# Lecture 6b: Logit Probit

Chris Conlon

Sunday 5th October, 2025

NYU Stern

## Binary Choice: Overview

Many problems we are interested in look at discrete rather than continuous outcomes:

- Entering a Market/Opening a Store
- Working or a not
- Being married or not
- Exporting to another country or not
- Going to college or not
- Smoking or not
- etc.

## Simplest Example: Flipping a Coin

Suppose we flip a coin which is yields heads ($Y = 1$) and tails ($Y = 0$). We want to estimate the probability $p$ of heads:

$$Y_i = \begin{cases} 1 \text{ with probability } p \\ 0 \text{ with probability } 1 - p \end{cases}$$

We see some data $Y_1, \dots, Y_N$ which are (i.i.d.)

We know that $Y_i \sim Bernoulli(p)$.

## Simplest Example: Flipping a Coin

We can write the likelihood of $N$ Bernoulli trials as a Binomial:

$$\text{Pr}(Y_1 = y_1, Y_2 = y_2, \dots, Y_N = y_N) = f(y_1, y_2, \dots, y_N | p)$$

$$= \prod_{i=1}^{N} p^{y_i} (1-p)^{1-y_i}$$

$$= p^{\sum_{i=1}^{N} y_i} (1-p)^{N - \sum i=1^{N} y_i}$$

And then take logs to get the log likelihood:

$$\ln f(y_1, y_2, \dots, y_N | p) = \left( \sum_{i=1}^{N} y_i \right) \ln p + \left( N - \sum_{i=1}^{N} y_i \right) (1-p)$$

## Simplest Example: Flipping a Coin

Differentiate the log-likelihood to find the maximum:

$$
\begin{aligned}
\ln f(y_1, y_2, \ldots, y_N | p) &= \left(\sum_{i=1}^{N} y_i\right) \ln p + \left(N - \sum_{i=1}^{N} y_i\right) \ln(1-p) \\
\rightarrow 0 &= \frac{1}{\hat{p}} \left(\sum_{i=1}^{N} y_i\right) + \frac{-1}{1-\hat{p}} \left(N - \sum_{i=1}^{N} y_i\right) \\
\frac{\hat{p}}{1-\hat{p}} &= \frac{\sum_{i=1}^{N} y_i}{N - \sum_{i=1}^{N} y_i} = \frac{\overline{Y}}{1-\overline{Y}} \\
\hat{p}^{MLE} &= \overline{Y}
\end{aligned}
$$

That was a lot of work to get the obvious answer: fraction of heads.

## More Complicated Example: Adding Covariates

We probably are interested in more complicated cases where $p$ is not the same for all observations but rather $p(X)$ depends on some covariates. Here is an example from the Boston HMDA Dataset:

- 2380 observations from 1990 in the greater Boston area.
- Data on: individual Characteristics, Property Characteristics, Loan Denial/Acceptance (1/0).
- Mortgage Application process circa 1990-1991:
    - Go to bank
    - Fill out an application (personal+financial info)
    - Meet with loan officer
    - Loan officer makes decision
        - Legally in race blind way (discrimination is illegal but rampant)
        - Wants to maximize profits (ie: loan to people who don't end up defeaulting!)

## Loan Officer's Decision

Financial Variables:

- $P/I$ ratio
- housing expense to income ratio
- loan-to-value ratio
- personal credit history (FICO score, etc.)
- Probably some nonlinearity:
  - Very high $LTV > 80\%$ or $> 95\%$ is a bad sign (strategic defaults?)
  - Credit Score Thresholds

# Loan Officer's Decision

Goal Pr$(Deny = 1|black, X)$

- Lots of potential omitted variables which are correlated with race
  - Wealth, type of employment
  - family status
  - credit history
  - zip code of property
- Many redlining cases hinge on whether or not black applicants were treated in a discriminatory way.

