UCI Paul Merage School of Business

Leadership for a Digitally Driven World™

MFIN 290: Financial Econometrics

Lecture 7-1

Last Time

- Introduction to MLE – Intuition, Normal Distribution example
- Wald, LR, Lagrange multiplier tests - Intuition
- Pseudo R-squared and interpretation (and what it isn't!)

Binary Outcomes

- One place where MLE methods often appear is with binary dependent data:
- "Did the Company Acquire the target?"
- "Did the firm default on their debt?"

Binary Outcomes

- Here, we have a situation where our dependent variable has the form:
- $y_i = \begin{cases} 1 & \text{if the i} - \text{th observation is true} \\ 0 & \text{otherwise} \end{cases}$

- We can think of each y_i as the realization of a random variable that can take values 1 with probability π_i and 0 with probability $(1 - \pi_i)$

Binary Outcomes

- Here, we have a situation where our dependent variable has the form:
- $y_i = \begin{cases} 1 \text{ if the i} - \text{th observation is true} \\ 0 \qquad \text{otherwise} \end{cases}$

- We can think of each y_i as the realization of a random variable Y_i that can take values 1 with probability π_i and 0 with probability $(1 - \pi_i)$

- $\Pr(Y_i = y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$

Binary Outcomes

- $\Pr(Y_i = y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$
- $if\ y_i = 1 => \pi_i^1(1-\pi_i)^{1-1} = \pi_i$

$if\ y_i = 0 => \pi_i^0(1-\pi_i)^{1-0} = 1-\pi_i$

Binary Outcomes

- $\Pr(Y_i = y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$

- $\Pr(Y_i = 1) = \pi_i$
- $\Pr(Y_1 = 0) = 1 - \pi_i$

- $E(Y_i) = \pi_i$
- $Var(Y_i) = \pi_i(1-\pi_i)$

Binary Outcomes

- $\Pr(Y_i = y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$

- *What is the probability I get n* $1's$ *and m zeros for* y_i *if they are all independent?*

Linear Probability Model

- We want our probabilities π_i to be a function of covariates, something like
- $\pi_i = x_i\beta$
- As before. This is called a linear probability model and we can estimate this through OLS.
- We did this in our schooling example when we regressed the class size dummy on covariate controls.

Linear Probability Model

- But π_i we know has to be between zero and one, and in general, $x_i\beta$ does not .
- In some (but not all!) contexts this is an issue.
- One possibility is to transform the independent variables so that they can range from $-\infty$ to ∞ but output a value that ranges from zero to one…

Odds and Log Odds

- First, we can calculate the odds of the event:
- $\frac{\pi_i}{(1-\pi_i)}$ = ratio of probability of occurrence to probability of non-occurence.

- This ranges from $[0, \infty)$

- Sometimes it's more sensible to talk about odds (gambling payoffs for example)…

Odds and Log Odds

- Then we take the (natural) logarithm:

- $\log\left(\frac{\pi_i}{1-\pi_i}\right) = logit(\pi_i)$

- As the odds go to zero, the logit goes to $-\infty$
- As the odds go to ∞, the logit goes to ∞

Odds and Log Odds

- This is obviously monotone and invertible… if we let $z_i = logit(\pi_i)$:

- $z_i = \log\left(\frac{\pi_i}{1-\pi_i}\right)$

- $\pi_i = \frac{\exp(z_i)}{(1+\exp(z_i))}$

Odds and Log Odds

- Example:
- $\pi_i = \frac{1}{3}$
- $\text{Odds} = \frac{\frac{1}{3}}{\frac{2}{3}} = ½$
- $\text{Logit} = \ln\left(\frac{1}{2}\right) = -0.69$
- $\frac{\exp(-0.69)}{1+\exp(-0.69)} = \frac{1}{3}$

Logistic Regression Model

- In this setting, we could model the logit of the probability as a linear function of the predictors:

- $logit\left(\frac{\pi_i}{(1-\pi_i)}\right) = x_i\beta$

- $\Rightarrow \pi_i = \frac{\exp(x_i\beta)}{1+\exp(x_i\beta)}$

Interpreting the Coefficients

- $\Rightarrow \pi_i = \frac{\exp(x_i\beta)}{1+\exp(x_i\beta)}$
- Can differentiate (quotient rule!) to recover:
- $\frac{\partial \pi_i}{\partial x_{ij}} = \beta_j \pi_i (1 - \pi_i)$
- The effect of changing the x_i on the probability is a function of the level of the probability itself!

