MLE - Binary Choice Models

Introduction

- Very often we want to model some qualitative outcome.
  - Whether someone decides to get married.
  - Whether a company will go bankrupt.
  - Whether a defendant will be found guilty.
  - Whether a football team wins a game.
  - Whether a stock price will be higher tomorrow.
- The variable we want to explain can be written as a binary variable (0 or 1), hence a **binary choice model.**
- These types of model are particularly prevalent in economics. This is because we are very interested in people's choices. What are the causes of individual's choices... Why did someone decide to buy a car?

Binary Choice Model

- Because Y can only take the value 0 or 1, we have

$$E[Y|X] = Pr(Y = 1|X) \equiv p(X)$$

  i.e. our regression will model the probability of 'success', however that is defined.

- It is also useful to note, that such a binary variable has a Bernoulli distribution and therefore has:

$$Var[y|X] = p(X)[1 - p(X)].$$

Linear Probability Model (LPM)

- What happens if we use the standard linear regression model? That is, we assume

$$E[Y|X] = p(X) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

- There are a couple of issues with this, however, it can still be used...

- Using data mroz, we try to determine the factors which affect whether a woman works in formal employment.

```
data_labour = mroz
data_labour$age2 = (data_labour$age - mean(data_labour$age))^2

LM = lm(data = data_labour, inlf ~ kidslt6 + kidsge6 + age + age2 + educ +
          hushrs + husage + huseduc + huswage + mtr)
summary(LM)
```

Linear Probability Model (LPM)