**TABLE 11.1 Variables Included in Regression Models of Mortgage Decisions**

| Variable | Definition | Sample Average |
|---|---|---|
| ***Financial Variables*** | | |
| *P/I ratio* | Ratio of total monthly debt payments to total monthly income | 0.331 |
| *housing expense-to-income ratio* | Ratio of monthly housing expenses to total monthly income | 0.255 |
| *loan-to-value ratio* | Ratio of size of loan to assessed value of property | 0.738 |
| *consumer credit score* | 1 if no "slow" payments or delinquencies<br>2 if one or two slow payments or delinquencies<br>3 if more than two slow payments<br>4 if insufficient credit history for determination<br>5 if delinquent credit history with payments 60 days overdue<br>6 if delinquent credit history with payments 90 days overdue | 2.1 |
| *mortgage credit score* | 1 if no late mortgage payments<br>2 if no mortgage payment history<br>3 if one or two late mortgage payments<br>4 if more than two late mortgage payments | 1.7 |
| *public bad credit record* | 1 if any public record of credit problems (bankruptcy, charge-offs, collection actions)<br>0 otherwise | 0.074 |
| ***Additional Applicant Characteristics*** | | |
| *denied mortgage insurance* | 1 if applicant applied for mortgage insurance and was denied, 0 otherwise | 0.020 |
| *self-employed* | 1 if self-employed, 0 otherwise | 0.116 |
| *single* | 1 if applicant reported being single, 0 otherwise | 0.393 |
| *high school diploma* | 1 if applicant graduated from high school, 0 otherwise | 0.984 |
| *unemployment rate* | 1989 Massachusetts unemployment rate in the applicant's industry | 3.8 |
| *condominium* | 1 if unit is a condominium, 0 otherwise | 0.288 |
| *black* | 1 if applicant is black, 0 if white | 0.142 |
| *deny* | 1 if mortgage application denied, 0 otherwise | 0.120 |

First thing we might try is OLS

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- What does $\beta_1$ mean when $Y$ is binary? Is $\beta_1 = \frac{\Delta Y}{\Delta X}$?
- What does the line $\beta_0 + \beta_1 X$ when $Y$ is binary?
- What does the predicted value $\hat{Y}$ mean when $Y$ is binary? Does $\hat{Y} = 0.26$ mean that someone gets approved or denied for a loan?

OLS is called the linear probability model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

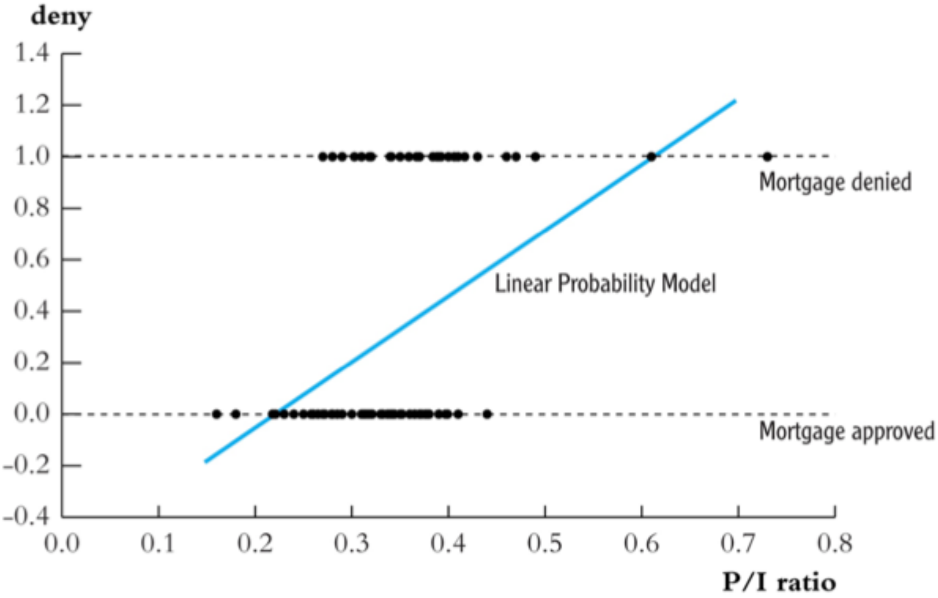
because:

$$\mathbb{E}[Y|X] = 1 \cdot \text{Pr}(Y = 1|X) + 0 \cdot \text{Pr}(Y = 0|X)$$
$$\text{Pr}(Y = 1|X) = \beta_0 + \beta_1 X_i + \varepsilon_i$$

The predicted value is a probability and

$$\beta_1 = \frac{\text{Pr}(Y = 1|X = x + \Delta x) - \text{Pr}(Y = 1|X = x)}{\Delta x}$$

So $\beta_1$ represents the average change in probability that $Y = 1$ for a unit change in $X$.


deny
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
P/I ratio
Mortgage denied
Linear Probability Model
Mortgage approved

## That didn't look great

- Is the marginal effect $\beta_1$ actually constant or does it depend on $X$?
- Sometimes we predict $\hat{Y} > 1$ or $\hat{Y} < 0$. What does that even mean? Is it still a probability?
- Fit in the middle seems not so great – what does $\hat{Y} = 0.5$ mean?

