Week 6

Interpretation, Interaction, & Heteroskedastic Probit

Rich Frank
University of New Orleans
September 27, 2012

Interpretation

- How do we turn the output that our statistical software gives us and interpret it in a way that is meaningful?
- What effect is theoretically the most important for your argument?
- How can you interpret this estimated effect?

General Tests

- Likelihood Ratio
- Wald
- Lagrange Multiplier (“Score”)

- All asymptotically equivalent when H_0 is true.

- No consensus which is preferred in small samples.

- LR test only involves subtraction while Wald required matrix manipulation.

- I have not seen many instances of the Lagrange Multiplier used in the literature.

- All use nested models.

- A model (let's call it A) is considered nested in another (B) if you can create Model A by restricting Model B's parameters in some way.

- Usually this restriction is accomplished by setting a parameter (β_i) equals to zero.

- In Stata, you can do this by dropping this variable (X_i)from the model.

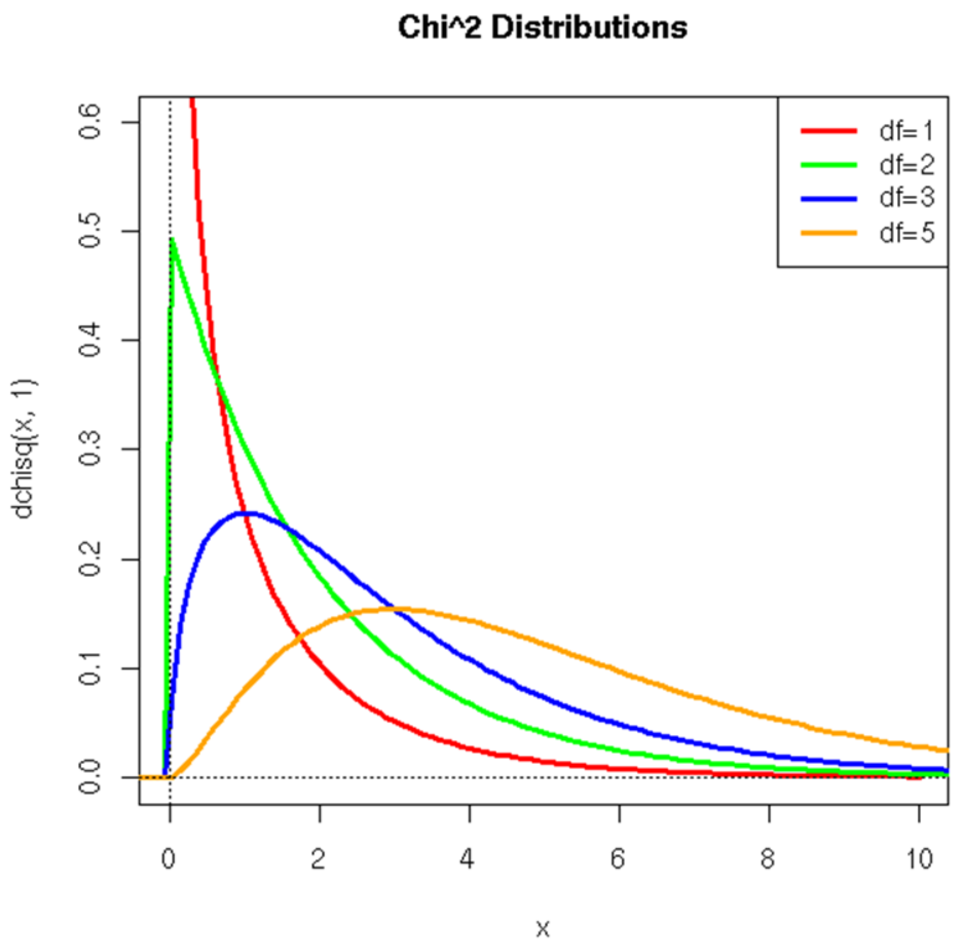
What is the Chi-Squared distribution?

- The χ^2-distribution (a PDF) with *n* degrees of freedom is given by the equation:

$$\Phi(x,v) = \frac{1}{2^{\frac{v}{2}}\Gamma(\frac{v}{2})}\left(e^{\frac{-x}{2}}\right)\left(x^{\frac{(v-2)}{2}}\right)$$

Where *v* = degrees of freedom

- For df over 100, the χ^2 approximates the normal distribution.



See also Long(1997: 89)

- All of these statistics are described well in Long (1997).
- Especially useful for understanding the intuition between the different approaches see Figure 4.2 (page 88).

Likelihood Ratio Test

- Does the likelihood change much under the null hypothesis versus our alternative?

- You actually have to run two models to run the lr test.

- Stata command: `lrtest`

- The likelihood ratio statistic is calculated as follows:

$$G^2 = (M_C |\, M_U) = 2\mathrm{lnL}(M_U) - 2\mathrm{lnL}(M_C)$$

Let’s try this in Stata and by hand

```
** LR Test of Ethnic Fractionalization using Fearon and Laitin (2003) **
. logit onset warl lgdpenl1 lpopl1 lmtnest ncontig Oil nwstate instab anocl deml ///
>         ethfrac relfrac

Iteration 0:   log likelihood = -534.46236
Iteration 1:   log likelihood = -489.16648
Iteration 2:   log likelihood = -477.13991
Iteration 3:   log likelihood = -476.83237
Iteration 4:   log likelihood = -476.83175
Iteration 5:   log likelihood = -476.83175

Logistic regression                               Number of obs   =       6326
                                                  LR chi2(12)     =     115.26
                                                  Prob > chi2     =     0.0000
Log likelihood = -476.83175                       Pseudo R2       =     0.1078

------------------------------------------------------------------------------
       onset |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        warl |  -.8826025   .3120193    -2.83   0.005    -1.494149   -.2710558
    lgdpenl1 |  -.6441048   .1241874    -5.19   0.000    -.8875076    -.400702
      lpopl1 |   .2566972   .0759579     3.38   0.001     .1078224    .4055719
     lmtnest |   .1816345   .0841573     2.16   0.031     .0166893    .3465797
     ncontig |   .3343231   .2800065     1.19   0.232    -.2144795    .8831258
         Oil |   .7628113   .2812081     2.71   0.007     .2116535    1.313969
     nwstate |   1.656059   .3452964     4.80   0.000       .97929    2.332827
      instab |   .4926794   .2464971     2.00   0.046      .009554    .9758048
       anocl |   .6148623   .2406123     2.56   0.011     .1432709    1.086454
        deml |   .0955259   .3141988     0.30   0.761    -.5202923    .7113441
     ethfrac |   .2201434    .367535     0.60   0.549    -.5002119    .9404987
     relfrac |  -.0441585   .5111071    -0.09   0.931     -1.04591    .9575929
       _cons |  -2.749393   1.191852    -2.31   0.021    -5.085381   -.4134058
------------------------------------------------------------------------------

. estimates store A
```

```
 ** Restricted Model does not include ethfrac **
. logit onset warl lgdpenl1 lpopl1 lmtnest ncontig Oil nwstate instab anocl deml ///
>         relfrac

Iteration 0:   log likelihood = -534.46236
Iteration 1:   log likelihood = -489.23147
Iteration 2:   log likelihood = -477.31961
Iteration 3:   log likelihood = -477.01331
Iteration 4:   log likelihood =  -477.0127
Iteration 5:   log likelihood =  -477.0127

Logistic regression                               Number of obs   =       6326
                                                  LR chi2(11)     =     114.90
                                                  Prob > chi2     =     0.0000
Log likelihood =  -477.0127                       Pseudo R2       =     0.1075

------------------------------------------------------------------------------
       onset |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        warl |  -.8691268   .3109841    -2.79   0.005    -1.478644   -.2596092
    lgdpenl1 |  -.6582621   .1213198    -5.43   0.000    -.8960446   -.4204795
      lpopl1 |   .2619946   .0745104     3.52   0.000      .115957    .4080323
     lmtnest |   .1758016   .0836563     2.10   0.036     .0118382     .339765
     ncontig |   .3372547   .2787012     1.21   0.226    -.2089895    .8834989
         Oil |   .7898462   .2776031     2.85   0.004     .2457542    1.333938
     nwstate |   1.666944   .3449221     4.83   0.000      .990909    2.342979
      instab |   .4936941    .246361     2.00   0.045     .0108354    .9765529
       anocl |   .6196298   .2405215     2.58   0.010     .1482162    1.091043
        deml |   .0916092   .3145002     0.29   0.771    -.5247999    .7080183
     relfrac |   .0180241   .5006194     0.04   0.971     -.963172    .9992202
       _cons |  -2.615902   1.162481    -2.25   0.024    -4.894322   -.3374821
------------------------------------------------------------------------------

. estimates store B
```

```
. lrtest A

Likelihood-ratio test                                 LR chi2(1)  =      0.36
(Assumption: B nested in A)                           Prob > chi2 =    0.5475
```

- We can also easily do this by hand:

$$G^2 = (M_C | M_U) = 2\text{lnL}(M_U) – 2\text{lnL}(M_C)$$
$$= 2(-477.0127) – 2(-476.83175)$$
$$= (-954.0254) – (953.6635)$$
$$= .3619$$

