Binary dependent variable models

EC 339

Marcio Santetti

Fall 2022

Motivation

The road so far

So far, we have studied models with **binary** variables on the regression's right-hand-side, as an *explanatory* factor.

But what if we want to have a **qualitative** indicator as the model's *dependent variable*?

Several decisions made by individuals and firms are *either-or* in nature.

For instance, what are the factors that determine an individual's decision to **join the labor force**, **enroll in a course**, or **drink Coke over Pepsi**?

To do that, we turn to **binary dependent variable models**.

The road so far

The problem now becomes setting up a statistical model of **binary** choices.

We represent these choices by an **indicator** variable that equals **1** if the outcome is chosen, and **0** otherwise.

Unlike flipping a *coin* or rolling a *die*, the probability of an individual choosing an outcome depends on **many factors**.

- Let these factors be denoted by $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{ik})$.

The road so far

Then, the **conditional probability** that the i^{th} individual **chooses** a given outcome is given by

$$P(y_i = 1 \mid \mathbf{x}_i) = p(\mathbf{x}_i)$$

And the **conditional probability** that the i^{th} individual **does not** choose a given outcome is given by

$$P(y_i = 0 \mid \mathbf{x}_i) = 1 - p(\mathbf{x}_i)$$

where $0 \leq p(\mathbf{x}_i) \leq 1$.

In *general*, we can write a **conditional probability function**:

$$f(y_i \mid \mathbf{x}_i) = p(\mathbf{x}_i)^{y_i} \left[1 - p(\mathbf{x}_i)\right]^{1-y_i} \qquad y_i = 0, 1$$

The Linear Probability Model

The Linear Probability Model

The **Linear Probability Model** (LPM) is the first alternative to estimate binary choice models.

It simply consists in estimating a model with $p(y_i \mid \mathbf{x}_i)$ as the dependent variable via **OLS**.

And since the left-hand side of the regression now has a **probability function**, we have

$$\mathbb{E}(y_i \mid \mathbf{x}_i) = \sum_{y_i=0}^{1} y_i f(y_i|\mathbf{x}_i) = 0 \times f(0|\mathbf{x}_i) + 1 \times f(1|\mathbf{x}_i) = p(\mathbf{x}_i)$$

$$p(\mathbf{x}_i) = \mathbb{E}(y_i \mid \mathbf{x}_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}$$

and $u_i = y_i - \mathbb{E}(y_i \mid \mathbf{x}_i)$.

The Linear Probability Model

Therefore, the **full** Linear Probability Model is:

$$y_i = \mathbb{E}(y_i \mid \mathbf{x}_i) + u_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i$$

And the **marginal effect** of a one-unit change in a variable *j* changes the *probability of success*, $p(y_i = 1 \mid x_j)$, by

$$\frac{\partial\, \mathbb{E}(y_i \mid \mathbf{x}_i)}{\partial\, x_j} = \beta_j$$

Problem!: Suppose $\beta_j > 0$. Its interpretation implies that increasing x_{ji} by one unit will increase the probability of y_i being equal to 1 by a constant amount β_j.

- What is **wrong** with this?

The Linear Probability Model

Moreover, the residuals from an **LPM** model will likely be **heteroskedastic**:

$$Var(u_i \mid \mathbf{x}_i) \neq \sigma^2$$

Therefore, LPM models should always be estimated with *robust standard errors*.

The Linear Probability Model

An example:

```
lpm_model <- lm(inlf ~ nwifeinc + educ  + exper +
                  I(exper^2) + age + kidslt6 + kidsge6, data = mroz)
lpm_model %>% tidy()
```

```
#> # A tibble: 8 × 5
#>   term        estimate std.error statistic  p.value
#>   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)  0.586     0.154         3.80  1.58e- 4
#> 2 nwifeinc    -0.00341   0.00145     -2.35  1.90e- 2
#> 3 educ         0.0380    0.00738      5.15  3.32e- 7
#> 4 exper        0.0395    0.00567      6.96  7.38e-12
#> 5 I(exper^2)  -0.000596  0.000185    -3.23  1.31e- 3
#> 6 age         -0.0161    0.00248     -6.48  1.71e-10
#> 7 kidslt6     -0.262     0.0335      -7.81  1.89e-14
#> 8 kidsge6      0.0130    0.0132       0.986 3.24e- 1
```

When interpreting this model's *estimates*, recall that a change in the independent variable changes the probability that `inlf == 1`.