## Results

$$\widehat{deny}_i = -.091 + .559 \cdot \text{P/I ratio} \qquad + .177 \cdot \text{black}$$
$$(0.32)(.098) \qquad\qquad (.025)$$

Marginal Effects:

- Increasing $P/I$ from $0.3 \rightarrow 0.4$ increases probabilty of denial by $5.59$ percentage points. (True at all level of $P/I$).
- At all $P/I$ levels blacks are $17.7$ percentage points more likely to be denied.
- But still some omitted factors.
- True effects are likely to be nonlinear can we add polynomials in $P/I$? Dummies for different levels?

## Moving Away from LPM

Problem with the LPM/OLS is that it requires that marginal effects are constant or that probability can be written as linear function of parameters.

$$\Pr(Y = 1|X) = \beta_0 + \beta_1 X + \varepsilon$$

Some desirable properties:

- Can we restrict our predictions to $[0, 1]$?
- Can we preserve monotonicity so that $\Pr(Y = 1|X)$ is increasing in $X$ for $\beta_1 > 0$?
- Some other properties (continuity, etc.)
- Want a function $F(z) : (-\infty, \infty) \to [0, 1]$.
- What function will work?

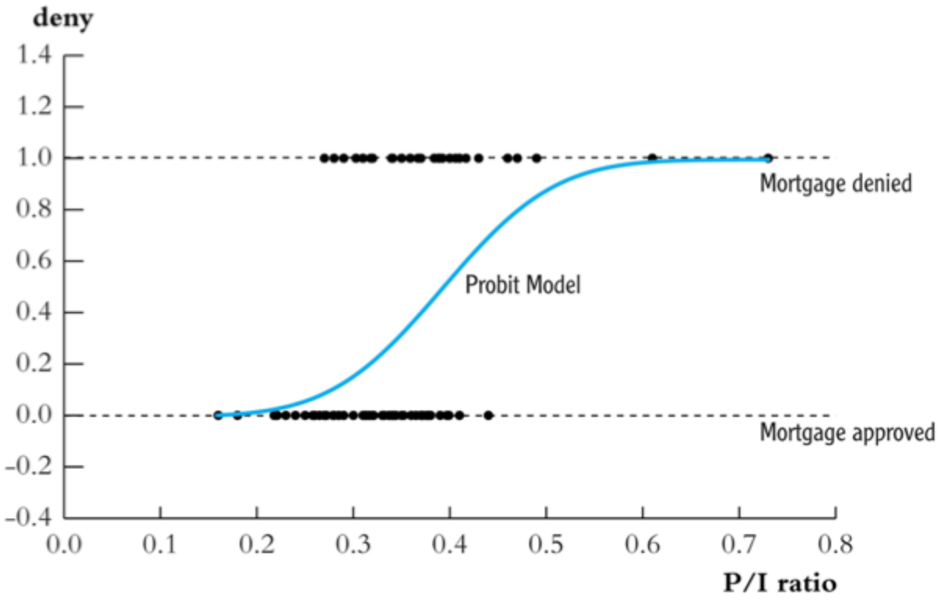
# Moving Away from LPM

Problem with the LPM/OLS is that it requires that marginal effects are constant or that probability can be written as linear function of parameters.

$$\Pr(Y = 1|X) = \beta_0 + \beta_1 X + \varepsilon$$

Some desirable properties:

- Can we restrict our predictions to $[0, 1]$?
- Can we preserve monotonicity so that $\Pr(Y = 1|X)$ is increasing in $X$ for $\beta_1 > 0$?
- Some other properties (continuity, etc.)
- Want a function $F(z) : (-\infty, \infty) \to [0, 1]$.
- What function will work?


deny
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
Mortgage denied
Probit Model
Mortgage approved
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
P/I ratio

## Choosing a transformation

$$\text{Pr}(Y = 1|X) = F(\beta_0 + \beta_1 X)$$

- One $F(\cdot)$ that works is $\Phi(z)$ the normal CDF. This is the probit model.
  - Actually any CDF would work but the normal is convenient.
- One $F(\cdot)$ that works is $\frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$ the logistic function . This is the logit model.
- Both of these give 'S'-shaped curves.
- The LPM is $F(\cdot)$ is the identity function (which doesn't satisfy my $[0, 1]$ property).
- This $F(\cdot)$ is often called a link function. Why?