- This null finding is not surprising because the beta for ethnic fractionalization is not significant. Let's try it with "New State."

```
logit onset warl lgdpenl1 lpopl1 lmtnest ncontig Oil nwstate instab anocl deml ///
>          ethfrac relfrac

Iteration 0:   log likelihood = -534.46236
Iteration 1:   log likelihood = -489.16648
Iteration 2:   log likelihood = -477.13991
Iteration 3:   log likelihood = -476.83237
Iteration 4:   log likelihood = -476.83175
Iteration 5:   log likelihood = -476.83175

Logistic regression                               Number of obs   =       6326
                                                  LR chi2(12)     =     115.26
                                                  Prob > chi2     =     0.0000
Log likelihood = -476.83175                       Pseudo R2       =     0.1078

------------------------------------------------------------------------------
       onset |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        warl |  -.8826025   .3120193    -2.83   0.005    -1.494149   -.2710558
    lgdpenl1 |  -.6441048   .1241874    -5.19   0.000    -.8875076    -.400702
      lpopl1 |   .2566972   .0759579     3.38   0.001     .1078224    .4055719
     lmtnest |   .1816345   .0841573     2.16   0.031     .0166893    .3465797
     ncontig |   .3343231   .2800065     1.19   0.232    -.2144795    .8831258
         Oil |   .7628113   .2812081     2.71   0.007     .2116535    1.313969
     nwstate |   1.656059   .3452964     4.80   0.000       .97929    2.332827
      instab |   .4926794   .2464971     2.00   0.046      .009554    .9758048
       anocl |   .6148623   .2406123     2.56   0.011     .1432709    1.086454
        deml |   .0955259   .3141988     0.30   0.761    -.5202923    .7113441
     ethfrac |   .2201434    .367535     0.60   0.549    -.5002119    .9404987
     relfrac |  -.0441585   .5111071    -0.09   0.931     -1.04591    .9575929
       _cons |  -2.749393   1.191852    -2.31   0.021    -5.085381   -.4134058
------------------------------------------------------------------------------

. estimates store A
```

```
 ** Restricted Model does not include New State **
. logit onset warl lgdpenl1 lpopl1 lmtnest ncontig Oil  instab anocl deml ///
>         ethfrac relfrac

Iteration 0:   log likelihood = -534.46236
Iteration 1:   log likelihood = -507.63265
Iteration 2:   log likelihood = -485.83417
Iteration 3:   log likelihood =  -485.7061
Iteration 4:   log likelihood = -485.70598
Iteration 5:   log likelihood = -485.70598

Logistic regression                               Number of obs   =       6326
                                                  LR chi2(11)     =      97.51
                                                  Prob > chi2     =     0.0000
Log likelihood = -485.70598                       Pseudo R2       =     0.0912

------------------------------------------------------------------------------
       onset |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        warl |  -.9659943   .3115093    -3.10   0.002    -1.576541   -.3554473
    lgdpenl1 |  -.7129984   .1223772    -5.83   0.000    -.9528532   -.4731436
      lpopl1 |   .2159782   .0754739     2.86   0.004      .068052    .3639043
     lmtnest |   .1804607    .084162     2.14   0.032     .0155062    .3454152
     ncontig |    .430106    .278506     1.54   0.123    -.1157557    .9759678
         Oil |   .8182815   .2775055     2.95   0.003     .2743808    1.362182
      instab |   .3143875   .2407495     1.31   0.192    -.1574728    .7862479
       anocl |   .7433193   .2372684     3.13   0.002     .2782818    1.208357
        deml |   .2285169   .3150897     0.73   0.468    -.3890476    .8460814
     ethfrac |   .2775607   .3646804     0.76   0.447    -.4371997    .9923211
     relfrac |    .030431   .5085138     0.06   0.952    -.9662377      1.0271
       _cons |  -1.867458   1.162471    -1.61   0.108    -4.145859     .410944
------------------------------------------------------------------------------

. estimates store B
```

```
. lrtest A

Likelihood-ratio test                                 LR chi2(1)  =     17.75
(Assumption: B nested in A)                           Prob > chi2 =    0.0000
```

We can also calculate the LR test by hand:

$$G^2 = (M_C | M_U) = 2\text{lnL}(M_U) - 2\text{lnL}(M_C)$$

$$= 2(-476.83175) - 2(-485.70598)$$

$$= (-953.6635) - (-971.41196)$$

$$= 17.74846$$

Wald statistic

- Are the estimated parameters far away from what they would be under the null hypothesis?

- The Wald statistic is calculated as follows:

$$W = [\mathbf{Q}\hat{\boldsymbol{\beta}} - \mathbf{r}]'[\mathbf{Q}\widehat{Var}(\hat{\boldsymbol{\beta}})\mathbf{Q}']^{-1}[\mathbf{Q}\hat{\boldsymbol{\beta}} - \mathbf{r}]$$

- W is also distributed chi-square.

- In Stata: `test`

```
. ** Wald **
. ** Using Fearon and Laitin Table 1 Model #3 **
. qui logit onset warl lgdpenl1 lpopl1 lmtnest ncontig Oil nwstate instab anocl deml
///
>   ethfrac relfrac, robust cluster(ccode)

. test

 ( 1)  [onset]warl = 0
 ( 2)  [onset]lgdpenl1 = 0
 ( 3)  [onset]lpopl1 = 0
 ( 4)  [onset]lmtnest = 0
 ( 5)  [onset]ncontig = 0
 ( 6)  [onset]Oil = 0
 ( 7)  [onset]nwstate = 0
 ( 8)  [onset]instab = 0
 ( 9)  [onset]anocl = 0
 (10)  [onset]deml = 0
 (11)  [onset]ethfrac = 0
 (12)  [onset]relfrac = 0

           chi2( 12) =  142.13
         Prob > chi2 =    0.0000
```

- We can also test one parameter:

```
. test warl=0

 ( 1)  [onset]warl = 0

           chi2(  1) =   12.62
         Prob > chi2 =    0.0004
```

- Or two jointly:

```
. test (warl=0)(lpopl1=0)

 ( 1)  [onset]warl = 0
 ( 2)  [onset]lpopl1 = 0

           chi2(  2) =   27.37
         Prob > chi2 =    0.0000
```

Lagrange Multiplier Test

- If I had a less restrictive likelihood function, would its derivative be close to zero here at the restricted ML estimate?

- In Stata: `enumopt, testomit`
  - These are add-ons you have to install
  - Type "find it enumopt" and "find it testomit" to install

```
.qui logit onset  lgdpenl1 lmtnest ncontig Oil nwstate instab anocl
deml ///
        ethfrac relfrac, robust cluster(ccode)
        predict test, score

.testomit warl lpopl1, score(test)

.logit: score tests for omitted variables

Term                 |    score  df      p
---------------------+----------------------
    warl (as factor) |     0.00   1   1.0000
              lpopl1 |     0.00   1   1.0000
---------------------+----------------------
   simultaneous test |     0.00   2   1.0000
---------------------+----------------------
            adjusted for clustering on ccode
```

- There are a number of other ways of interpreting estimated coefficients.

- For a useful and much more in depth guide see Ch. 3 and 4 of Long and Freese (2005).