- To interpret the coefficients we say, for example, if the woman has an extra year of education, her probability of working increases by 4 percentage points.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.815e+00  3.599e-01  10.599  < 2e-16 ***
kidslt6     -2.532e-01  3.528e-02  -7.177 1.74e-12 ***
kidsge6      1.072e-02  1.413e-02   0.759   0.4483
age         -7.587e-03  4.449e-03  -1.705   0.0886 .
age2        -2.091e-04  2.680e-04  -0.780   0.4355
educ         4.194e-02  9.095e-03   4.612 4.70e-06 ***
hushrs      -2.321e-04  3.194e-05  -7.267 9.35e-13 ***
husage      -5.817e-03  4.398e-03  -1.323   0.1863
huseduc     -9.979e-03  7.166e-03  -1.393   0.1641
huswage     -6.013e-02  6.516e-03  -9.228  < 2e-16 ***
mtr         -2.970e+00  3.351e-01  -8.862  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4361 on 742 degrees of freedom
Multiple R-squared:  0.2362,    Adjusted R-squared:  0.2259
F-statistic: 22.95 on 10 and 742 DF,  p-value: < 2.2e-16
```

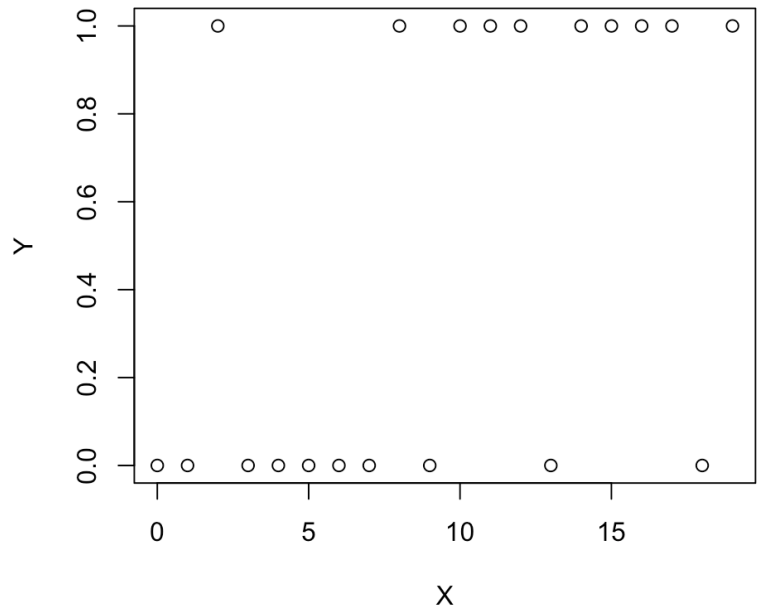
Shortcomings of LPM

- So, what's wrong with the LPM??
- First, check out the fitted values from the previous regression using: LM$fitted.values.
- You should be able to see that they don't all lie in $[0, 1]$. This is a little weird because the fitted value is our best guess of the probability that $Y = 1$.
- Secondly, the model must contain heteroskedasticity because

$$Var\left[\epsilon|X\right] = Var\left[y|X\right] = p(X)[1 - p(X)].$$

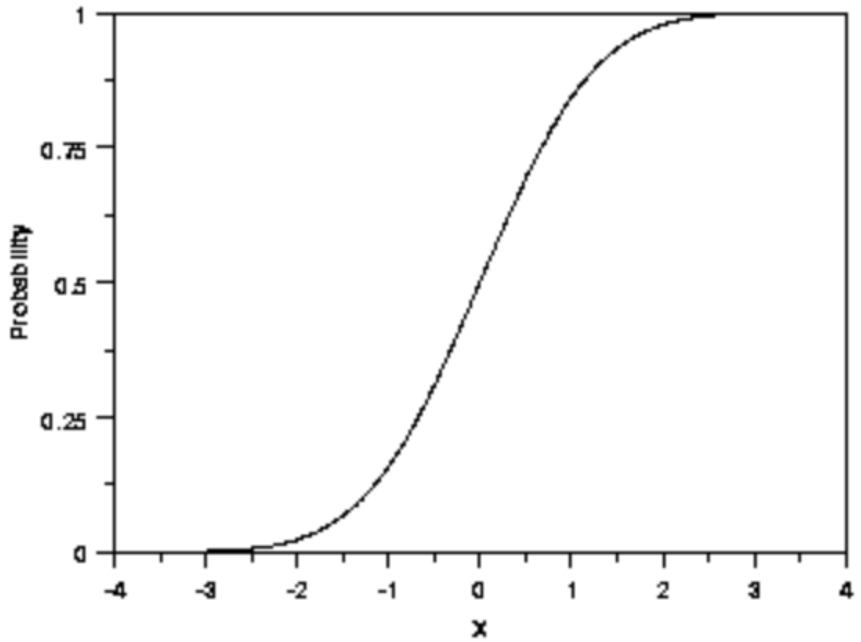
Shortcomings of LPM

- Finally, look at a simple bivariate plot of one of the X on the binary Y. Here is an example to illustrate the issue:



- Is a linear model appropriate? No!
- So, what would be more appropriate? Ideally, we want a function which has the following kind of shape:

Alternative to LPM



- But this is exactly the shape of the Standard Normal CDF! (And in fact, also the logistic function)
- Bonus: it constrains the outcome to be in $[0, 1]$.

Probit and Logit Models

- So our new model becomes

$$E[Y|X] = Pr[Y = 1|X] = F(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)$$

  where $F(\cdot)$ is either the CDF of the Standard Normal, $\Phi(\cdot)$, or the logistic function, $\Lambda(z) = \frac{exp(z)}{1+exp(z)}$.

- The two resulting estimators are known as the probit and logit estimators, respectively.

- In practice, there is very little difference in the results from these two estimators and they have the same asymptotic properties. It is typically a matter of taste which people use.

MLE Estimation

- Because we are now in a nonlinear set-up, we cannot use the method of moments or least squares to estimate the unknown parameters, $\beta_0, \ldots, \beta_k$.