# Why use the normal CDF?

Has some nice properties:

- Gives us more of the ‘S’ shape
- Pr$(Y = 1|X)$ is increasing in $X$ if $\beta_1 > 0$.
- Pr$(Y = 1|X) \in [0, 1]$ for all $X$
- Easy to use – you can look up or use computer for normal CDF.
- Relatively straightforward interpretation
  - $Z = \beta_0 + \beta_1 X$ is the $z$-value.
  - $\beta_1$ is the change in the $z$-value for a change in $X_1$.

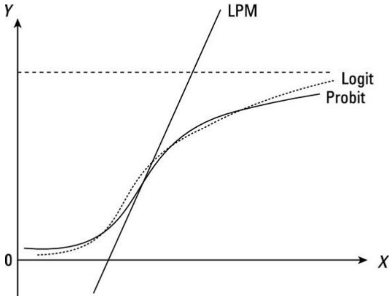
## Why use the logistic CDF?

Has some nice properties:

- Gives us more of the ‘S’ shape
- Pr$(Y=1|X)$ is increasing in $X$ if $\beta_1 > 0$.
- Pr$(Y=1|X) \in [0,1]$ for all $X$
- Easy to compute: $\frac{1}{1+e^{-z}} = \frac{e^z}{1+e^z}$ has analytic derivatives too.
- Log odds interpretation
  - $\log(\frac{p}{1-p}) = \beta_0 + \beta_1 X$
  - $\beta_1$ tells us how log odds ratio responds to $X$.
  - $\frac{p}{1-p} \in (-\infty, \infty)$ which fixes the $[0,1]$ problem in the other direction.
  - more common in other fields (epidemiology, biostats, etc.).
- Also has the property that $F(z) = 1 - F(-z)$.
- Similar to probit but different scale of coefficients
- Logit/Logistic are sometimes used interchangeably but sometimes mean different things depending on the literature.

# A quick comparison

- LPM prediction departs greatly from CDF long before $[0, 1]$ limits.
- We get probabilities that are too extreme even for $X\hat{\beta}$ "in bounds".
- Some (MHE) argue that though $\hat{Y}$ is flawed, constant marginal effects are still OK.
- Logit and Probit are highly similar



# HMDA: Results

| | OLS | Probit | Logit |
|---|---|---|---|
| Constant | -0.174*** | -2.96*** | -5.56*** |
| | (0.026) | (0.205) | (0.406) |
| Black | 0.081*** | 0.367*** | 0.657*** |
| | (0.017) | (0.097) | (0.177) |
| Debt/Income | 0.471*** | 2.58*** | 5.03*** |
| | (0.087) | (0.546) | (1.03) |
| Housing/Income | -0.069 | -0.328 | -0.405 |
| | (0.096) | (0.652) | (1.24) |
| LTV: Medium | 0.028** | 0.193** | 0.428*** |
| | (0.012) | (0.081) | (0.158) |
| LTV: High | 0.189*** | 0.779*** | 1.48*** |
| | (0.033) | (0.175) | (0.309) |
| Consumer Credit | 0.031*** | 0.153*** | 0.286*** |
| | (0.004) | (0.021) | (0.040) |
| Mortgage Credit | 0.019* | 0.134* | 0.258* |
| | (0.011) | (0.074) | (0.141) |
| Public Bad Credit | 0.200*** | 0.712*** | 1.25*** |
| | (0.023) | (0.118) | (0.205) |
| Denied Mortgage Ins | 0.701*** | 2.54*** | 4.53*** |
| | (0.041) | (0.284) | (0.554) |
| Pseudo $R^2$ | 0.51868 | 0.26407 | 0.26586 |
| Observations | 2,380 | 2,380 | 2,380 |

*IID standard-errors in parentheses*

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

- We cannot compare parameter estimates across specifications
- Rule of thumb: $\beta_{logit} \approx 1.81 \cdot \beta_{probit}$
- It is better to compare marginal effects $\frac{\partial \mathbb{E}[Y_i = 1 | X]}{\partial x_k}$
- For LPM these are constant $\beta_k$, for logit/probit they depend on $X_i\beta$.