The Linear Probability Model

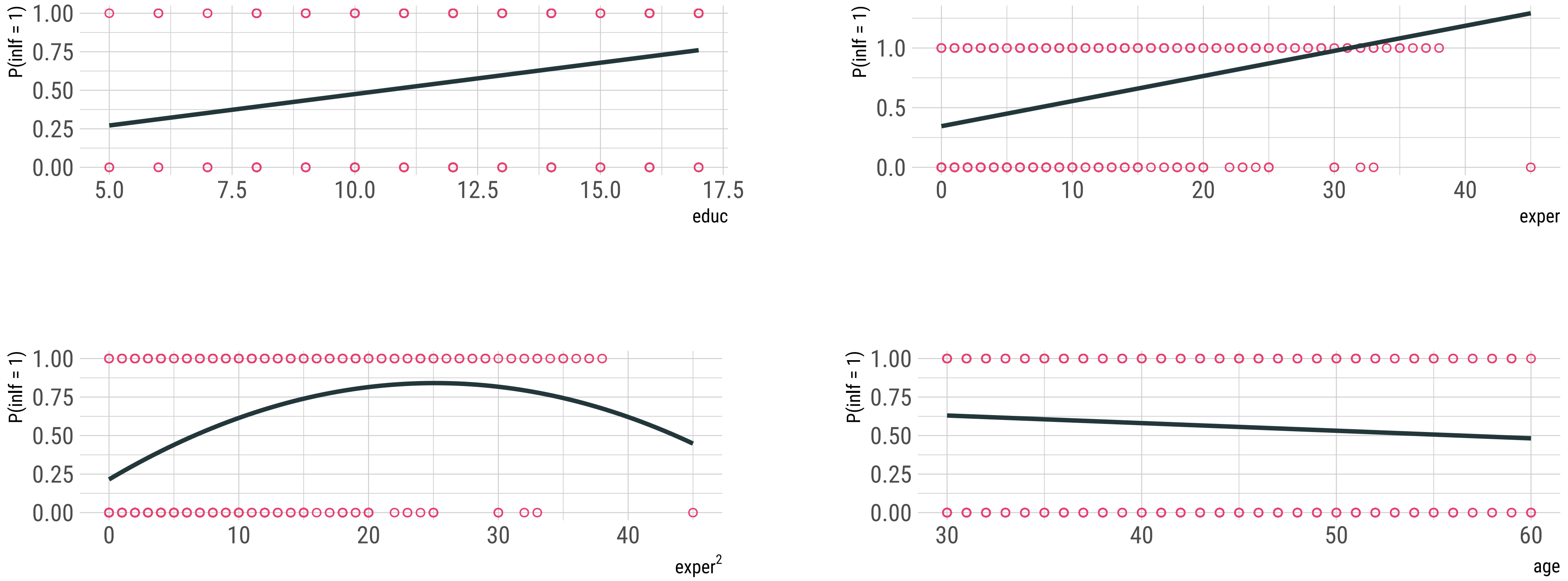
An example:

```
. reg inlf nwifeinc educ exper expersq age kidslt6 kidsge6

      Source |       SS           df       MS      Number of obs   =       753
-------------+----------------------------------   F(7, 745)       =     38.22
       Model |  48.8080578         7  6.97257969   Prob > F        =    0.0000
    Residual |  135.919698       745  .182442547   R-squared       =    0.2642
-------------+----------------------------------   Adj R-squared   =    0.2573
       Total |  184.727756       752  .245648611   Root MSE        =    .42713

------------------------------------------------------------------------------
        inlf | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    nwifeinc |  -.0034052   .0014485    -2.35   0.019    -.0062488   -.0005616
        educ |   .0379953    .007376     5.15   0.000      .023515    .0524756
       exper |   .0394924   .0056727     6.96   0.000     .0283561    .0506287
     expersq |  -.0005963   .0001848    -3.23   0.001    -.0009591   -.0002335
         age |  -.0160908   .0024847    -6.48   0.000    -.0209686    -.011213
     kidslt6 |  -.2618105   .0335058    -7.81   0.000    -.3275875   -.1960335
     kidsge6 |   .0130122    .013196     0.99   0.324    -.0128935    .0389179
       _cons |   .5855192    .154178     3.80   0.000     .2828442    .8881943
------------------------------------------------------------------------------
```

When interpreting this model's *estimates*, recall that a change in the independent variable changes the probability that `inlf == 1`.