Interpreting the Coefficients

- $\frac{\partial \pi_i}{\partial x_{ij}} = \beta_j \pi_i (1 - \pi_i)$

- The effect of changing the x_i on the probability is a function of the level of the probability itself!

- When these marginal effects are reported, where they are reported in the distribution matters. Often, software will default to the mean of x; but this may not make sense (especially if the independent variable is a categorical variable as well).

Logits via Maximum Likelihood

- Imagine that we have a binary variable y whose behavior is governed by some latent variable y^* such that

- $y = 1$ if $y^* = x\beta + \epsilon \geq 0$ and zero otherwise

- Note that this is scale independent: we don't identify the variation in y^*, only its sign. What do we mean by this?

Logits via Maximum Likelihood

- Imagine that we have a binary variable y whose behavior is governed by some latent variable y^* such that

- $y = 1$ if $y^* = x\beta + \epsilon \geq 0$ and zero otherwise

- If we have a new error term, e, with variance one:

$$y^* = x\beta + \sigma e$$

- $\frac{y^*}{\sigma} = x\frac{\beta}{\sigma} + e$ is still positive with exactly the same x, e combinations.
- $\Rightarrow$ σ cannot be estimated here!

Binary Outcomes via Maximum Likelihood

- $y = 1$ if $y^* = x\beta + \epsilon \geq 0$ and zero otherwise

- $\Pr(\ y^* > 0|x) = \Pr(\epsilon > -x\beta|x)$
- If the distribution of the errors is symmetric, we have:
- $\Pr(\epsilon > -x\beta|x) = \Pr(\epsilon < x\beta|x) = F(x\beta)$
- Where F( ) is the CDF of the residuals.

Probits via Maximum Likelihood

- $\Pr(y = 1|x) = F(x\beta)$
- $\Pr(y = 0|x) = 1 - F(x\beta)$

- Note: no stars – this is the binary variable y

-

Probits via Maximum Likelihood

- A natural choice for the distribution of the ϵ here would be normal. This yields the "Probit" model.

- If $\phi(x)$ is the pdf of the normal distribution, and $\boldsymbol{\Phi}(x)$the cdf, then
- $\Pr(y = 1|x) = \boldsymbol{\Phi}(x)$

-

Probits via Maximum Likelihood

- $\Pr(y = 1|x) = \boldsymbol{\Phi}(x)$
- $\Pr(y = 1\ k\ times\ out\ of\ n\ observations\ |x) = \boldsymbol{\Phi}(x)^k\left(1 - \boldsymbol{\Phi}(x)\right)^{n-k}$
- Maximizing this involves some arduous derivatives (but is certainly possible!).
-

Logits via Maximum Likelihood

- Different distributional assumptions will make for a better introductory example, no matter how much we love the normal distribution in practice.
- If instead, we assume that the errors follow a logit distribution, we recover/recall the following expression for π_i:
- $\Pr(y = 1|x) = \frac{\exp(x'\beta)}{1+\exp(x'\beta)}$

- This is often abbreviated as $\boldsymbol{\Lambda}(t)$ (it was π before). We know:
- $\Pr(y = 1 \text{ k times out of n observations } |x) = \boldsymbol{\Lambda}(t)^k\left(1 - \boldsymbol{\Lambda}(t)\right)^{n-k}$ and have already mentioned the pdf:
- $\frac{\partial \pi_i}{\partial(x'\beta)} = \pi_i(1 - \pi_i) \Rightarrow \frac{\partial \boldsymbol{\Lambda}(t)}{\partial t} = \boldsymbol{\Lambda}(t)\ (1- \boldsymbol{\Lambda}(t)\ )$

Differentiating

- $\frac{\partial F(y|x)}{\partial x} = \mathbf{\Lambda}(x'\beta)[1 - \mathbf{\Lambda}(x'\beta)]\beta$

- This is one of the reasons the logit is particularly popular: if we know the inverse log odds, we know the derivative basically immediately… though it varies with x in general (the probit does too!)

Aside: Logit vs. Probit

- In practice, the derivatives/partial effects of these models end up being quite close most of the time. They can vary in the tails (as this is where the normal and logit vary the most), but seeing materially different results is unusual.

