1. do "C:\web\PS4dofile.txt"
2. insheet using "C:\web\PS4.txt"
   (19 vars, 1636 obs)
3. *2 -- Estimate Linear Probability Model of diabetes on bmi and income*
4. regress diabetes bmi income, robust
   Linear regression                                      Number of obs =    1367
   F(  2,  1364) =     15.19
   Prob > F      =             0.0000
   R-squared     =      0.0335
   Root MSE      =   0.23504

   diabetes    Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
   bmi         .0062544    .0015    4.17  0.000       .0033119    .0091969
   income      -.0124606    .0034855  -3.58  0.000     -.0192981   -.0056231
   _cons       -.0393018    .043229   -0.91  0.363     -.1241044    .0455008

5. *3 -- Do same thing using bmierr which is bmi + a mean 0 variance 100 random > error*
6. regress diabetes bmierr income, robust
   Linear regression                                      Number of obs =    1367
   F(  2,  1364) =      7.600
   Prob > F      =             0.0005
   R-squared     =      0.0141
   Root MSE      =   0.23738

   diabetes    Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
   bmierr      .0003839    .0005335   0.72   0.472     -.0006627    .0013404
   income      -.0135402    .0035193  -3.85  0.000     -.0204441   -.0066363
   _cons       .1227672    .0255838   4.80   0.000     .0725793    .1729551

7. *the coefficient here is much closer to zero/ smaller in magnitude*
8. *measurement error in x variables leads to attenuation bias i.e., bias toward > zero*
9. *while measurement error in y simply adds to the noise of the model*
10. *measurement error in x leads to a bias toward zero*
11. *see Wooldridge pp. 318-320 for a formal presentation*
12. *but one intuitive way to think about it is that*
13. *as the noise in the x variable gets big relative to the signal*
**it's like regressing y on an error term**

which is going to lead to an estimated effect that approaches zero

what would happen if you used bmi measured with error but where the error is lower variance

well then the signal gets stronger relative to the noise

so the bias toward zero will be smaller

4 -- redo #2 with logit

first #2 again

regress diabetes bmi income, robust

| diabetes | Robust Coef. | Std. Err. | t     | P>|t|     | [95% Conf. Interval] |
|----------|--------------|-----------|-------|---------|---------------------|
| bmi      | 0.066254     | 0.0015    | 4.17  | 0.000   | 0.003319 - 0.0091969|
| income   | -0.014606    | 0.0034855 | -3.58 | 0.000   | -0.0192981 - 0.0056231|
| _cons    | -0.0393018   | 0.043229  | -0.91 | 0.363   | -0.1241044 - 0.0455008|

Logit diabetes bmi income, robust

Iteration 0:  log pseudolikelihood = -312.9549
Iteration 1:  log pseudolikelihood = -296.66849
Iteration 2:  log pseudolikelihood = -292.74888
Iteration 3:  log pseudolikelihood = -292.7108
Iteration 4:  log pseudolikelihood = -292.7108

Logistic regression

| diabetes | Robust Coef. | Std. Err. | z     | P>|z|     | [95% Conf. Interval] |
|----------|--------------|-----------|-------|---------|---------------------|
| bmi      | 0.0868625    | 0.0168344 | 5.16  | 0.000   | 0.0538678 - 0.1198572|
| income   | -0.2083366   | 0.054089  | -3.85 | 0.000   | -0.3143492 - 0.102324|
| _cons    | -4.142655    | 0.5736033 | -7.22 | 0.000   | -5.266897 - 3.018413|

*remember that we can only compare the sign and significance across*

*the models; if we want to compare magnitude size we need to estimate the log it*

at a certain x vector, namely the mean

to do that for the logit command, we type mfx*
Marginal effects after logit
   y = Pr(diabetes) (predict) = .0501379

| variable | dy/dx   | Std. Err. | z     | P>|z| | [ 95% C.I. ] | X     |
|----------|---------|-----------|-------|------|-------------|-------|
| bmi      | .0041367 | .00081    | 5.08  | 0.000 | .00254      | .005734 | 26.6061 |
| income   | -.0099218 | .00247    | -4.01 | 0.000 | -.014766    | -.005078 | 5.32772 |

31. *we could have calculated the effects at any values of the X's that we wanted
32. *now the probit*
33. probit diabetes bmi income, robust

Iteration 0:  log pseudolikelihood = -312.9549
Iteration 1:  log pseudolikelihood = -292.58751
Iteration 2:  log pseudolikelihood = -292.10174
Iteration 3:  log pseudolikelihood = -292.10142