A number of other statistics can be calculated

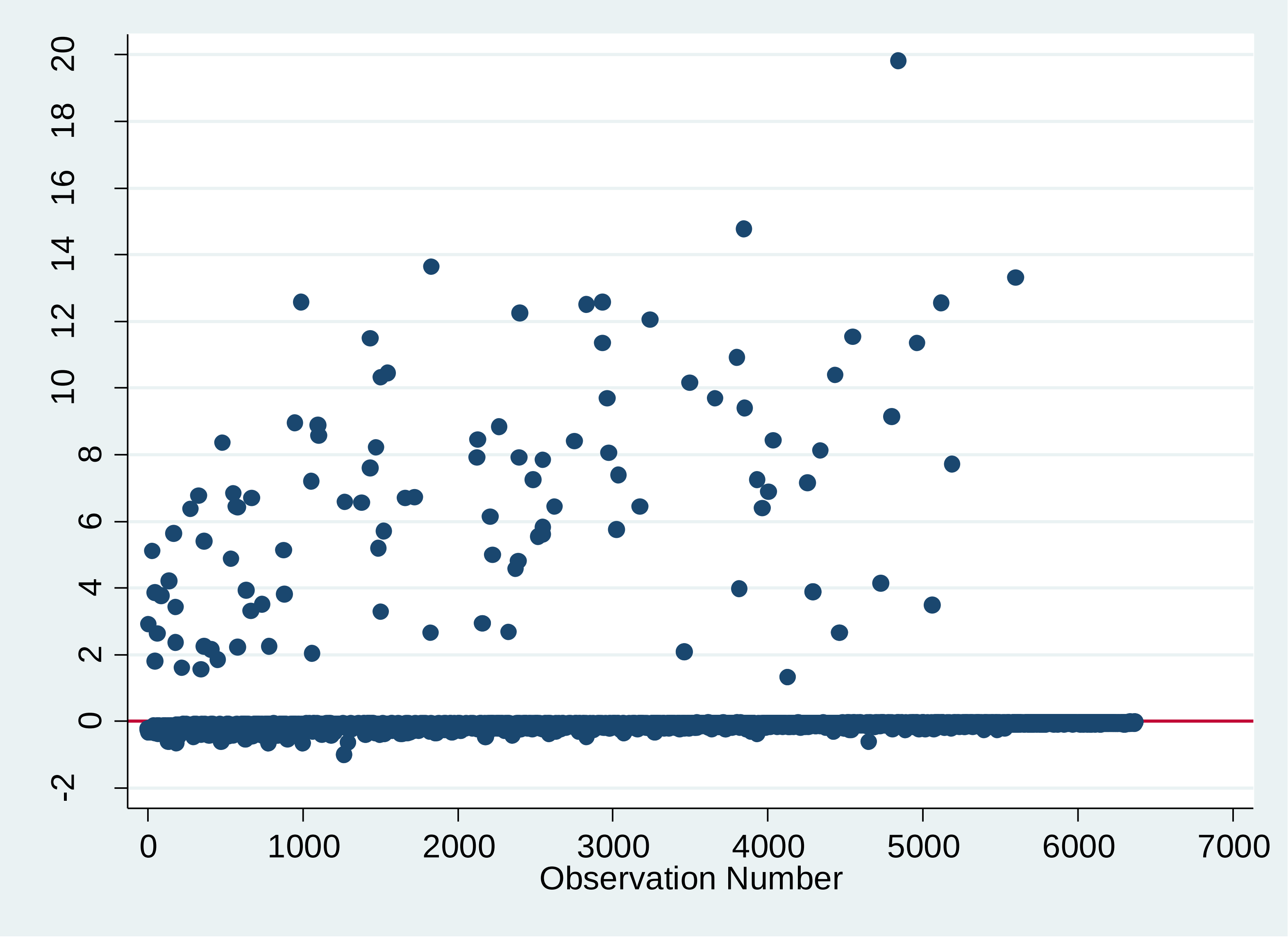
```
. ** Long and Freese Spost **
. quietly logit onset warl lgdpenl1 lpopl1 lmtnest ncontig Oil
nwstate instab anocl deml ///
>         ethfrac relfrac, robust cluster(ccode) nolog
. estimates store model1
. quietly fitstat, save
. end of do-file
. quietly logit onset warl lgdpenl1 lpopl1 lmtnest ncontig Oil
instab anocl deml ///
>         ethfrac relfrac, robust cluster(ccode) nolog

. estimates store model2

. fitstat, diff
```

```
Measures of Fit for logit of onset

                              Current            Saved         Difference
Model:                          logit            logit
N:                               6326             6326                  0
Log-Lik Intercept Only       -534.462         -534.462              0.000
Log-Lik Full Model           -485.706         -476.832             -8.874
D                             971.412(6314)    953.664(6313)       17.748(1)
LR                             97.513(11)      115.261(12)         17.748(1)
Prob > LR                       0.000            0.000              0.000
McFadden's R2                   0.091            0.108             -0.017
McFadden's Adj R2               0.069            0.084             -0.015
ML (Cox-Snell) R2               0.015            0.018             -0.003
Cragg-Uhler(Nagelkerke) R2      0.098            0.116             -0.018
McKelvey & Zavoina's R2         0.210            0.221             -0.012
Efron's R2                      0.028            0.041             -0.014
Variance of y*                  4.164            4.225             -0.062
Variance of error               3.290            3.290              0.000
Count R2                        0.983            0.983              0.000
Adj Count R2                    0.000            0.000              0.000
AIC                             0.157            0.155              0.002
AIC*n                         995.412          979.664             15.748
BIC                        -54291.389       -54300.385              8.996
BIC'                           -1.236          -10.232              8.996
BIC used by Stata            1076.441         1067.445              8.996
AIC used by Stata             995.412          979.664             15.748

Difference of    8.996 in BIC' provides strong support for saved model.

Note: p-value for difference in LR is only valid if models are nested.
```

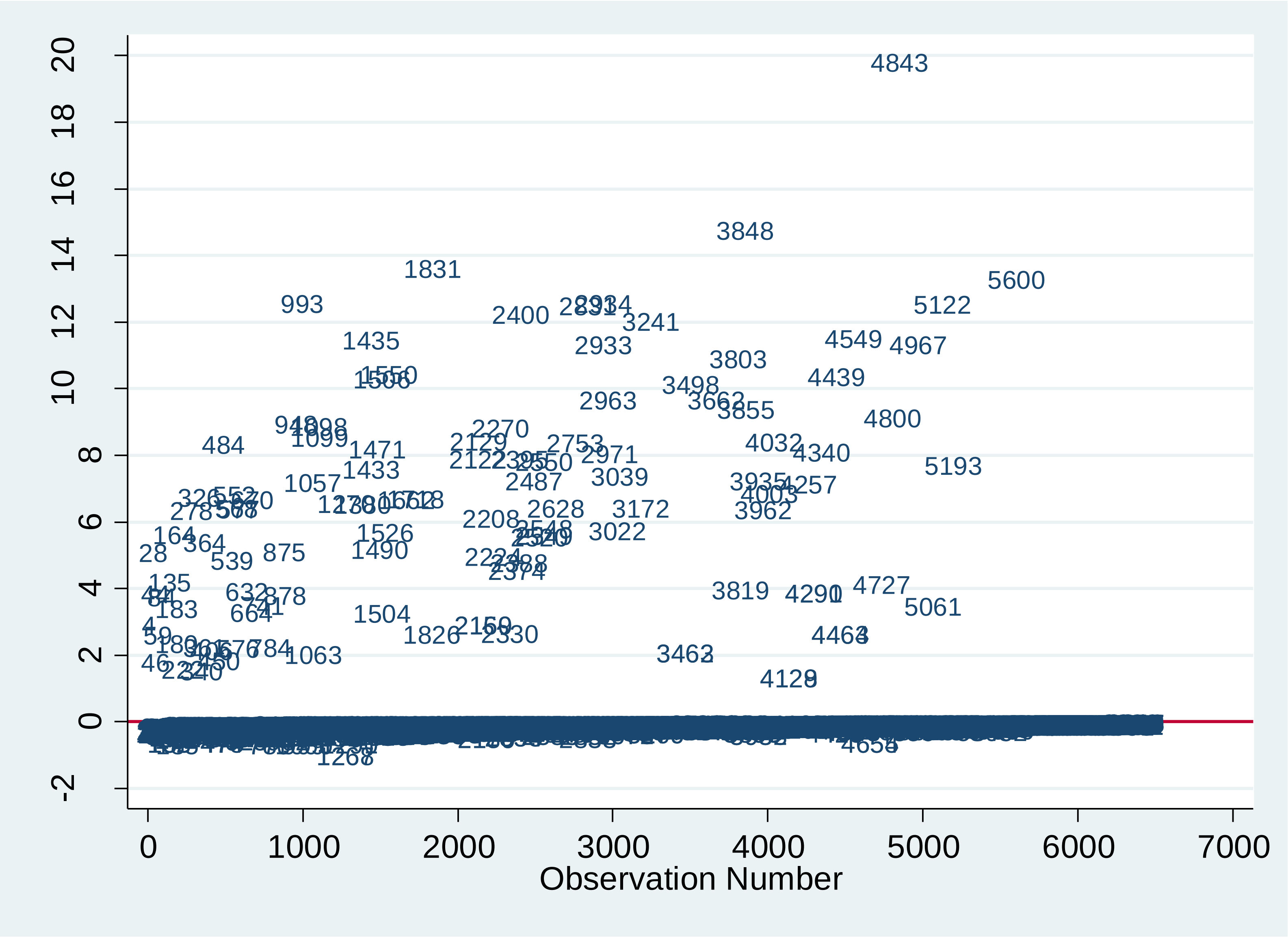
- Truly understanding your data, your model, and how the two are connected is a time-intensive and subtle process that is often overlooked by busy researchers.

- However, the better you know your data, your model, and your link function, the more confidence you can have in your findings (or try and figure out why you are not getting any results).

Let's start by taking a look at outliers



By observation number



Stata code for these two graphs

```
** Residuals **
** From Long and Freese page 147-149) **

quietly logit onset warl lgdpenl1 lpopl1 lmtnest ncontig Oil nwstate
instab anocl deml ///
        ethfrac relfrac, robust cluster(ccode) nolog

predict rstd, rs
label var rstd "Standardized Residual"
sort lgdpenl1, stable

generate index=_n
label var index "Observation number"
graph twoway scatter rstd index, xlabel(0(1000)7000) ylabel(-2(2)20) ///
xtitle("Observation Number") yline(0) msymbol(0h)

graph twoway scatter rstd index, xlabel(0(1000)7000) ylabel(-2(2)20) ///
        xtitle("Observation Number") yline(0) msymbol(none)
///
        mlabel(index) mlabposition(0)
```