- The RESET test is used to get at the appropriateness of this distributional assumption (called the "link" function).

- The similarity in outputs, combined with computational ease contribute to the logit's popularity.

Partial Effects/Interpreting Coefficients

- When the derivative varies at each X, we have to do some work to calculate the net effect. One option is to present the derivative at the mean value of X. This is called the "partial effect at the average":

- $PEA = f(\bar{x}'\hat{\beta})\hat{\beta}$

- Where $\hat{\beta}$ is the coefficient estimated through the MLE procedure.
- This may not always be interesting or helpful (consider a dummy variable: the mean will occur with probability zero!)

Partial Effects/Interpreting Coefficients

- $PEA = f(\bar{x}'\hat{\beta})\hat{\beta}$

- Contrast this with the "Average Partial Effect" which is the average partial effect calculated across all of the X's in the sample:

- $APE = \frac{1}{n}\sum f(x_i\hat{\beta})\hat{\beta}$

- This is the sample analog to the expected value of the partial effect, and *usually* what we care about.

Partial Effects/Interpreting Coefficients

- $PEA = f(\bar{x}'\hat{\beta})\hat{\beta}$

- $APE = \frac{1}{n}\sum f(x_i\hat{\beta})\hat{\beta}$

- Can easily calculate either of these "by hand" and be sure of what you are looking at. …If you remember this is an issue!

Dummy Variables

With categorical variables, the right sensitivity might be the change in the probability as I move between categories (i.e., effect on defaults by provenance), or it might actually be the sensitivity for that category (sensitivity of defaults for home owners in Las Vegas).

These are two different things.

Have to be comfortable either reading help files to figure out exactly what is being reported, or just calculating the information you need by hand.

Logits and Odds Ratios

- There is an immediate interpretation to the logit coefficients, however.
- Take the odds ratio: $\frac{\Pr(y=1|x)}{\Pr(y=0|x)} = \frac{\frac{\exp(x'\beta)}{1+\exp(x'\beta)}}{\frac{1}{1+\exp(x'\beta)}} = \exp(x'\beta)$
- The change in the odds ratio when x changes is given by $\beta \exp(x'\beta)$
- => The coefficient is the multiplicative change in the odds! This should make sense somewhat since the logit started as the log odds ratio…

Stata Example: LV Defaults, marginal effects


```
. *1000 mortgages on single family homes in Las Vegas in 2008
.
. summ lvr ref insur rate amount credit term arm delinquent
```

Variable	Obs	Mean	Std. Dev.	Min	Max
lvr	1,000	70.5893	23.30734	5	99.7
ref	1,000	.576	.4944375	0	1
insur	1,000	.72	.4492236	0	1
rate	1,000	7.939915	1.721108	4.75	14.599
amount	1,000	1.757963	1.012117	.0735	8.6665
credit	1,000	622.069	59.52792	445	809
term	1,000	28.125	4.825211	10	30
arm	1,000	.716	.4511624	0	1
delinquent	1,000	.199	.3994478	0	1