- Instead, we use maximum likelihood. (Since you've covered MLE in previous courses I will skip this here).
- To construct our MLE we need to write out the likelihood function, that is, the joint density of the data. If our observations are drawn independently, our likelihood function can be written as

$$L(\beta; y_1, \ldots y_n, x_1, \ldots, x_n) = \prod_{i=1}^{n} f(y_i, x_i; \beta).$$

MLE Estimation - Probit

- Let $p(x;\beta) = Pr(y = 1|x;\beta) = \Phi(\beta_0 + \beta_1 x)$ (for ease of notation I'm just using a single regressor).
- Our dependent variable y has a Bernoulli distribution

$$\begin{aligned} f(y, x;\beta) &= p(x;\beta)^y \left[1 - p(x;\beta)\right]^{1-y} \\ &= \Phi(\beta_0 + \beta_1 x)^y \left[1 - \Phi(\beta_0 + \beta_1 x)\right]^{1-y} \end{aligned}$$

- We construct the likelihood and then the log-likelihood from this, giving

$$\begin{aligned} ln\left(L(\beta; y_1, \ldots, y_n, x_1, \ldots, x_n)\right) = \sum_{i=1}^{n} \Big\{ & y_i ln\left(\Phi(\beta_0 + \beta_1 x)\right) + \\ & (1 - y_i) ln\left(1 - \Phi(\beta_0 + \beta_1 x)\right) \Big\}. \end{aligned}$$

MLE Estimation - Probit

- We can then maximise this w.r.t β using some numerical optimisation algorithm. Note that we cannot obtain a closed-form solution for this estimator.
- The same procedure is used for the logit estimator but just replacing the Normal CDF with the logistic function.
- Because we have used maximum likelihood, it means that our estimators enjoy all the nice asymptotic properties that MLE has. That is, the estimators are consistent, asymptotically normal, invariant to transformation, and asymptotically achieve the Cramer-Rao Lower Bound (i.e. they are the most efficient in the class of unbiased estimators).

Probit and Logit - Partial Effects

- Our parameters of interest are not necessarily the β's themselves. Remember, we are interested in the marginal effect of X on Y, or, more specifically

$$\frac{\partial E[Y|X = x]}{\partial x}.$$

- In our case, this means we are interested in

$$\frac{\partial p(x)}{\partial x_j} = \frac{\partial F(X\beta)}{\partial x_j} = \beta_j F'(X\beta).$$

- So although β_j gives us the correct sign of the effect (since the derivative of both the Normal CDF and the logistic function are always non-negative), they do not give us the marginal effect.
- Notice that the marginal effect depends on the value of every regressor. As such, we typically calculate the marginal effect at the mean of all regressors.

Example: Conviction of Defendants (justice_data.csv)

- We are going to use some of my own data to see if we can predict which defendants will be convicted. You can download this from blackboard.