This lets us find and understand outliers

```
. **Then you can list influential observations **
. list in 4844, noobs

  +-----------------------------------------------------------------------------------------------------+
  | ccode | country |  cname | cmark | year | wars | war | warl | onset | ethonset | durest | aim  |
  |   352 |  CYPRUS | CYPRUS |     0 | 1979 |    0 |   0 |    0 |     0 |        0 |      . |   .  |
  |-----------------------------------------------------------------------------------------------------|
  | casename | ended | ethwar | waryrs | pop  |     lpop  | polity2  | gdpen  | gdptype  | gdpenl  |
  |          |     . |      . |        | 623  | 6.434546  |      10  | 5.091  |       0  |  4.697  |
  |----------+-----------------------------------------------------+-------------------+---------|
  | lgdpenl1 |   lpopl1 |    region | western  | eeurop  | lamerica  | ssafrica  | asia  | nafrme  |
  | 8.454679 | 6.426488 | n. africa |       0  |      0  |        0  |        0  |    0  |      1  |
  |-----------------------------------------------------------------------------------------------------|
  | colbrit | colfra | mtnest |  lmtnest | elevdiff | Oil | ncontig |  ethfrac |       ef | plural |
  |       1 |      0 |      9 | 2.302585 |     1952 |   0 |       0 | .3485829 | .3592001 |    .78 |
  |-----------------+------------------+---------+----------------------------------------------------|
  | second | numlang | relfrac | plurrel | minrelpc | muslim | nwstate | polity2l | instab | anocl |
  |    .18 |       2 |   .3576 |      78 |       18 |     18 |       0 |       10 |      0 |     0 |
  |----------------------------------------+---------+----------------------------------------------------|
  | deml | empeth~c | empwarl | emponset | empgdp~l | emplpopl | emplmt~t  | empnco~g  | empol~2l  |
  |    1 | .3485829 |       0 |        0 |    4.697 | 6.426488 | 2.302585  |        0  |      10   |
  |-----------------------------------------------------------------------------------------------------|
  | sdwars | sdonset | colwars  | colonset  | cowwars  | cowonset  | cowwarl  | sdwarl  | colwarl  |
  |      0 |       0 |       0  |        0  |       0  |        0  |       0  |      0  |       0  |
  |-----------------+---------------------+----------------------------------------------------|
  |     cyr | _est_A | _est_B  |      test  | _F1_1  | _F1_2  | _est_m~1  | _est_m~2  |      rstd  |
  | 3521979 |      1 |      1  | -.0025459  |     1  |     0  |        1  |        1  | -.0505275  |
  |-----------------------------------------------------------------------------------------------------|
  |                                          index                                                      |
  |                                           4844                                                      |
  +-----------------------------------------------------------------------------------------------------+
```

```
. list country year onset warl Oil ethfrac if rstd>12 &
rstd~=.

       +--------------------------------------------------------+
       |     country   year   onset   warl   Oil     ethfrac |
       |--------------------------------------------------------|
  993. |     SOMALIA   1991       1      1     0    .0767429 |
 1831. |    PARAGUAY   1947       1      0     0    .1449224 |
 2400. |   COSTARICA   1948       1      0     0      .07109 |
 2831. |      JORDAN   1970       1      0     0    .0466286 |
 2934. | AFGHANISTAN   1992       1      1     0     .658281 |
       |--------------------------------------------------------|
 3241. |   SRI LANKA   1987       1      1     0    .4670414 |
 3848. |   NICARAGUA   1978       1      0     0    .1793287 |
 4843. |      CYPRUS   1974       1      0     0    .3485829 |
 5122. |   ARGENTINA   1973       1      0     0    .3073192 |
 5600. |          UK   1969       1      0     0    .3254545 |
       +--------------------------------------------------------+
```

- We can also look for influential observations using Cook's statistic:

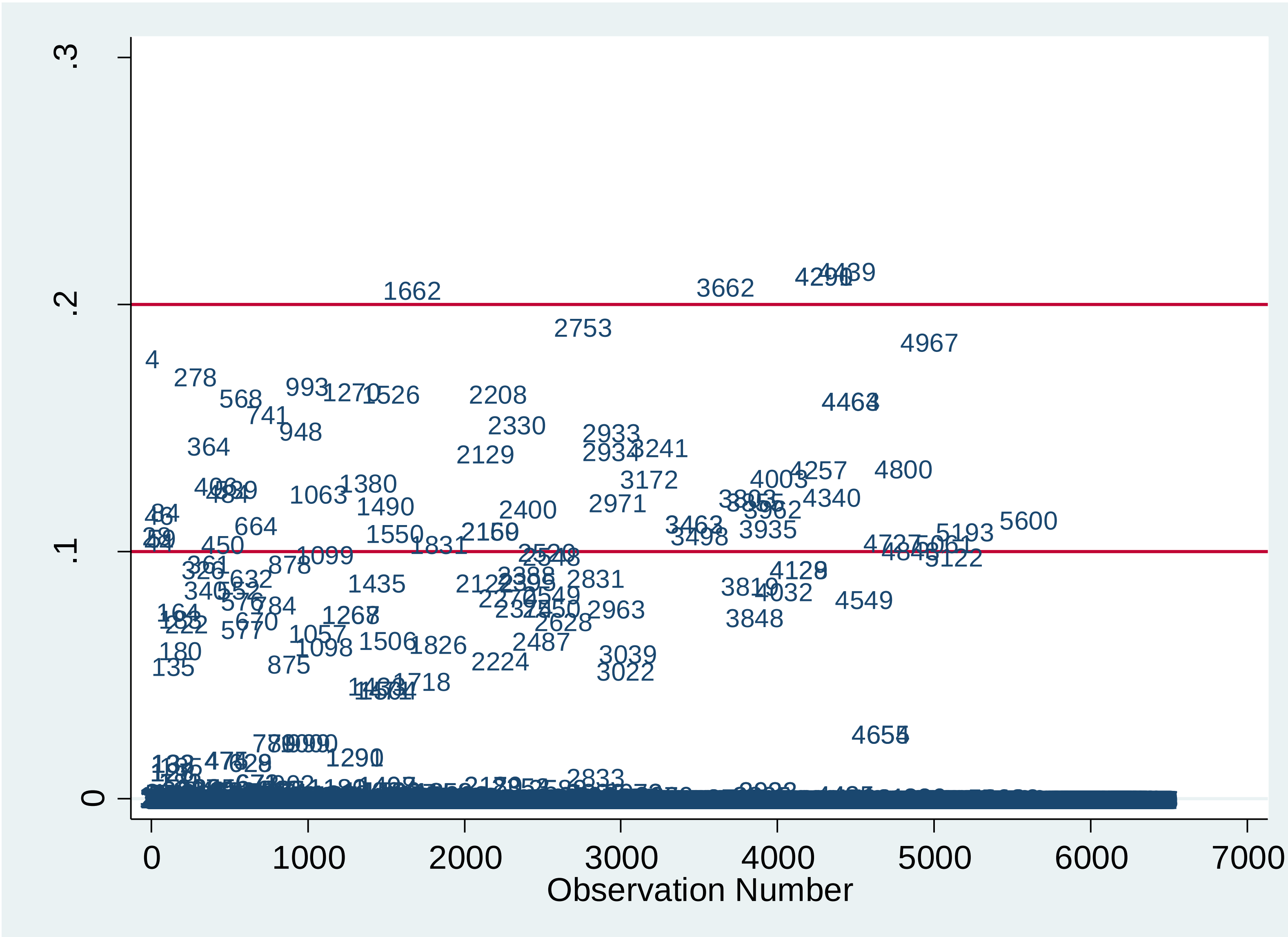
$$P_i = \frac{r^2 h_{ii}}{(1 - h_{ii})^2}$$

Where:

$$h_{ii} = \hat{\pi}_i(1 - \hat{\pi}_i)\, x_i \widehat{Var}(\hat{\beta}) x_i'$$

- This measures the effect of removing observation *i* on the vector $\hat{\beta}$.

Visually



.3
.2
.1
0
0
1000
2000
3000
4000
5000
6000
7000
Observation Number

Then in more detail

```
. list country year onset warl Oil ethfrac if cook>.2 &
cook~=.

      +-----------------------------------------------------+
      |    country   year   onset   warl   Oil    ethfrac |
      |-----------------------------------------------------|
1662. | BANGLADESH   1976       1      0     0   .0049751 |
3662. |    LEBANON   1958       1      0     0   .1348051 |
4290. |    MOLDOVA   1991       0      0     0   .5515695 |
4291. |    MOLDOVA   1992       1      0     0   .5515695 |
4439. |    LEBANON   1975       1      0     0   .1348051 |
      +-----------------------------------------------------+
```

Stata code for above figures

```
** INFLUENTIAL CASES **
predict cook, dbeta
label var cook "Cook's Statistic"
graph twoway scatter cook index,  xlabel(0(1000)7000)  ylabel(0(.1).3) ///
        xtitle("Observation Number") yline(.1 .2)              ///
        msymbol(none) mlabel(index) mlabposition(0)

list country year onset warl Oil ethfrac if cook>.2 & cook~=.
```

Least Likely Observations

- You can also look for those cases that are least likely (i.e. have large residuals) given your model.