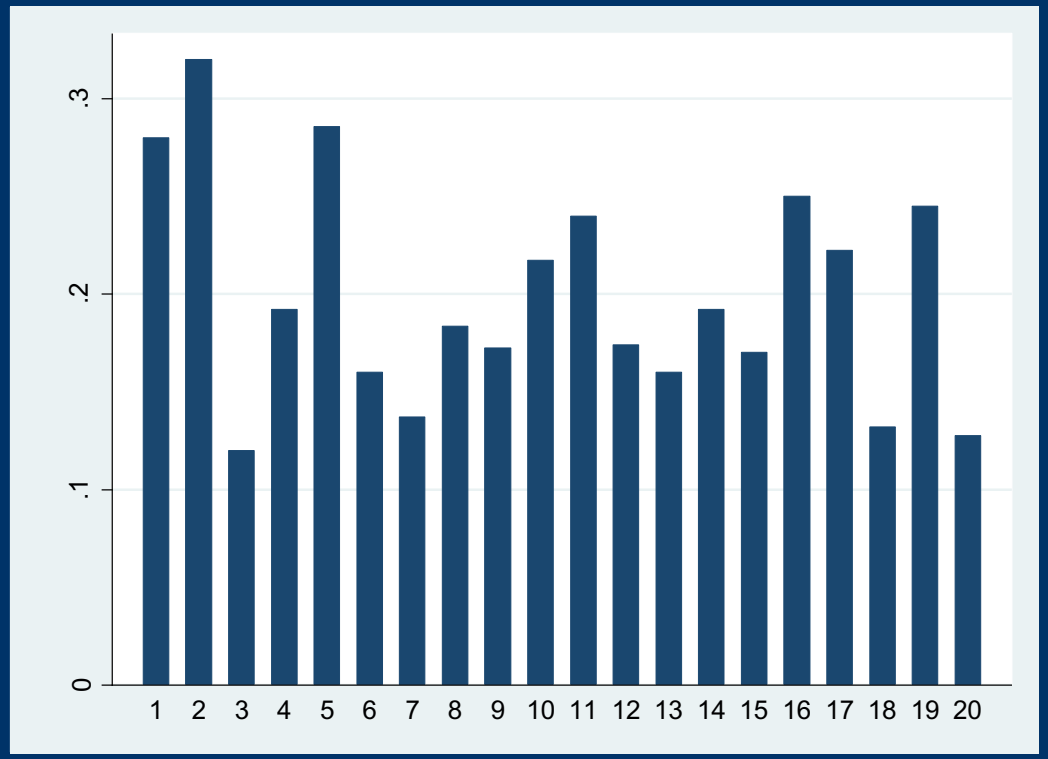
Credit Quality?

```
. xtile credit_bins = credit, nquantiles(20)

. table credit_bins, c(mean delinquent)
```

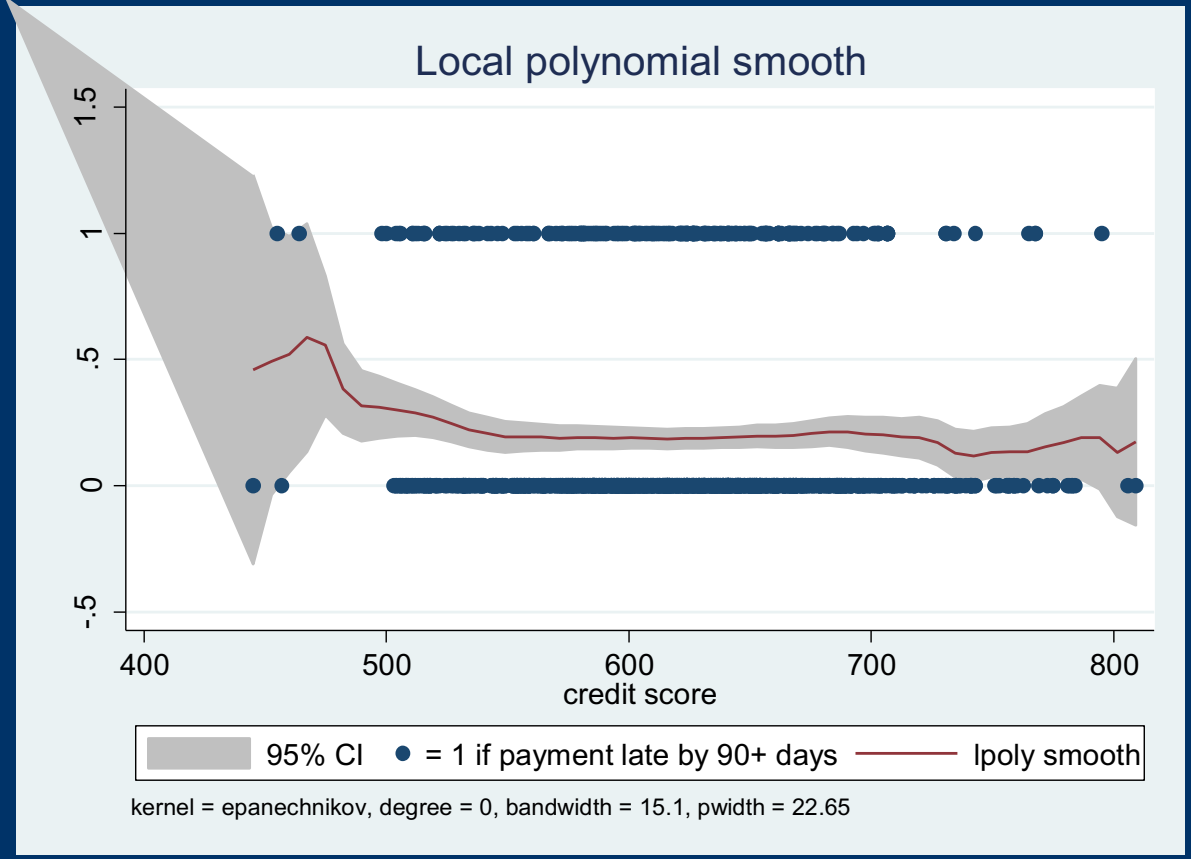
20 quantiles of credit	mean(delinq~t)
1	.28
2	.32
3	.12
4	.192308
5	.285714
6	.16
7	.137255
8	.183673
9	.172414
10	.217391
11	.24
12	.173913
13	.16
14	.192308
15	.170213
16	.25
17	.222222
18	.132075
19	.244898
20	.12766