```
install.packages('mfx')
library(mfx)

justice_data = read.csv("~/Documents/Justice_data.csv")

justice_data$age2 = (justice_data$age - mean(justice_data$age))^2

probit1 = glm(data = justice_data, formula = convicted ~
                female + black + age  + age2 + average_prior_arrests,
              family = binomial(link = "probit"))
summary(probit1)
```

Example: Conviction of Defendants

- We get the following output. Note the strong racial discrimination (can we conclude this is discrimination?) There's also a gender bias too. Can we interpret these as marginal effects? No!

```
Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)           -6.637e-01  7.372e-02  -9.003   <2e-16 ***
female                -1.055e-01  6.045e-02  -1.745   0.0810 .
black                  1.037e-01  4.305e-02   2.409   0.0160 *
age                    7.129e-03  2.831e-03   2.518   0.0118 *
age2                  -7.644e-05  1.418e-04  -0.539   0.5899
average_prior_arrests  1.120e-02  8.013e-03   1.398   0.1621
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6314.5  on 4999  degrees of freedom
Residual deviance: 6295.0  on 4994  degrees of freedom
AIC: 6307

Number of Fisher Scoring iterations: 4
```

Marginal Effects in R

- We have to use a different function to get out the marginal effects.

```
probit2 = probitmfx(data = justice_data, formula = convicted ~
                    female + black + age + age2 + average_prior_arrests)

probit2$mfxest
```

```
                            dF/dx    Std. Err.          z      P>|z|
female                -3.724672e-02 0.0208946676 -1.7825944 0.07465235
black                  3.779002e-02 0.0158542344  2.3835919 0.01714460
age                    2.567910e-03 0.0010195905  2.5185702 0.01178324
age2                  -2.753495e-05 0.0000510833 -0.5390205 0.58987269
average_prior_arrests  4.035119e-03 0.0028865481  1.3979048 0.16214167
```

Goodness-of-Fit (Causality/Prediction)

- Notice that there was no R^2 in the output. How do we know if our model is any good? We see if it predicts well.
- But we said at the start of the course that we're not concerned with prediction, so why are we judging our causal analysis using prediction? (R^2 is also a prediction measure by the way).
- Essentially, because we have no other means to evaluate the quality of our model. There's no statistical method to determine if the relationship found is causal, and as a result, no means to test 'how much causality you've found'.
- If you've found factors which explain (cause) a good deal of the variation in your outcome, your prediction will also be good.

Goodness-of-Fit (Causality/Prediction)

- Equally, if your factors do not cause much change in your outcome, your prediction will be poor.
- The key distinction is that just because your model predicts well, does not mean that you have found causality.
- Analogously, if your model predicts poorly, it does not mean that you don't have causality. You may have a very small causal effect, but just because it is small, does not mean it is not causal.
- Ok, back to goodness-of-fit....

Goodness-of-Fit

- We use hit-ratios to determine how well we have done.
- The hit-ratio is simply the percentage of correct predictions. To get the predictions ('convicted' or 'not convicted'), we say that if the fitted value is above 0.5, we predict $y = 1$ and $y = 0$ otherwise.

```
fitted = probit1$fitted.values
pred = ifelse(fitted>0.5, 1, 0)
true = justice_data$convicted

length(pred[pred==true]) / length(pred)
```

- This gives a hit-ratio of 67.4%. Not bad.... or is it??

Goodness-of-Fit

- We want some benchmark with which to compare this to.
- We use an empty model (or a naive model), which has no regressors. This is the equivalent of always predicting everyone to be convicted, or everyone to not be convicted, depending on which is more prevalent in the data.
- In our case, 67.4% of people are not convicted.... this is exactly the same as our model. If we look more closely, our model actually does simply predict everyone to be let off. So, not a very good model!

Example: Surviving the Titanic (titanic_data.csv)

- We are trying to explain the causes of why people survived the sinking of the titanic. The data can be dowloaded from blackboard.

```
titanic_data = read.csv("~/Documents/titanic_data.csv")
titanic_data = titanic_data[complete.cases(titanic_data), ]

logit = glm(data = titanic_data, formula = survived ~
              male + first_class + second_class + male +
              age + sibs_spouse + parents_children,
            family = binomial(link = "logit"))
summary(logit)
```

- We're using a logit in this example just to see how it works (it's pretty much exactly the same).
- *sib – spouse* indicates the number of siblings or spouse(s!) you have on board. *Parents – children* is similarly defined.

Example: Surviving the Titanic

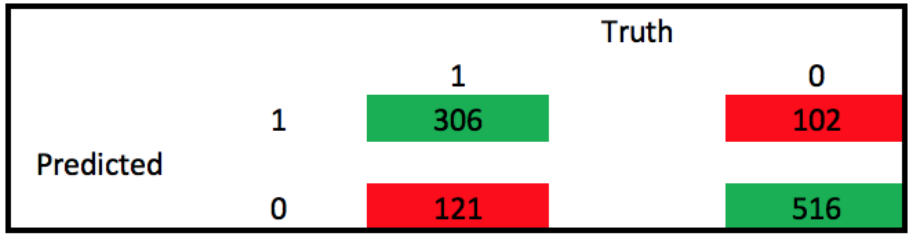
- Pretty much what we would expect... (remember these are not the marginal effects though)

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       1.554769   0.246348   6.311 2.77e-10 ***
male             -2.556860   0.173270 -14.756  < 2e-16 ***
first_class       2.352022   0.228803  10.280  < 2e-16 ***
second_class      0.985266   0.199384   4.942 7.75e-07 ***
age              -0.039489   0.006634  -5.952 2.64e-09 ***
sibs_spouse      -0.352915   0.105349  -3.350 0.000808 ***
parents_children  0.074361   0.099907   0.744 0.456695
```