We sometimes call these single index models or threshold crossing models

$$Z_i = X_i\beta$$

- We start with a potentially large number of regressors in $X_i$ but $X_i\beta = Z_i$ is a scalar
- We can just calculate $F(Z_i)$ for Logit or Probit (or some other CDF).
- $Z_i$ is the index. if $Z_i = X_i\beta$ we say it is a linear index model.

# Marginal effects

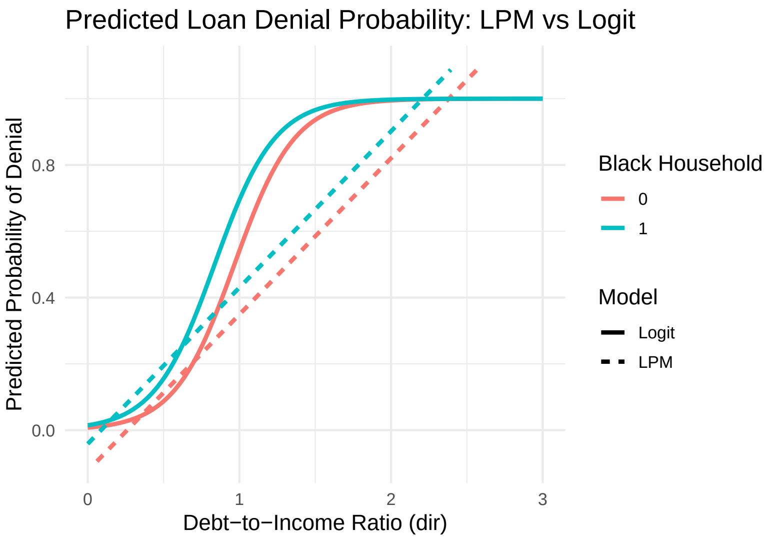
$$\frac{\partial \mathbb{E}[Y_i|X_i]}{\partial X_{ik}} = f(Z_i)\beta_k$$

- The whole point was that we wanted marginal effects not to be constant
- So where do we evaluate?
  - Software often plugs in mean or median values for each component
  - Alternatively we can integrate over $X$ and compute:

$$\mathbb{E}_{X_i}[f(Z_i)\beta_k]$$

  - The right thing to do is probably to plot the response surface (either probability) or change in probability over all $X$.

# Predicted Effects



An alternative way to think about this problem is that there is a continuously distributed $Y^*$ that we as the econometrician don't observe.

$$Y_i = \begin{cases} 1 \text{ if } Y^* > 0 \\ 0 \text{ if } Y^* \leq 0 \end{cases}$$

- Instead we only see whether $Y^*$ exceeds some threshold (in this case $0$).
- We can think about $Y^*$ as a latent variable.
- Sometimes you will see this description in the literature, everything else is the same!

# What does software do?

- One temptation might be nonlinear least squares:

$$\hat{\beta}^{NLLS} = \arg\min_{\beta} \sum_{i=1}^{N} (Y_i - \Phi(X_i \beta))^2$$

- Turns out this isn't what people do.
- We can't always directly estimate using the log-odds

$$\log\left(\frac{p}{1-p}\right) = \beta X_i + \varepsilon_i$$

- The problem is that $p$ or $p(X_i)$ isn't really observed.

- Can construct an MLE:

$$\hat{\beta}^{MLE} = \arg\max_{\beta} \prod_{i=1}^{N} F(Z_i)^{y_i}(1 - F(Z_i))^{1-y_i}$$

$$Z_i = \beta_0 + \beta_1 X_i$$

- Probit: $F(Z_i) = \Phi(Z_i)$ and its derivative (density) $f(Z_i) = \phi(Z_i)$. Also is symmetric so that $1 - F(Z_i) = F(-Z_i)$.
- Logit: $F(Z_i) = \frac{1}{1+e^{-z}}$ and its derivative (density) $f(Z_i) = \frac{e^{-z}}{(1+e^{-z})^2}$ a more convenient property is that $\frac{f(z)}{F(z)} = 1 - F(z)$ this is called the hazard rate.