graph bar delinquent, over(credit_bins)



Credit Quality?

lpoly delinquent credit, ci

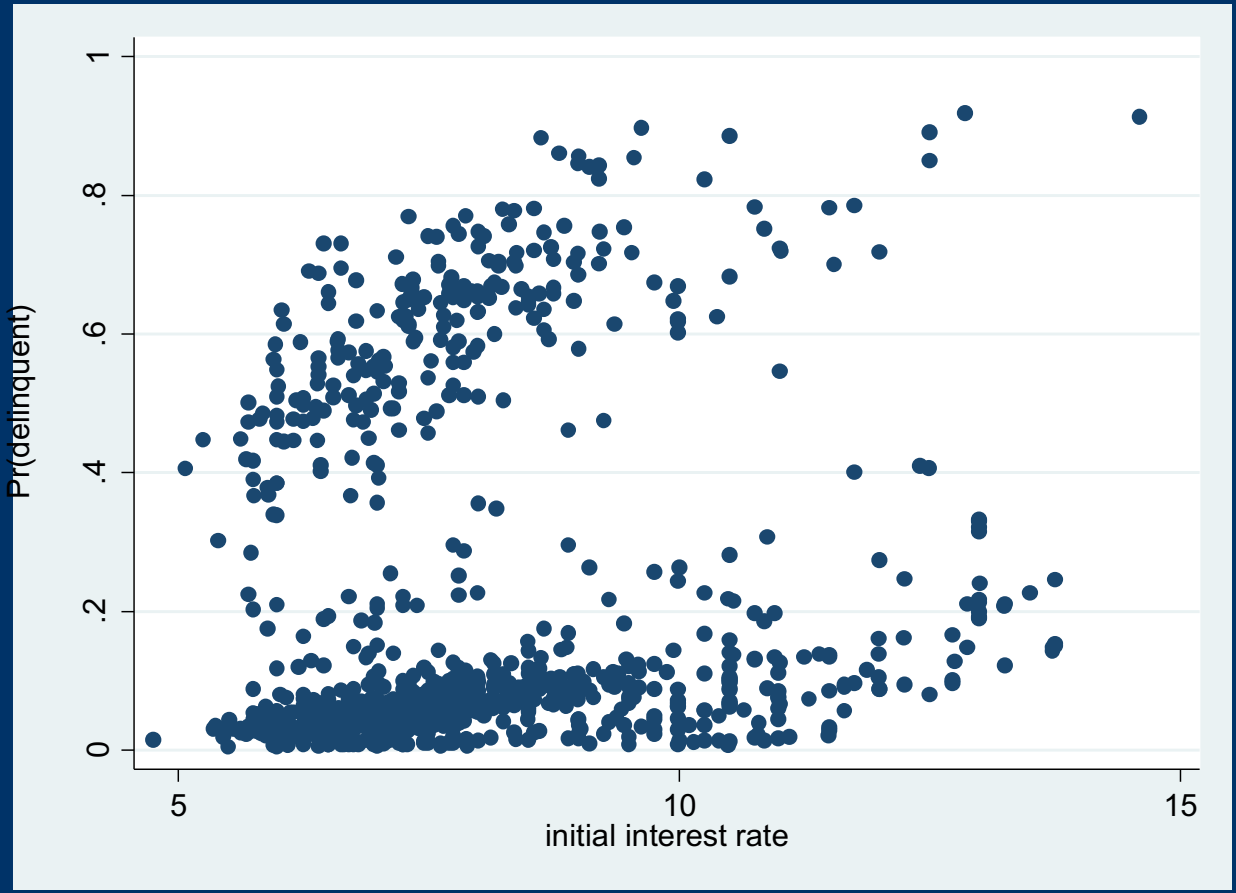


Categorical splits…

logit delinquent arm term credit amount rate insur ref lvr, robust

predict del_hat
scatter del_hat rate

Clearly something going on that splits up these groups… will ignore this for now.. But these are the sorts of things you want to look at and think about