Probit regression  Number of obs = 1367
                        Wald chi2( 2) = 43.16
                        Prob > chi2 = 0.0000

Log pseudolikelihood = -292.10142 Pseudo R2 = 0.0666

| diabetes | Robust Coef. | Std. Err. | z     | P>|z| | [95% Conf. Interval] |
|----------|--------------|-----------|-------|------|----------------------|
| bmi      | .0452375     | .0089391  | 5.06  | 0.000 | .0277172             | .0627578 |
| income   | -.1023977    | .0264308  | -3.87 | 0.000 | -.1542011            | -.0505944 |
| _cons    | -2.296131    | .2935368  | -7.82 | 0.000 | -2.871453            | -1.72081 |

34. *marginal effects from the probit can be calculated again using the mfx comm
35. *or, dprobit does it directly (if we didn't really care about the underlying*
36. *model parameters, which is usually the case*
37. dprobit diabetes bmi income, robust

Iteration 0:  log pseudolikelihood = -312.9549
Iteration 1:  log pseudolikelihood = -292.58751
Iteration 2:  log pseudolikelihood = -292.10174
Iteration 3:  log pseudolikelihood = -292.10142

Probit regression, reporting marginal effects  Number of obs = 1367
                        Wald chi2( 2) = 43.16
                        Prob > chi2 = 0.0000

Log pseudolikelihood = -292.10142 Pseudo R2 = 0.0666

| diabetes | Robust dF/dx | Std. Err. | z     | P>|z| | x-bar [ 95% C.I. ] |
|----------|--------------|-----------|-------|------|--------------------|
| bmi      | .0047177     | .0009442  | 5.06  | 0.000 | 26.6061            | .002867 | .006568 |
| income   | -.0106789    | .0027017  | -3.87 | 0.000 | 5.32772            | -.015974 | -.005384 |

| obs. P  | .0607169     | (at x-bar) |
| pred. P | .0507021     |            |

z and P>|z| correspond to the test of the underlying coefficient being 0
38. "as you can see, all three models are producing comparable effect sizes"
39. "graph predictions from linear probability model, logit, probit"
40. `regress diabetes bmi income, robust`
    `logit diabetes bmi income, robust`
    `predict LPM`
2
41. `graph predictions from linear probability model, logit, probit`
    `probit diabetes bmi income, robust`

### Linear Regression

| diabetes | Coef.   | Std. Err. | t     | P>|t|    | [95% Conf. Interval] |
|----------|---------|-----------|-------|--------|---------------------|
| bmi      | .0062544 | .0015     | 4.17  | .000   | .0033119 to .0091969 |
| income   | -.0124606 | .0034855  | -3.58 | .000   | -.0192981 to -.0056231 |
| _cons    | -.0393018 | .043229   | -0.91 | .363   | -.1241044 to .0455008 |

### Logistic Regression

| diabetes | Coef.   | Std. Err. | z     | P>|z|    | [95% Conf. Interval] |
|----------|---------|-----------|-------|--------|---------------------|
| bmi      | .0868625 | .0168344  | 5.16  | .000   | .0538678 to .1198572 |
| income   | -.2083366 | .054089   | -3.85 | .000   | -.3143492 to -.102324 |
| _cons    | -4.142655 | .5736033  | -7.22 | .000   | -5.266897 to -3.018413 |

### Probit Regression

| diabetes | Coef.   | Std. Err. | z     | P>|z|    | [95% Conf. Interval] |
|----------|---------|-----------|-------|--------|---------------------|
| bmi      | .0868625 | .0168344  | 5.16  | .000   | .0538678 to .1198572 |
| income   | -.2083366 | .054089   | -3.85 | .000   | -.3143492 to -.102324 |
| _cons    | -4.142655 | .5736033  | -7.22 | .000   | -5.266897 to -3.018413 |

41. `predict LPM`  
    (option xb assumed; fitted values)  
    (268 missing values generated)

42. `logit diabetes bmi income, robust`
    
    *Iteration 0: log pseudolikelihood = -312.9549*
    *Iteration 1: log pseudolikelihood = -296.66849*
    *Iteration 2: log pseudolikelihood = -292.7488*
    *Iteration 3: log pseudolikelihood = -292.7108*
    *Iteration 4: log pseudolikelihood = -292.7108*