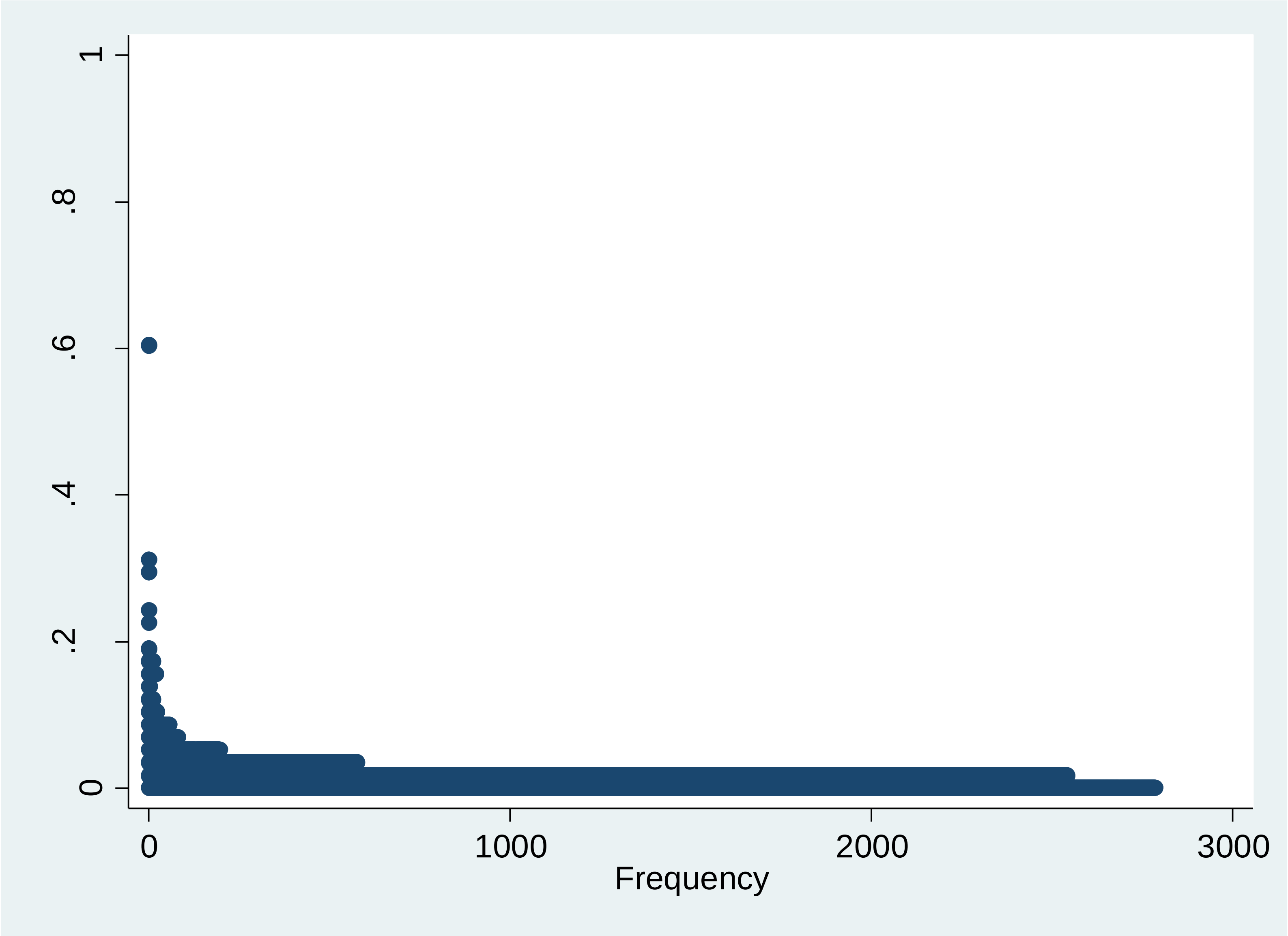
```
. qui logit onset warl lgdpenl1 lpopl1 lmtnest ncontig Oil nwstate instab anocl deml ///
>         ethfrac relfrac, robust cluster(ccode) nolog

. leastlikely country year warl lpopl1

Outcome: 0

       +----------------------------------------------+
       |     Prob     country   year   warl     lpopl1 |
       |----------------------------------------------|
 127.  | .3946847   INDONESIA   1949      0   11.22257 |
 185.  | .6983526   INDONESIA   1952      0   11.24785 |
 902.  |  .765673   INDONESIA   1962      0   11.47718 |
1267.  | .6832954    PAKISTAN   1947      0   11.18728 |
1268.  | .6832954    PAKISTAN   1948      0   11.18728 |
       +----------------------------------------------+

Outcome: 1

       +----------------------------------------------+
       |     Prob     country   year   warl     lpopl1 |
       |----------------------------------------------|
 993.  |  .006282     SOMALIA   1991      1   8.958411 |
1831.  | .0053564    PARAGUAY   1947      0   7.150702 |
3848.  | .0045628   NICARAGUA   1978      0   7.852828 |
4843.  | .0025399      CYPRUS   1974      0   6.415097 |
5600.  | .0056291          UK   1969      0   10.91792 |
       +----------------------------------------------+
```

Predicting Pr(Y)

- Again, what we are usually interested in is how our model predicts the probability of our Y over our main independent variable of interest.

- However, before we do that it can also be informative to get a rough sense of the range of our predicted latent variable.

As you can see, civil war data is highly skewed



```
*** Predicted Probabilities with PREDICT **
.qui logit onset warl lgdpenl1 lpopl1 lmtnest ncontig Oil nwstate
instab anocl deml ///
        ethfrac relfrac, robust cluster(ccode) nolog
.predict prlogit
.label var prlogit "Logit: Pr(DV)"
.dotplot prlogit, ylabel(0(.2)1)
```

- However, there are still a few cases with an over 50% chance of conflict.

- As most of our readings indicate, there is little difference between logit and probit estimates.

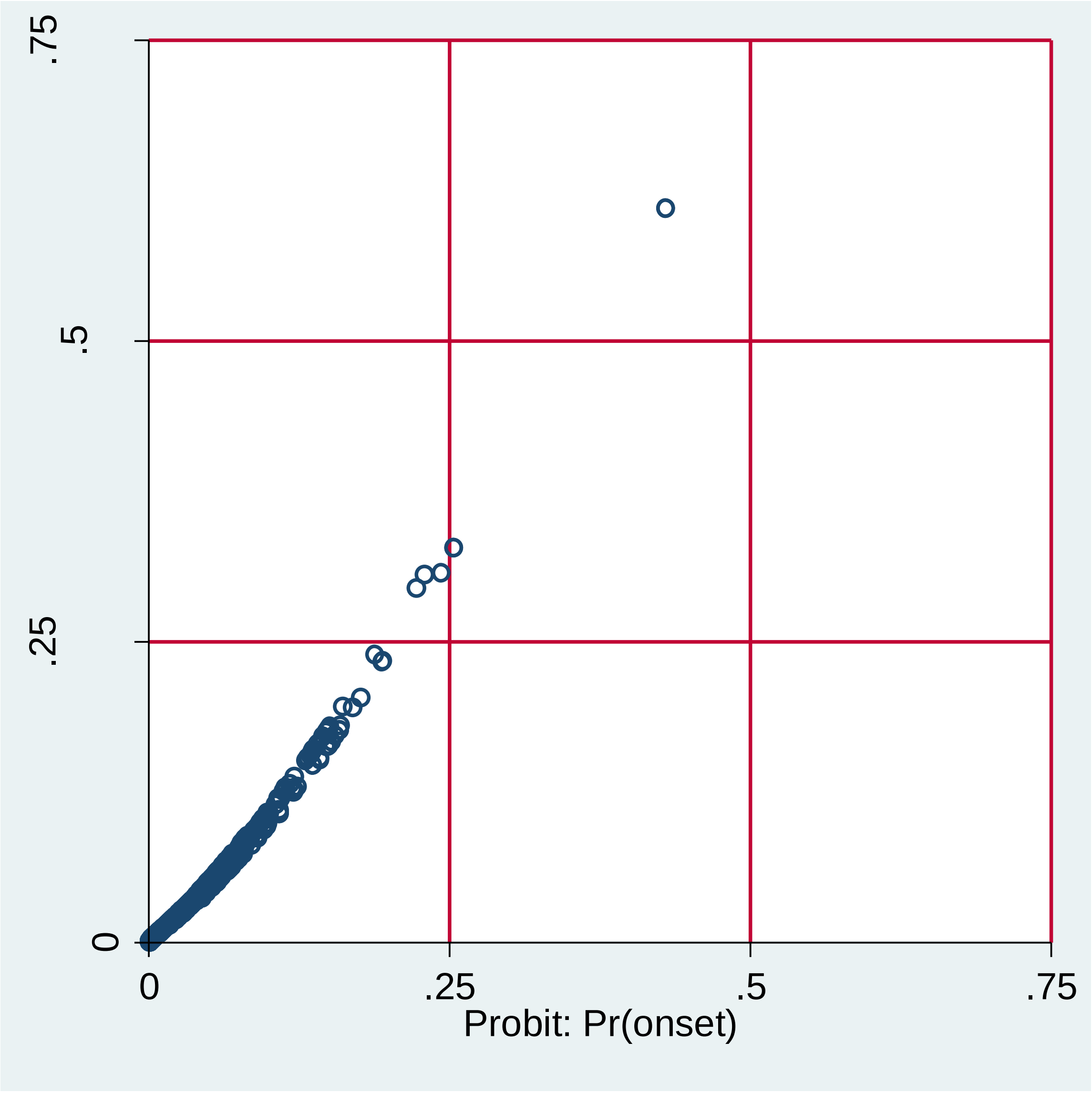
- However, in cases where the data are heavily skewed, it is possible.