Assessing Fit with Binary Responses

- We will end up with predicted probabilities $\hat{y}$, actual 0-1 information, y
- Given the structure of y, R^2 statistics leave something to be desired….
- We already discussed some likelihood based measures that may be helpful (Pseudo R^2, AIC, BIC), though the special categorical nature of the dependent variable affords us a few more options worth exploring.
- To be concrete, consider the problem of predicting which loans are likely to default, or predicting which individuals are ill.

Possible Outcomes

	Actual Default	Actual No Default
Predict Default	True Positive	False Positive
Predict No Default	False Negative	True Negative

If we assign some threshold percentage, above which we say that observation is predicted to be a 1 (= a default), then we can classify the results into true and false positives and negatives as in the table above.

Called confusion matrices. Ironic as these are probably the easiest matrices to understand in the entire course!

Stata Example: LV Defaults Pt 2

- Confusion matrix example at right. What is the cutoff they assign for a default predictions?
- What's special about that one?

```
. estat classification

Logistic model for delinquent

                 True
Classified |         D          ~D |      Total
-----------+----------------------+-----------
     +     |       118          65 |        183
     -     |        81         736 |        817
-----------+----------------------+-----------
   Total   |       199         801 |       1000

Classified + if predicted Pr(D) >= .5
True D defined as delinquent != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   59.30%
Specificity                     Pr( -|~D)   91.89%
Positive predictive value       Pr( D| +)   64.48%
Negative predictive value       Pr(~D| -)   90.09%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    8.11%
False - rate for true D         Pr( -| D)   40.70%
False + rate for classified +   Pr(~D| +)   35.52%
False - rate for classified -   Pr( D| -)    9.91%
--------------------------------------------------
Correctly classified                        85.40%
--------------------------------------------------
```

Confusion Matrices

- Confusion matrix example at right. What is the cutoff they assign for a default predictions?
- What's special about that one?
- Nothing!
- We could use the mean?
- What's the right level??

```
. summ delinquent

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
  delinquent |      1,000        .199    .3994478         0          1

. estat classification, cut(0.199)

Logistic model for delinquent

                 -------- True --------
Classified |         D             ~D |      Total
-----------+--------------------------+-----------
         + |       154            126 |        280
         - |        45            675 |        720
-----------+--------------------------+-----------
     Total |       199            801 |       1000

Classified + if predicted Pr(D) >= .199
True D defined as delinquent != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   77.39%
Specificity                     Pr( -|~D)   84.27%
Positive predictive value       Pr( D| +)   55.00%
Negative predictive value       Pr(~D| -)   93.75%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   15.73%
False - rate for true D         Pr( -| D)   22.61%
False + rate for classified +   Pr(~D| +)   45.00%
False - rate for classified -   Pr( D| -)    6.25%
--------------------------------------------------
Correctly classified                        82.90%
--------------------------------------------------
```

Possible Outcomes: Example

	Actual Default	Actual No Default
Predict Default	.005	.015
Predict No Default	.05	0.93

Let's say we have the following results…

Is this classifier any good?

Possible Outcomes: Example

	Actual Default	Actual No Default
Predict Default	.005	.015
Predict No Default	.05	0.93

Likely to have far fewer defaulters than non-defaulters ("unbalanced" classes), which makes interpretation not obvious.

Possible Outcomes: Example

	Actual Default	Actual No Default
Predict Default	.005	.015
Predict No Default	.05	0.93

How often am I right?
What proportion of the actual defaulters am I getting right?
What proportion of the actual no defaulters am I getting right?
What proportion of my predicted defaults is right?
What proportion of my predicted no defaults is right?

Do I care about these outcomes equally?
Are the answers the same for the entire dataset?