delinquency by 0.00011748. #Alternatively we could say a $10,000 increase in net base value decrease the probability of delinquency by 0.011748 percentage points.

The statistically significant negative sign on the coefficient is likely related to income. However, we do not have income data. If we believe that more wealthy individuals buy more expensive parcels of land, then Net base value can serve as a proxy for income.

I ought to look at total_taxes and control for net base value. As taxes increase we should expect delinquency to increase. However, as we have seen, delinquency decreases as net base value increases. higher Net base value means higher total tax.

By controlling for the correlation of net base value and total taxes we can get a more accurate account of whats happening

```r
cd <- cd %>% mutate(TOTAL_TAX100 = TOTAL_TAXES/100)
m2 <- lm(DELINQUENT ~ TOTAL_TAX100, data = cd)
m3 <- lm(DELINQUENT ~ NBV_10k + TOTAL_TAX100, data = cd)
stargazer::stargazer(m1, m2, m3, type="text", header = TRUE, omit.stat= c("f", "ser"))
```

```
##
## ============================================
##                     Dependent variable:
##                ------------------------------
##                         DELINQUENT
##                   (1)        (2)        (3)
## --------------------------------------------
## NBV_10k       -0.0001***             -0.020***
##               (0.00000)              (0.0002)
##
## TOTAL_TAX100            -0.0002*** 0.042***
##                          (0.00000)  (0.0004)
##
## Constant        0.099***   0.098***  0.101***
```

```
##                (0.0001)   (0.0001)  (0.0001)
##
## ---------------------------------------------
## Observations 5,813,763  5,813,842  5,813,750
## R2             0.001      0.001      0.003
## Adjusted R2    0.001      0.001      0.003
## =============================================
## Note:             *p<0.1; **p<0.05; ***p<0.01
```

a $100 increase in total taxes decrease the likelihood of delinquency by .02 percentage point. This is a biased estimator since total tax is a function of net base value.

the combined regression is a much more clear description. Here we see that net base value (income proxy) is negatively correlated with delinquency and, total tax is now positively correlated with delinquecny.

I'd like to mess with the lag variable I created earlier as well. I suspect it will be a strong predictor.

I'll compare multiple models where I control for net base value and total taxes as well.

```r
m4 <- lm(DELINQUENT ~ LAG1, data = cd)
m5 <- lm(DELINQUENT ~ LAG1 + NBV_10k, data=cd)
m6 <- lm(DELINQUENT ~ LAG1 + NBV_10k + TOTAL_TAXES, data=cd)
stargazer::stargazer(m4,m5,m6, type="text", header = TRUE, omit.stat= c("f", "ser"))
```

```
##
## =============================================
##                    Dependent variable:
##              --------------------------------
##                        DELINQUENT
##                 (1)        (2)        (3)
## ---------------------------------------------
## LAG1         0.793***   0.797***   0.797***
##              (0.0003)   (0.0003)   (0.0003)
##
## NBV_10k                -0.00003*** -0.005***
##                         (0.00000)  (0.0001)
##
## TOTAL_TAXES                        0.0001***
##                                    (0.00000)
##
## Constant     0.026***   0.028***   0.029***
##              (0.0001)   (0.0001)   (0.0001)
##
## ---------------------------------------------
```

```
## Observations 6,184,614  5,813,763  5,813,750
## R2              0.575      0.579      0.579
## Adjusted R2    0.575      0.579      0.579
## ===============================================
## Note:              *p<0.1; **p<0.05; ***p<0.01
```

the delinquency lag indicator is an excellent predictor of deliquency. all predictors are statistically significant but LAG1 has a much greater magnitude. If a parcel is delinquent last year the probability of the parcel being delinquent this year is 0.795!

I wanna look more into this. What is the probability of being delinquent today if a parcel was delinquent two years ago, three years ago, …, n years ago? Can we see if a class or group of parcels is persistently delinquent?