Estimating Probit and Logit

```
qui logit onset warl lgdpenl1 lpopl1 lmtnest ncontig Oil
nwstate instab anocl deml ///
       ethfrac relfrac, robust cluster(ccode) nolog
predict prlogit
label var prlogit "Logit: Pr(onset)"
qui probit onset warl lgdpenl1 lpopl1 lmtnest ncontig Oil
nwstate instab anocl deml ///
       ethfrac relfrac, robust cluster(ccode) nolog
predict prprobit
label var prprobit "Probit: Pr(onset)"
pwcorr prlogit prprobit

* graphing predicted probabilities from logit and probit
graph twoway scatter prlogit prprobit, ///
    xlabel(0(.25).75) ylabel(0(.25).75) ///
    xline(.25(.25).75) yline(.25(.25).75) ///
    plotregion(margin(zero)) msymbol(Oh) ///
    ysize(4.0413) xsize(4.0413)
```

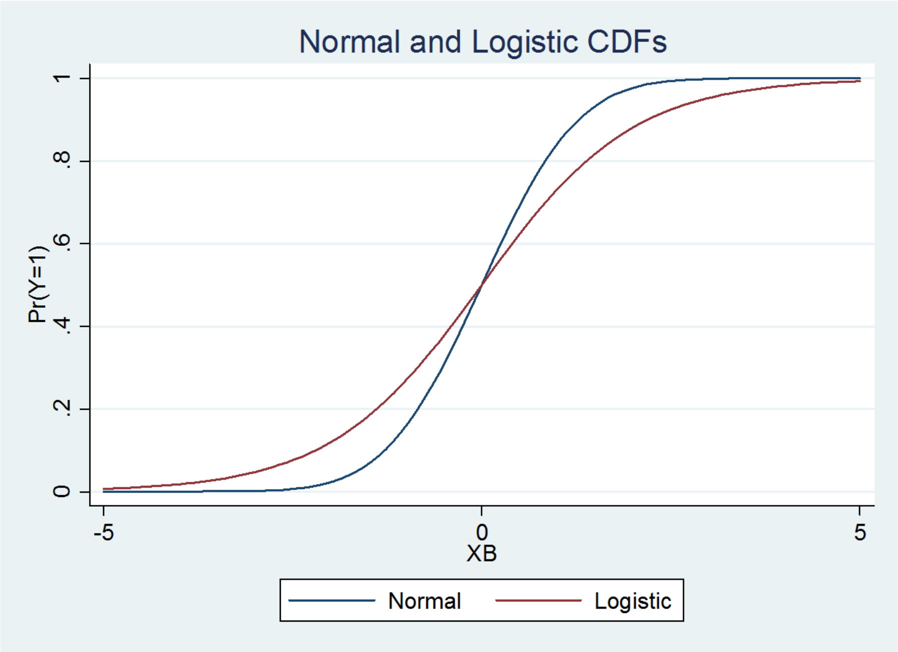
- They are still highly correlated.

```
. pwcorr prlogit prprobit

             |  prlogit prprobit
-------------+------------------
     prlogit |   1.0000
    prprobit |   0.9880   1.0000

. tab onset

1 for civil |
  war onset |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      6,499       98.34       98.34
          1 |        110        1.66      100.00
------------+-----------------------------------
      Total |      6,609      100.00
```



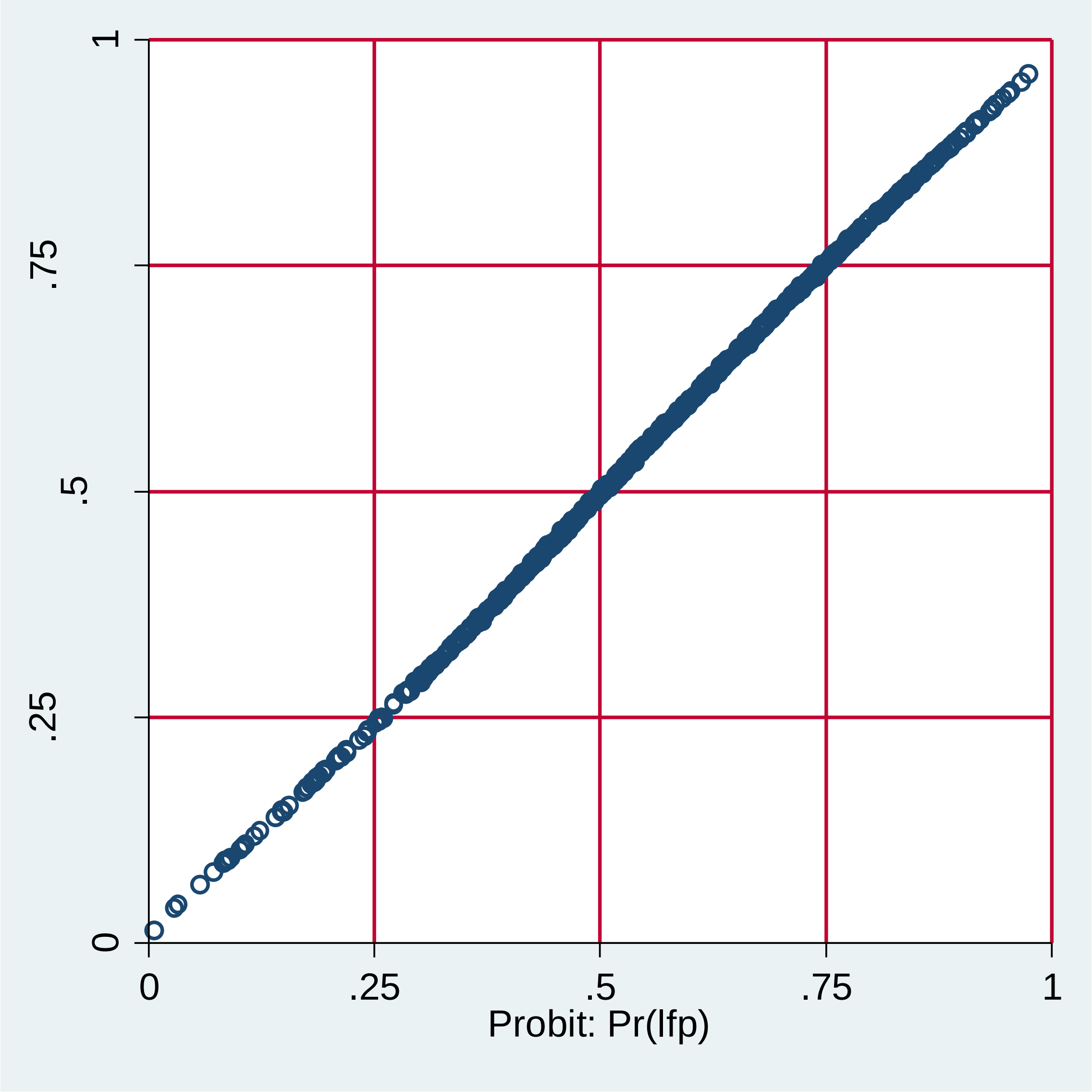
.75
.5
.25
0
0
.25
.5
.75
Probit: Pr(onset)

- So Logit gives slightly higher predictions of Pr(Y).
- Why?
- The slight difference in the normal and logistic CDFs we saw last week.