## A probit trick

Let $q_i = 2y_i - 1$

$$F(q_i \cdot Z_i) = \begin{cases} F(Z_i) & \text{when } y_i = 1 \\ F(-Z_i) = 1 - F(Z_i) & \text{when } y_i = 0 \end{cases}$$

So that

$$\ell(y_1, \ldots, y_n | \beta) = \sum_{i=1}^{N} \ln F(q_i \cdot Z_i)$$

$$\begin{aligned}
\ell(y_1, \dots, y_n|\beta) &= \sum_{i=1}^{N} y_i \ln F(Z_i) + (1 - y_i) \ln(1 - F(Z_i)) \\
\frac{\partial l}{\partial \beta} &= \sum_{i=1}^{N} \frac{y_i}{F(Z_i)} \frac{dF}{d\beta}(Z_i) - \frac{1 - y_i}{1 - F(Z_i)} \frac{dF}{d\beta}(Z_i) \\
&= \sum_{i=1}^{N} \frac{y_i \cdot f(Z_i)}{F(Z_i)} \frac{dZ_i}{d\beta} - \sum_{i=1}^{N} \frac{(1 - y_i) \cdot f(Z_i)}{1 - F(Z_i)} \frac{dZ_i}{d\beta} \\
&= \sum_{i=1}^{N} \left[ \frac{y_i \cdot f(Z_i)}{F(Z_i)} X_i - \frac{(1 - y_i) \cdot f(Z_i)}{1 - F(Z_i)} X_i \right]
\end{aligned}$$

# FOC of Log-Likelihood (Logit)

This is the score of the log-likelihood:

$$\frac{\partial l}{\partial \beta} = \nabla_{\beta} \cdot \ell(\mathbf{y}; \beta) = \sum_{i=1}^{N} \left[ y_i \frac{f(Z_i)}{F(Z_i)} - (1 - y_i) \frac{f(Z_i)}{1 - F(Z_i)} \right] \cdot X_i$$

It is technically also a moment condition. It is easy for the logit

$$\begin{aligned} \nabla_{\beta} \cdot \ell(\mathbf{y}; \beta) &= \sum_{i=1}^{N} [y_i(1 - F(Z_i)) - (1 - y_i)F(Z_i)] \cdot X_i \\ &= \sum_{i=1}^{N} \underbrace{[y_i - F(Z_i)]}_{\varepsilon_i} \cdot X_i \end{aligned}$$

This comes from the hazard rate.

## FOC of Log-Likelihood (Probit)

This is the score of the log-likelihood:

$$\frac{\partial l}{\partial \beta} = \nabla_\beta \cdot \ell(\mathbf{y};\beta) = \sum_{i=1}^{N}\left[y_i \frac{f(Z_i)}{F(Z_i)} - (1-y_i)\frac{f(Z_i)}{1-F(Z_i)}\right] \cdot X_i$$
$$= \sum_{y_i=1} \frac{\phi(Z_i)}{\Phi(Z_i)} X_i + \sum_{y_i=0} \frac{-\phi(Z_i)}{1-\Phi(Z_i)} X_i$$

Using the $q_i = 2y_i - 1$ trick

$$\nabla_\beta \cdot \ell(\mathbf{y};\beta) = \sum_{i=1}^{N} \underbrace{\frac{q_i \phi(q_i Z_i)}{\Phi(Z_i)}}_{\lambda_i} X_i$$

We could also take second derivatives to get the Hessian matrix:

$$\frac{\partial l^2}{\partial\beta\partial\beta'} = -\sum_{i=1}^{N} y_i \frac{f(Z_i)f(Z_i) - f'(Z_i)F(Z_i)}{F(Z_i)^2} X_i X_i'$$

$$+\sum_{i=1}^{N}(1-y_i)\frac{f(Z_i)f(Z_i) - f'(Z_i)(1-F(Z_i))}{(1-F(Z_i))^2} X_i X_i'$$

This is a $K \times K$ matrix where $K$ is the dimension of $X$ or $\beta$.