### Logistic Regression

| diabetes | Coef.   | Std. Err. | z     | P>|z|    | [95% Conf. Interval] |
|----------|---------|-----------|-------|--------|---------------------|
| bmi      | .0868625 | .0168344  | 5.16  | .000   | .0538678 to .1198572 |
| income   | -.2083366 | .054089   | -3.85 | .000   | -.3143492 to -.102324 |
| _cons    | -4.142655 | .5736033  | -7.22 | .000   | -5.266897 to -3.018413 |

43. `predict logit`  
    (option pr assumed; Pr(diabetes))  
    (268 missing values generated)

44. `probit diabetes bmi income, robust`
    
    *Iteration 0: log pseudolikelihood = -312.9549*
    *Iteration 1: log pseudolikelihood = -292.58751*
    *Iteration 2: log pseudolikelihood = -292.10174*
    *Iteration 3: log pseudolikelihood = -292.10142*

### Probit Regression

| diabetes | Coef.   | Std. Err. | z     | P>|z|    | [95% Conf. Interval] |
|----------|---------|-----------|-------|--------|---------------------|
| bmi      | .0868625 | .0168344  | 5.16  | .000   | .0538678 to .1198572 |
| income   | -.2083366 | .054089   | -3.85 | .000   | -.3143492 to -.102324 |
| _cons    | -4.142655 | .5736033  | -7.22 | .000   | -5.266897 to -3.018413 |
Robust diabetes Coef. Std. Err. z P>|z| [95% Conf. Interval]
bmi .0452375 .0089391 5.06 0.000 .0277172 .0627578
income -.1023977 .0264308 -3.87 0.000 -.1542011 -.0505944
_cons -2.296131 .2935368 -7.82 0.000 -2.871453 -1.72081

45. predict probit
   (option pr assumed; Pr(diabetes))
   (268 missing values generated)

46. twoway (scatter LPM bmi, sort)
47. twoway (scatter logit bmi, sort)
48. twoway (scatter probit bmi, sort)

49. *regress bmi private which top codes bmi above 35 as 35 on income and educati
   > on*
50. *we need a censored regression model for a y variable like this*
51. *cnreg will work; check the help to see how it works*
52. *we need to create a variable to tell the computer whether an observation is
   > censored*
53. *and we need to tell it whether it's censored above or below*
54. *Stata uses 0 for uncensored, 1 for censored above, and -1 for censored below
   > *
55. generate cens =0
56. replace cens =1 if bmi > 35
   (177 real changes made)
57. cnreg bmipriv income educa, robust cens(cens)

Censored-normal regression
   Number of obs = 1414
   F(  2, 1412) = 5.86
   Prob > F = 0.0029
Log pseudolikelihood = -4074.435
   Pseudo R2 = 0.0015

Robust bmipriv Coef. Std. Err. t P>|t| [95% Conf. Interval]
income -.0684913 .0791855 -0.86 0.387 -.2238252 .0868427
educa -.3909632 .1502624 -2.60 0.009 -.6857247 -.0962017
_cons 29.05579 .6871785 42.28 0.000 27.70779 30.40379
/sigma 5.180393 .1066678 4.9711149 5.389638

Observation summary: 0 left-censored observations
1270 uncensored observations
144 right-censored observations

58. regress bmi income educa, robust

Linear regression
   Number of obs = 1367
   F(  2, 1364) = 5.49
   Prob > F = 0.0042
   R-squared = 0.0078
   Root MSE = 5.3431
59. *intuition -- well, since you're not using all of the information in a subset
> of the data*
60. *that might have an effect*
61. *the censored regression is estimating a smaller in magnitude effect of incom
> e*
62. *and a bigger in magnitude effect of educa, but the differential is*
63. *proportionately smaller for education*
64. *so one guess would be that while the > 35 bmi people are systematically lowe
> r*
65. *in income than the rest*
66. *they are not too much different than everyone else in education terms*
67. *at least conditional on income*
68. *so let's check that intuition*
69. regress cens income educa

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 1414</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>.830980721</td>
<td>2</td>
<td>.41549036</td>
<td>F( 2, 1411) = 4.56</td>
</tr>
<tr>
<td>Residual</td>
<td>128.504239</td>
<td>1411</td>
<td>.091073167</td>
<td>Prob &gt; F = 0.0106</td>
</tr>
<tr>
<td>Total</td>
<td>129.335219</td>
<td>1413</td>
<td>.091532356</td>
<td>Adj R-squared = 0.0050</td>
</tr>
</tbody>
</table>

| cens         | Coef.     | Std. Err. | t     | P>|t|     | [95% Conf. Interval] |
|--------------|-----------|-----------|-------|---------|---------------------|
| income       | -.0110143 | .0043164  | -2.55 | 0.011   | -.0194815 -.0025472 |
| educa        | -.0028227 | .008598   | -0.33 | 0.743   | -.019689 .0140436  |
| _cons        | .1741926  | .0384623  | 4.53  | 0.000   | .0987431 .249642   |