- Note that there's also a variable for the fare that was paid. Is it significant if you include it in the above regression? Why not? Try putting it in instead of first and second class dummies. Does it make sense to think of this in a causal way?

Surviving the Titanic - Goodness-of-Fit

- The hit-ratio is 78.7%. The empty model gives 59.1%. So we didn't do too badly.
- Let's look at one more thing... the confusion matrix:



- The code in R to calculate these numbers:

```
length(pred[pred==0 & true==0])
length(pred[pred==1 & true==0])
length(pred[pred==0 & true==1])
length(pred[pred==1 & true==1])
```

Testing in an MLE World

- We can test all the same hypotheses using these MLE estimators
- However, because we cannot assume normality, we cannot use our regular t and F tests.
- Instead we appeal to asymptotic approximations and use the Wald, Likelihood Ratio, and Lagrange Multiplier tests which we cover now.

Trinity of Tests

- The Trinity of Tests are:
  - Wald Test
  - Likelihood Ratio Test (LR)
  - Lagrange Multiplier Test (LM)
- Although any of these three tests can be used to test a hypothesis, each has its own advantages.
- Suppose we are interested in testing one or more linear restrictions of the parameter vector $\theta = (\theta_1 \ldots, \theta_k)'$, say $R\theta = c$. Where R is a fixed $J \times k$ matrix, and c is a fixed $J \times 1$ vector (J is the number of restrictions).

Example of Restrictions

- For example, suppose we have 3 regressors and an intercept. And we want to test $\beta_1 = \beta_2 = 0$:

$$R\theta = c$$

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

- Or, suppose we want to test $\beta_2 = \beta_3$:

$$R\theta = c$$

$$\begin{pmatrix} 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 0 \end{pmatrix}.$$