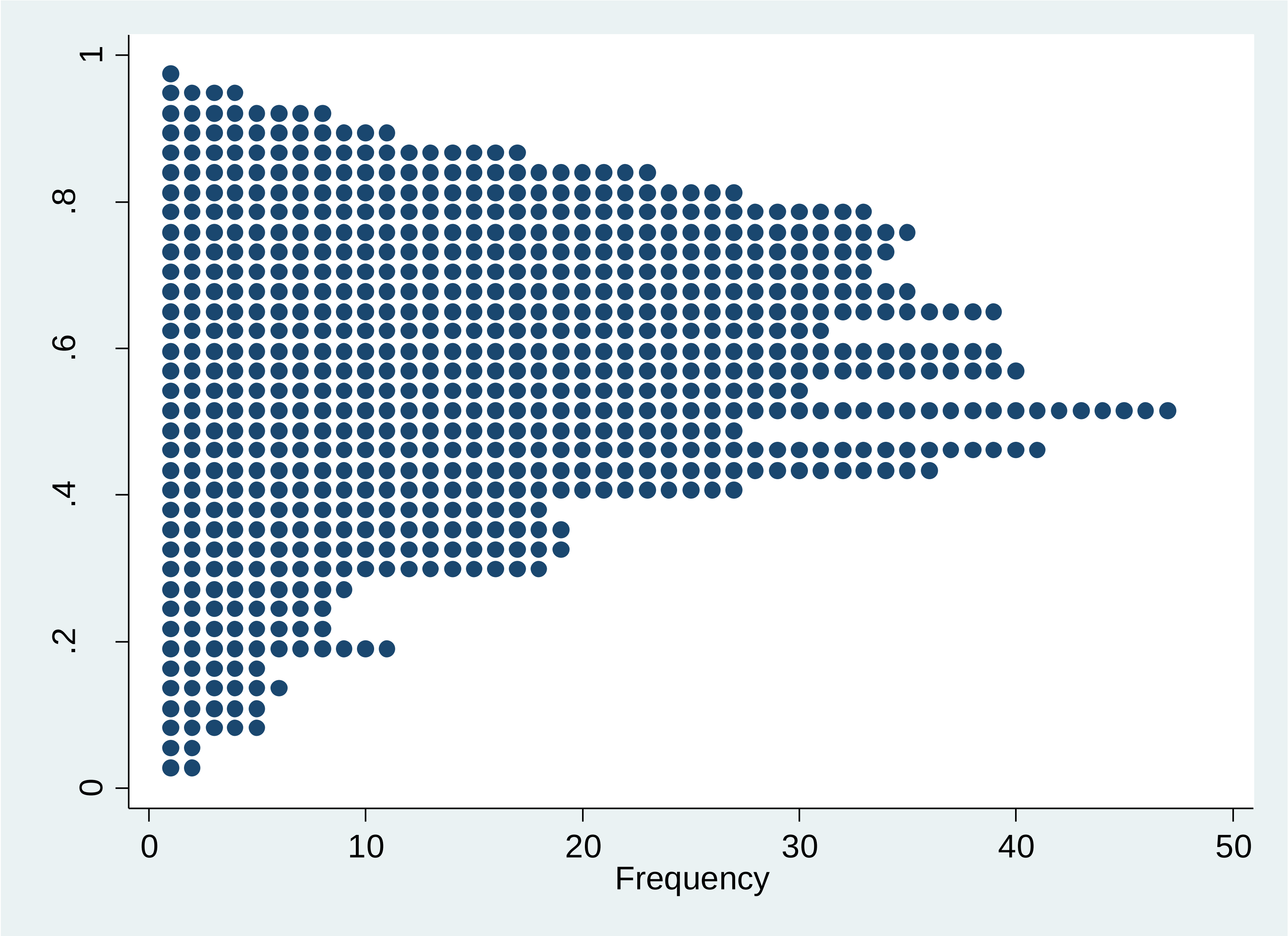


- Let's try it on less skewed data.
- Labor force participation data from Long (1997)

```
. pwcorr prlogit prprobit

             |  prlogit prprobit
-------------+------------------
     prlogit |   1.0000
    prprobit |   0.9998   1.0000


. tab lfp

 Paid Labor |
     Force: |
 1=yes 0=no |      Freq.     Percent        Cum.
------------+-----------------------------------
    NotInLF |        325       43.16       43.16
       inLF |        428       56.84      100.00
------------+-----------------------------------
      Total |        753      100.00
```



0
.25
.5
.75
1
0
.25
.5
.75
1
Probit: Pr(lfp)


1
.8
.6
.4
.2
0
0
10
20
30
40
50
Frequency

- As you can see the DV is more balanced.

- Towards the end of class we are going to talk a bit about ways of dealing with skewed variables.

Now we all remember out-of-sample predictions

- As described last week, what we are really interested in is understanding what effect our X variable of interest has on Y for different values of X while holding all other variables constant.

Let’s move to the meat of interpretation

- How do we really get a sense of the relation between my main IV and DV?
- After running a regression there are a number of ways of isolating the effect of my IV on my DV.
- Several ways can be implemented in Spost, an add-on to Stata written by Long and Freese.

- Let's say that I am most interested in capturing the probability of conflict as GDP varies.

- What you can do is look at the relationship graphically (I like pictures).

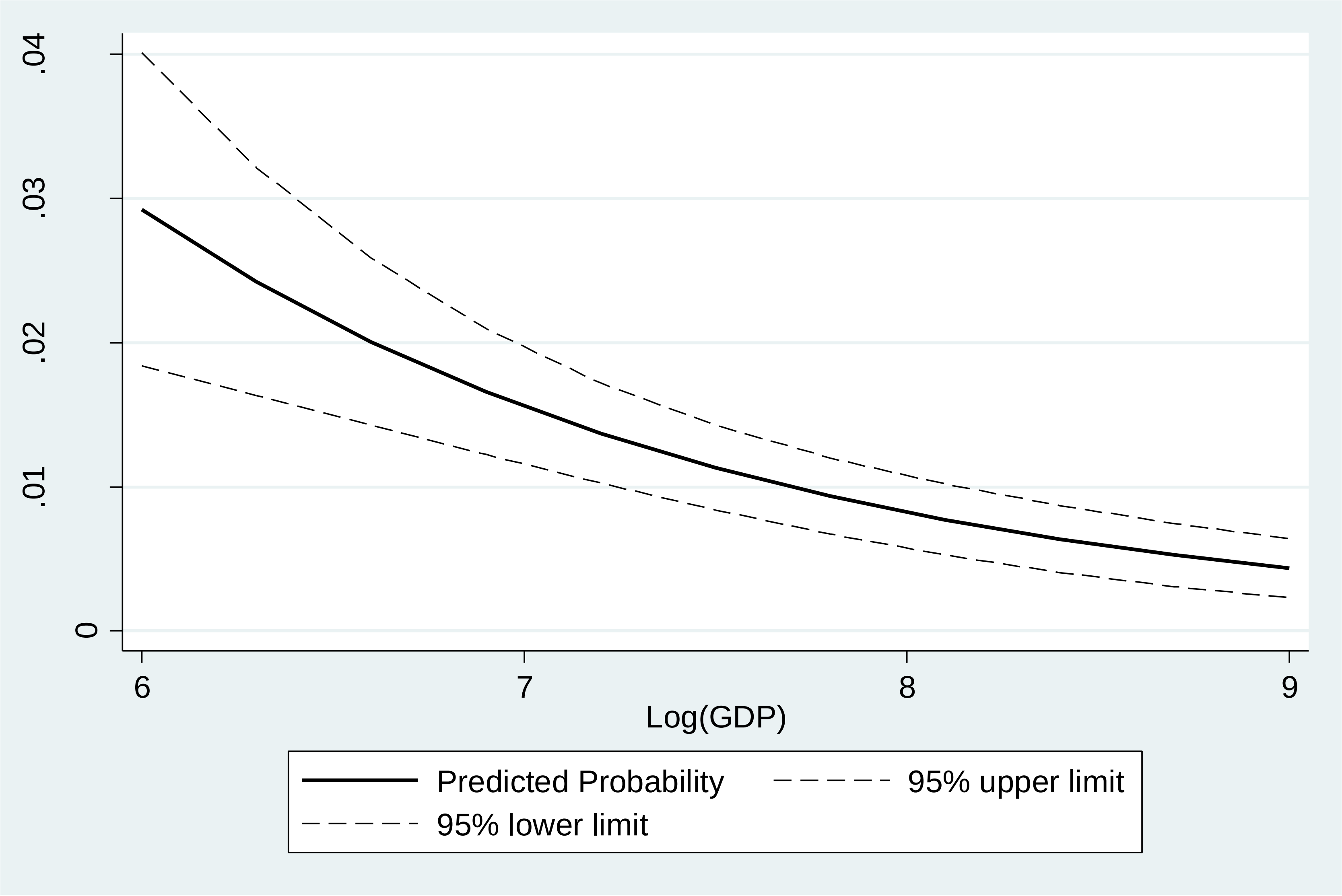
- To look at only GDP's effect on the probability of conflict I can set all other variables at their mean values (or modes for dichotomous variables) and calculate the predicted parameter $\hat{\pi}$.

Where:

$$\hat{\pi} = \frac{1}{1 + \exp[-\bar{X}\widehat{\beta^{-}} - GDP(\beta\beta_{GDP})}$$

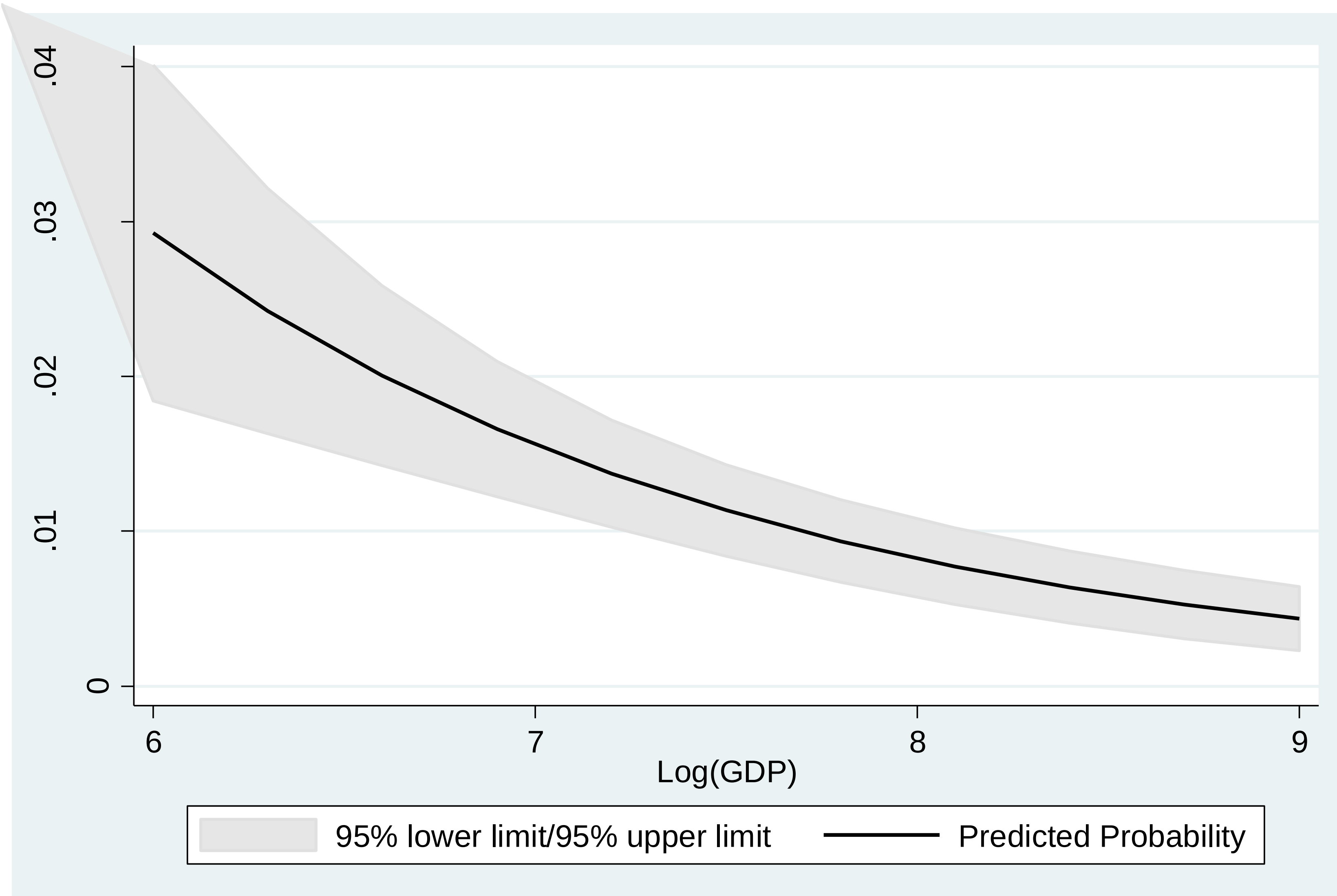
- Then we plug in different values for GDP and plot them.
- For more detail see King (1989: 105)