70. *our intuition is validated*
71. *7 -- run truncated regression model*
72. truncreg bminormal income educa, ll(18.5) ul(25) robust
(note: 6 obs. truncated)

Fitting full model:

Iteration 0:  log pseudolikelihood = -1013.2253
Iteration 1:  log pseudolikelihood = -986.92835
Iteration 2:  log pseudolikelihood = -986.74233
Iteration 3:  log pseudolikelihood = -986.54979
Iteration 4:  log pseudolikelihood = -986.54833
Iteration 5:  log pseudolikelihood = -986.54833

Truncated regression
Limit:  lower = 18.5  Number of obs = 541
       upper = 25  Wald chi2(2) = 1.82
Log pseudolikelihood = -986.54833  Prob > chi2 = 0.4027
When we chop off the tails, it's close to zero in the lower end, so the differences cancel out. But because we have less variation in X, intuitively, we're getting comparable results for income, which makes sense.

```
regress bmi income educa, robust
```

| Robust | Coef.   | Std. Err. | z     | P>|z|  | [95% Conf. Interval] |
|--------|---------|-----------|-------|------|----------------------|
| income | -.092955| .1153835  | -0.81| 0.420| -.3191025 .1331924  |
| educa  | .3115204| .2322346  | 1.34 | 0.180| -.143651 .7666919   |
| _cons  | 21.92178| .9864519  | 22.22| 0.000| 19.98837 23.85519   |
| /sigma | 2.789208| .3340342  | 8.35 | 0.000| 2.134513 3.443903   |

73. \texttt{. regress bmi income educa}

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs =</th>
<th>1367</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>304.479098</td>
<td>2</td>
<td>152.239549</td>
<td>F( 2, 1364) = 5.33</td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>38939.7992</td>
<td>1364</td>
<td>28.5482398</td>
<td>Prob &gt; F = 0.0049</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>39244.2783</td>
<td>1366</td>
<td>28.7293399</td>
<td>R-squared = 0.0063</td>
<td></td>
</tr>
</tbody>
</table>

| Coef.   | Std. Err. | t     | P>|t|  | [95% Conf. Interval] |
|---------|-----------|-------|------|----------------------|
| income  | -.1000753 | .0779809 | -1.28 | 0.200 | -.2530508 .0529001 |
| educa   | -.3308629 | .1551339 | -2.13 | 0.033 | -.6351898 -.026536 |
| _cons   | 28.73271  | .6945347 | 41.37 | 0.000 | 27.37024 30.09518 |

74. \texttt{. regress bmi income educa, robust}

Linear regression | Number of obs = | 1367 |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>F( 2, 1364) = 5.49</td>
<td>Prob &gt; F = 0.0042</td>
<td>R-squared = 0.0078</td>
</tr>
<tr>
<td>Root MSE = 5.3431</td>
<td>Root MSE = 5.3431</td>
<td></td>
</tr>
</tbody>
</table>

| Coef.   | Std. Err. | t     | P>|t|  | [95% Conf. Interval] |
|---------|-----------|-------|------|----------------------|
| income  | -.1000753 | .0831068 | -1.20 | 0.229 | -.2631063 .0629556 |
| educa   | -.3308629 | .1473205 | -2.25 | 0.025 | -.6198622 -.0418639 |
| _cons   | 28.73271  | .6766832 | 42.46 | 0.000 | 27.40526 30.06016 |

75. "Intuitively, we're getting comparable results for income, which makes sense since.*
76. "While we're cutting off the tails, bmi is basically symmetric with respect to income.*
77. "While the income effect is twice as big in the upper end of bmi.*
78. "It's close to zero in the lower end, so the differences cancel out.*
79. "When we chop off the tails.*
80. "But because we have less variation in X.*
81. "It makes sense that our SEs blow up.*
82. "Education, however, is a different story."
*while education is negatively related to bmi, conditional on income*
*on average throughout the sample*
in the truncated regression, it comes in with a positive sign*
*if you look at the relationship between education and bmi, conditional*
*on income in the "normal" range, it is actually positive*
*so the negative effect is driven by a larger in magnitude*
*negative effect in the tails*
*the truncated regression doesn't use the info from the tails*

end of do-file