# The Hessian Matrix (Logit)

For the logit this is even easier (use the simplified logit score):

$$\frac{\partial l^2}{\partial\beta\partial\beta'} = -\sum_{i=1}^{N} f(Z_i)X_iX_i'$$
$$= -\sum_{i=1}^{N} F(Z_i)(1 - F(Z_i))X_iX_i'$$

This is negative semi definite

## The Hessian Matrix (Probit)

Recall

$$\nabla_\beta \cdot \ell(\mathbf{y};\beta) = \sum_{i=1}^{N} \underbrace{\frac{q_i \phi(q_i Z_i)}{\Phi(Z_i)}}_{\lambda_i} X_i$$

Take another derivative and recall $\phi'(z_i) = -z_i \phi(z_i)$

$$\begin{aligned}\nabla_\beta^2 \cdot \ell(\mathbf{y};\beta) &= \sum_{i=1}^{N} \frac{q_i \phi'(q_i Z_i)\Phi(z_i) - q_i \phi(z_i)^2}{\Phi(z_i)^2} X_i X_i' \\ &= -\lambda_i(z_i + \lambda_i) \cdot X_i X_i'\end{aligned}$$

Hard to show but this is negative definite too.

## Inference

- If we have the Hessian Matrix, inference is straightforward.
- $\mathbf{H}_f(\widehat{\beta}^{MLE})$ tells us about the curvature of the log-likelihood around the maximum.
  - Function is flat $\rightarrow$ not very precise estimates of parameters
  - Function is steep $\rightarrow$ precise estimates of parameters
- Construct Fisher Information $I(\widehat{\beta}^{MLE}) = -\mathbb{E}[\mathbf{H}_f(\widehat{\beta}^{MLE})]$ where expectation is over the data.
  - Logit does not depend on $y_i$ so $\mathbb{E}[\mathbf{H}_f(\widehat{\beta}^{MLE})] = \mathbf{H}_f(\widehat{\beta}^{MLE})$.
  - Probit does depend on $y_i$ so $\mathbb{E}[\mathbf{H}_f(\widehat{\beta}^{MLE})] \neq \mathbf{H}_f(\widehat{\beta}^{MLE})$.
- Inverse Fisher information $-\mathbb{E}[\mathbf{H}_f(\widehat{\beta}^{MLE})]^{-1}$ is an estimate of the variance covariance matrix for $\widehat{\beta}$.
- $\sqrt{\text{diag}[\mathbb{E}[-\mathbf{H}_f(\widehat{\beta}^{MLE})]^{-1}]}$ is an estimate for $SE(\widehat{\beta})$.

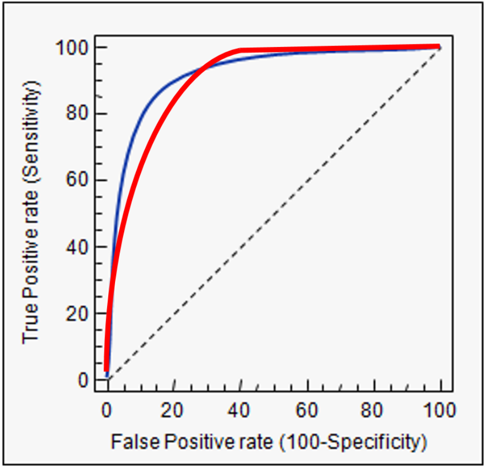
How well does the model fit the data?

- No $R^2$ measure (why not?).
- Well we have likelihood units so average likelihood tells us something but is hard to interpret.
- $\rho = 1 - \frac{\ell(\hat{\beta}^{MLE})}{\ell(\beta_0)}$ where $\ell(\beta_0)$ is the likelihood of a model with just a constant (unconditional probability of success).
  - If we don't do any better than unconditional mean then $\rho = 0$.
  - Won't ever get all of the way to $\rho = 1$.

## Goodness of Fit #2: Confusion Matrix

- Machine learning likes to think about this problem more like classification then regression.
- A caution: these are regression models not classification models.
- Predict either $\hat{y}_i = 1$ or $\hat{y}_i = 0$ for each observation.
- Predict $\hat{y}_i = 1$ if $\Pr(y_i = 1|X_i = x) \geq 0.5$ or $F(X_i\hat{\beta}) > 0.5$.
- Imagine for cells Prediction: $\{Success, Failure\}$, Outcome $\{Success, Failure\}$
- Can construct this using the R package `caret` and command `caret`.

| | | Actual Values | |
|---|---|---|---|
| | | Positive (1) | Negative (0) |
| Predicted Values | Positive (1) | TP | FP |
| | Negative (0) | FN | TN |




- At each predicted probability calculate both True Positive Rate and False Positive Rate.
- AOC is area under the curve

# Thanks!