This is what we get…



```
logit onset warl lgdpenl1 lpopl1 lmtnest ncontig Oil nwstate instab anocl deml ///
         ethfrac relfrac, robust cluster(ccode) nolog

prgen lgdpenl1, from(6) to(9) gen(gdp) rest(mean) ci

label var gdpp1 "Predicted Probability"
label var gdpp1ub "95% upper limit"
label var gdpp1lb "95% lower limit"
label var gdpx "GDP"

* using lines to show confidence interval
twoway ///
    (connected gdpp1 gdpx, ///
        clcolor(black) clpat(solid) clwidth(medthick) ///
        msymbol(i) mcolor(none)) ///
    (connected gdpp1ub gdpx, ///
        msymbol(i) mcolor(none) ///
     clcolor(black) clpat(dash) clwidth(thin)) ///
    (connected gdpp1lb gdpx, ///
         msymbol(i) mcolor(none) ///
         clcolor(black) clpat(dash) clwidth(thin)), ///
    ytitle("Probability of Civil War Onset") yscale(range(0 .03)) ///
    ylabel(, grid glwidth(medium) glpattern(solid)) ///
    xscale(range(6 9)) xlabel(6(1)9)  xtitle("Log(GDP)") ///
    ysize(4) xsize(6)
```

Alternatively…



```
* using shading to show confidence interval
graph twoway ///
    (rarea gdpp1lb gdpp1ub gdpx, color(gs14)) ///
    (connected gdpp1 gdpx, ///
        clcolor(black) clpat(solid) clwidth(medthick) ///
        msymbol(i) mcolor(none)), ///
    ytitle("Probability of Civil War Onset") yscale(range(0 .03)) ///
    ylabel(, grid glwidth(medium) glpattern(solid)) ///
   xscale(range(6 9)) xlabel(6(1)9) xtitle("Log(GDP)") ///
    ysize(4) xsize(6) //
    legend(label (1 "95% confidence interval"))
```

Or we can look at point predictions using Clarify

```
. estsimp logit onset warl lgdpenl1 lpopl1 lmtnest ncontig Oil nwstate instab anocl deml ///
          ethfrac relfrac, robust cluster(ccode) nolog

setx lgdpenl1 p25 anocl 1

. simqi

      Quantity of Interest |      Mean       Std. Err.    [95% Conf. Interval]
---------------------------+----------------------------------------------------
               Pr(onset=0) |   .9982832      .0011713      .9953762    .9995933
               Pr(onset=1) |   .0017168      .0011713      .0004067    .0046238
```

	GDP 25th percentile	GDP 50th percentile	GDP 75th percentile
Anocratic	.0017	.0010	.00062
Democratic	.0010	.0006	.00030
Percent Δ Pr(onset)	-41%	-40%	-48%

```
** USING CLARIFY **

set seed 1234567890
sort ccode year
estsimp logit onset warl lgdpenl1 lpopl1 lmtnest ncontig Oil nwstate
instab anocl deml ///
        ethfrac relfrac, robust cluster(ccode) nolog

setx lgdpenl1 p25 anocl 1
simqi
setx lgdpenl1 p50 anocl 1
simqi
setx lgdpenl1 p75 anocl 1
simqi

setx lgdpenl1 p25 anocl 0 deml 1
simqi
setx lgdpenl1 p50 anocl 0 deml 1
simqi
setx lgdpenl1 p75 anocl 0 deml 1
simqi
```

- Lastly, it is possible to set values to those actually observed in your data (say Nicaragua in 1993 or Ethiopia in 2008).

- In summary, there are a number of different ways of trying to tease out the relationship between your dependent and independent variables.

- Moving from interpretation back to modeling…

Heteroskedastic Probit

- Heteroskedasticity is a greater problem in ML than in OLS, so it is important to account for non-constant variance to inference and prediction.

- To begin it is important to understand the difference between heteroskedasticity and heterogeneity.

- **Heterogeneity** typically refers to a difference between groups in their <u>mean</u> in Y.

- **Heteroskedasticity** refers to a difference in the <u>variances</u> between groups.

- Again remember the Bernoulli distribution.

$$Y = \begin{cases} 1\,, & \pi \\ 0, & 1 - \pi \end{cases}$$

Where:

$$\pi_i = f(X\beta)$$

and for probit = $\Phi(X\beta)$

Where $P(Y = 1) = \Phi(X\beta)$

and $P(Y = 0) = 1 - \Phi(X\beta)$

- The probit maximum likelihood function again:

$$\text{Ln } L = \sum_{i=1}^{n} y_i \ln\Phi(X\beta) + (1\text{-} y_i)\ln\Phi(X\beta)$$

- And again remember that for both the logit and probit the variance was assumed to be a constant—1 for probit and $\frac{\pi^2}{3}$ for logit.

- But what if the model is not homoskedastic? In OLS our estimates would not be efficient, but they would still be unbiased.

- This is not the case in ML.

Consider an example…

- Supposed we have data on baseball players at-bats and whether they get hits.
- Whether to swing or not is determined by some X variables.
- We have data on both Major League players and AAA minor league players.
- If we run probit models on both types of players we find that they have the same βs.
- The effect of X on Pr(swinging) is the same for both samples and models.

- However, think about these two groups.
- The Major Leaguers are the best of the best (think Big Poppie David Ortiz) and are selective in what pitches they go for.
- The AAA players are also quite good, but maybe they are not as selective so they swing at more pitches than they should.
- This means that the error term is smaller for the ML players than the AAA players.

- Now suppose we put data on both types of players into one model (Yes! More observations).

- If we include a dummy for whether the player is ML or AAA, this variable should be insignificant.

- But remember the kernel of the probit likelihood function is a ratio:

$$ln\Phi\left(\frac{X\beta}{\sigma^2}\right)$$

- We are assuming that the variance is the same for both groups and that it is equals to 1. If either assumption fails then our estimates of β will be inconsistent.

Again:

$$ln\Phi\left(\frac{X\beta}{\sigma^2}\right)$$

So we can write:

$$\frac{\left(\frac{\beta_{AAA}}{\sigma^2{}_{AAA}}\right)}{\left(\frac{\beta_{ML}}{\sigma^2{}_{ML}}\right)} = 1$$

- But what if $\sigma^2{}_{ML}$ equals 1 but $\sigma^2{}_{AAA}$ does not equal 1 but rather δ?

- Then,

$$\frac{\left(\frac{\beta_{AAA}}{\delta_{AAA}}\right)}{\left(\frac{\beta_{ML}}{1}\right)} = \left(\frac{\beta_{AAA}}{\delta_{AAA}\beta_{ML}}\right)$$

- Thus as the variance γ in AAA players increases our estimate of $\beta_{\ AAA}$ decreases.

- This means that our estimates in ML are both inconsistent and inefficient.

- This in some ways is similar to omitted variable bias.
- Some difference between groups in our data is not being controlled for.
- However correcting for heteroskedasticity is not as simple as correcting for heterogeneity.

- To correct for heterogeneity then, you need to control for the difference in variance.

- For the probit model let

$$\text{Var}[\varepsilon]=\sigma^2 = [e^{\gamma z'}]^2$$

So

$$\sigma = e^{\gamma z'}$$

This gives the probit likelihood function:

$$\text{Ln } L = \sum_{i=1}^{n} y_i \ln\Phi\left(\frac{X\beta}{e^{\gamma Z'}}\right) + (1 - y_i)\ln\Phi\left(\frac{X\beta}{e^{\gamma Z'}}\right)$$

- Therefore we now have two unknowns, β and γ, where β represents the effects of X on the mean probability of Y, and γ represents the effects of Z on the variance of Y.

- It is possible for X and Z to be the same (e.g. ethnic fractionalization in Blimes 2006).

- Why $e^{\gamma z'}$? Basically it has some properties that the variance must maintain.

  - First, it must be positive.

  - Second, if the effect of Z on the variance is 0, then σ^2 must revert to 1. Exponentiating γ does both of these things.

- Note that as the estimated heteroskedasticity γ increases the effects of the independent variables **X** decreases. (why is this?)

- Now, let's turn to the Blimes (2006) and the Alvarez and Brehm (1995) articles to look at how researchers have used (and theoretically justified using) heteroskedastic probit.

Interaction Terms

- A detailed exploration of the debate about a) the need for and b) the interpretation of interaction terms is outside the scope of this class.

- I did want to show you that it is out there.

- Let's take the time we have left to work through Berry et al.'s (2010) argument and evidence.

- See you next week!