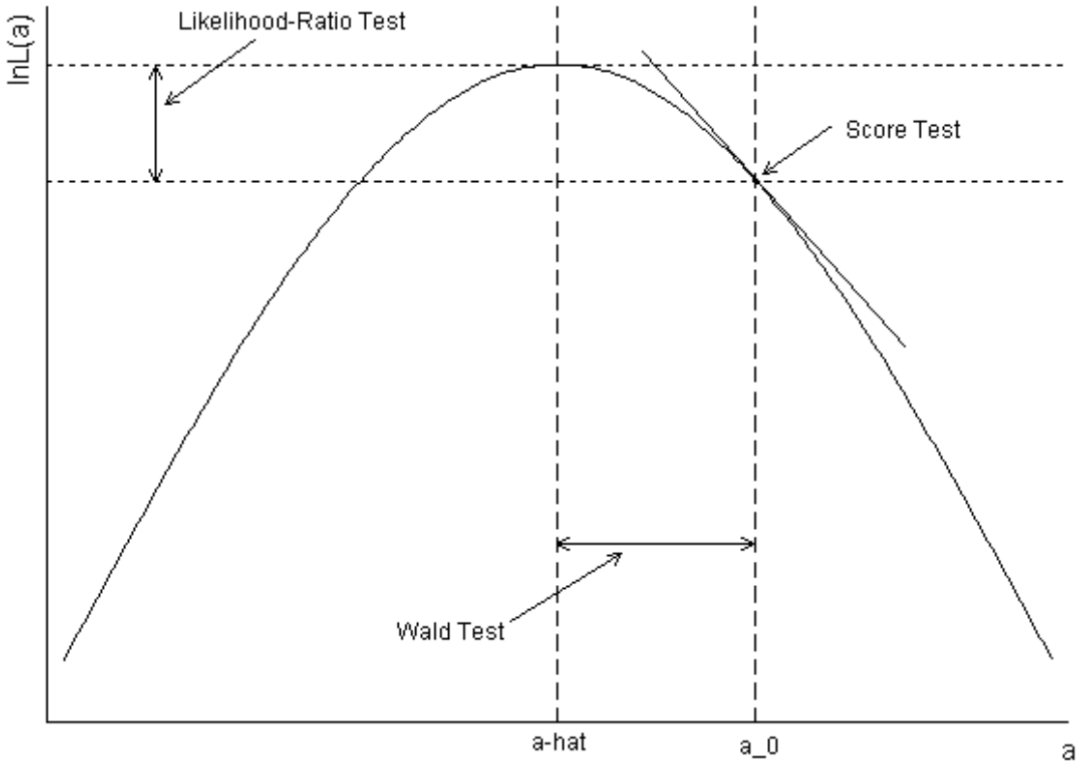
Trinity of Tests

- **Wald test**: Estimate θ by MLE and check whether the distance $R\hat{\theta} - c$ is close to zero. This is the same idea that underlies t- and F-tests.

- **Likelihood Ratio test**: Estimate the model twice. Once without the restriction imposed (giving $\hat{\theta}$) and once with the null imposed (giving the constrained estimator $\tilde{\theta}$, that is $R\tilde{\theta} = c$). We then check whether the difference in log-likelihood values $lnL(\hat{\theta}) - lnL(\tilde{\theta})$ is close to zero. This is a similar idea to comparing the restricted RSS and the unrestricted RSS.

- **Lagrange Multiplier test**: Estimate the model with the null imposed (giving $\tilde{\theta}$) and check if the first order condition from the unrestricted model, evaluated at $\tilde{\theta}$, is significantly violated, i.e. check if $\frac{\partial L(\theta)}{\partial \theta}\Big|_{\tilde{\theta}}$ is close to zero.

Trinity of Tests - Graphically



Credit: Fox, 1997, p 570)

Trinity of Tests

- Each test looks at a different aspect of the likelihood function (width, height, slope), however, they are asymptotically equivalent.
- Choosing which to use, generally comes down to ease of computation.
- If the restricted and unrestricted are easy to compute, we can apply the LR test.
- If the restricted model is difficult to estimate, we can use the Wald test as it only requires the unrestricted parameter.
- If the unrestricted model is difficult to estimate, we can use the LM test since it only requires estimation under the null hypothesis.

Wald Test

- The standard R output includes significance tests on the coefficients using the Wald test:

$$z = \frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)} \rightarrow_d N(0,1) \text{ under } H_0.$$

- The standard errors can be obtained from the first (or second) order derivatives of the log likelihood. This uses the efficiency property of MLE:

$$A\hat{V}ar(\hat{\theta}) = \left(\sum \frac{\partial^2 lnL_i(\theta)}{\partial\theta\partial\theta'} \Big|_{\hat{\theta}} \right)^{-1}.$$

- Notice that the test is asymptotic (z-test) rather than finite sample (t-test).

Wald Test

- This is a specific case of the more general Wald test:

$$W = d' \left(\hat{Var}(d) \right)^{-1} d \rightarrow_d \chi^2_J \text{ under } H_0,$$

  where $d = R\hat{\beta} - c$, and J is the number of restrictions.

- This general form is analogous to the F-test since it can test joint hypotheses.

- Again, notice how this test is asymptotic (χ^2_J) rather than finite sample (F-test).

Wald Test

- The assumptions we need for this test are just the standard MLE regularity conditions.
- Then we have that $\hat{\beta}_{MLE} \rightarrow_d N(\beta, V)$. Hence

$$d = R\hat{\beta}_{MLE} - c \rightarrow_d N(0, R'VR) \text{ under } H_0.$$

- Since we use the standardised square for our test, this gives it a chi-squared distribution.

Likelihood Ratio Test

- The test statistic is given by

$$LR = -2\left(lnL_R - lnL_U\right) \rightarrow_d \chi^2_J \text{ under } H_0.$$

- This comes from Wilks' theorem which proves $-2ln\frac{L_R}{L_U} \rightarrow_d \chi^2_J$.
- The intuition is that if the null is not very restrictive, then there shouldn't be a big difference between the likelihood of the restricted and unrestricted fits.
- Note that lnL_R will always be smaller than lnL_U, so that $LR \geq 0$.

Lagrange Multiplier Test

- We can conduct maximum likelihood under any constraint as:

$$\max_{\beta} lnL(\beta) \; s.t. \; R\beta = c.$$

- This gives the following Lagrangian:

$$lnL(\beta) + \lambda'(R\beta - c)$$

- Recall that λ measures the degree of constraint of the restriction. If $\lambda = 0$, then the constraint is not binding.
- The LM test is given by

$$LM = \tilde{\lambda}' \left(\hat{Var}(\tilde{\lambda}) \right)^{-1} \tilde{\lambda} \rightarrow_d \chi^2_J \text{ under } H_0.$$

- Note that it can be shown that $\lambda = \frac{\partial lnL(\beta)}{\partial \beta}\Big|_{\tilde{\beta}}$. So we have an equivalent way to write the test using the first derivative evaluated at the restricted estimate (known as a **score test in this form**)

Example: Alcoholism and Marriage (alcohol)

- In this example, we use the LR and Wald test to look at the effect of your parents' alcoholism on your likelihood of getting married.
- In particular, we will test to see whether the effect of your father being an alcoholic is the same as the effect of your mother being an alcoholic, i.e. $H_0 : \beta_{\text{FathAlc}} = \beta_{\text{MothAlc}}$.
- Of course, because the outcome is a binary variable, we will be using a binary choice model, and so we cannot use finite-sample tests.
- To carry out the LR test, we need to work out how to impose the null hypothesis in order to construct the restricted model.

Example: Alcoholism and Marriage (alcohol)

- Note that if $\beta_{\text{FathAlc}} = \beta_{\text{MothAlc}}$, then

$$\begin{aligned} Y &= \beta_{\text{FathAlc}} X_{\text{FathAlc}} + \beta_{\text{MothAlc}} X_{\text{MothAlc}} + U \\ &= \beta_{\text{Alc}} \left(X_{\text{FathAlc}} + X_{\text{MothAlc}} \right) + U. \end{aligned}$$

- So, we can impose the null for the LR test by including the two alcoholism variables as a sum, rather than as two separate variables.

Example: Alcoholism and Marriage (alcohol)

- For the Wald test, notice that the null is a little different to our usual $\beta = 0$ type null.
- It would be great if we could transform the model somehow so that we can express our null in this simple kind of $\beta = 0$ form.
- We know we want $\beta_{\text{FathAlc}} - \beta_{\text{MothAlc}} = 0$. So we write

$$\begin{aligned} Y &= \beta_{\text{FathAlc}} X_{\text{FathAlc}} + \beta_{\text{MothAlc}} X_{\text{MothAlc}} + U \\ &= (\beta_{\text{FathAlc}} - \beta_{\text{MothAlc}}) X_{\text{FathAlc}} + \beta_{\text{MothAlc}} (X_{\text{MothAlc}} + X_{\text{FathAlc}}) + U. \end{aligned}$$

- So we regress Y on X_{FathAlc} and $(X_{\text{MothAlc}} + X_{\text{FathAlc}})$, and do a simple test of significance on the coefficient on X_{FathAlc}.

Summary

- We have seen why the linear probability model is not really adequate.
- We discussed why probit/logit are more appropriate.
- We also saw how these estimators are calculated using MLE and how to find the marginal effects at the average in R.
- We considered how to measure the goodness-of-fit of our model.
- And finally, we looked at testing when maximum likelihood estimators are used.