The Linear Probability Model

Visually (assuming simple regression models):



`Logit` Models

`Logit` Models

The **main** issue with the Linear Probability Model is its incapacity to **constrain** the predicted probability between **0** and **1**.

The **`Logit`** and **`Probit`** models are examples of **nonlinear** models that address the above issue.

These models **ensure** that $p(y_i \mid \mathbf{x}_i)$ remains between 0 and 1.

This is made possible due to these models' ability to generate **S-shaped** (*sigmoid*) curves, which **do not** go beyond the [0,1] interval.

Think of a single-variable model with y as a binary outcome variable. If $\hat{\beta}_1 > 0$, as x increases, the probability of success increases rapidly at first, then begins to increase at a decreasing rate, keeping this probability below 1 no matter how large x becomes.

Moreover, **slope** coefficients are not *constant* anymore.

`Logit` Models

`Logit` models are based on a **logistic** random variable's *Cumulative Distribution Function* (CDF).

Consider a random variable L that follows a logistic distribution.

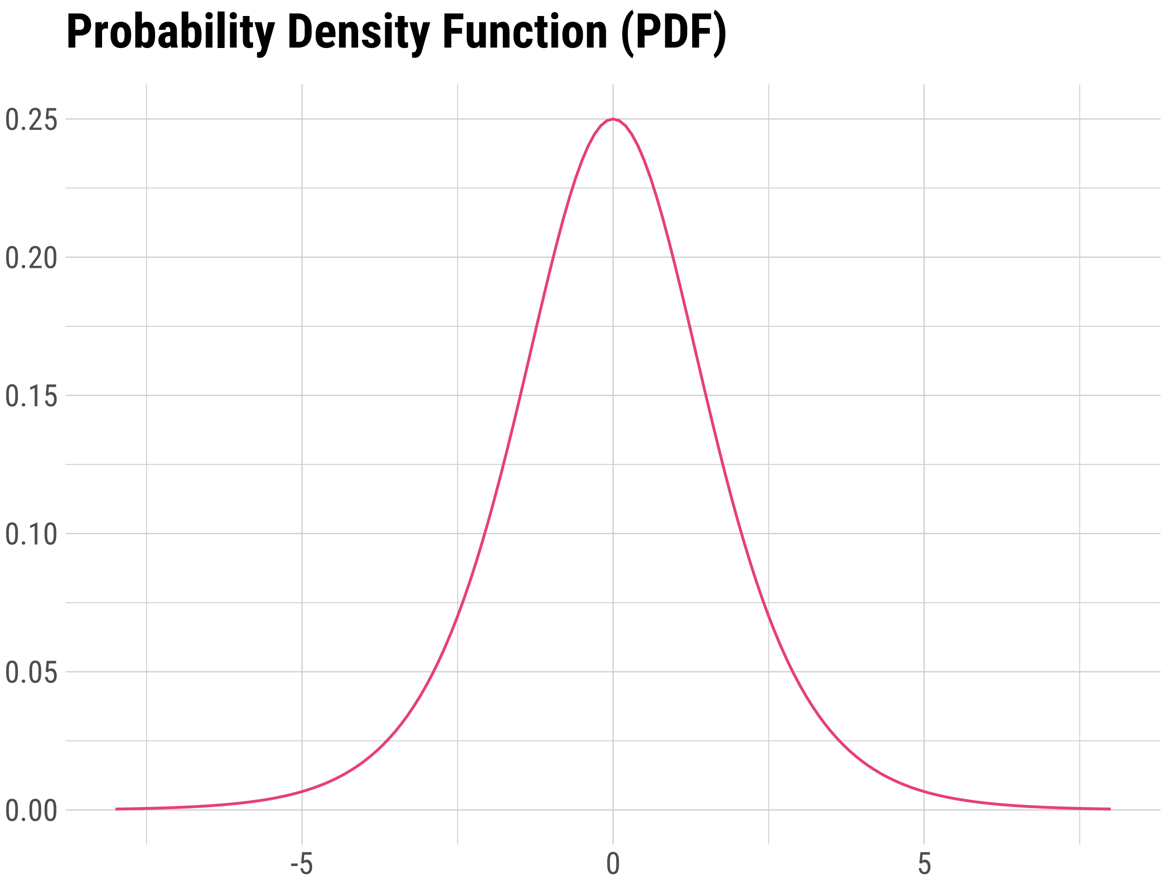
Then, its **Probability Density Function** (PDF) is given by

$$\lambda(l) = \frac{e^{-l}}{(1+e^{-l})^2} \qquad -\infty < l < \infty$$

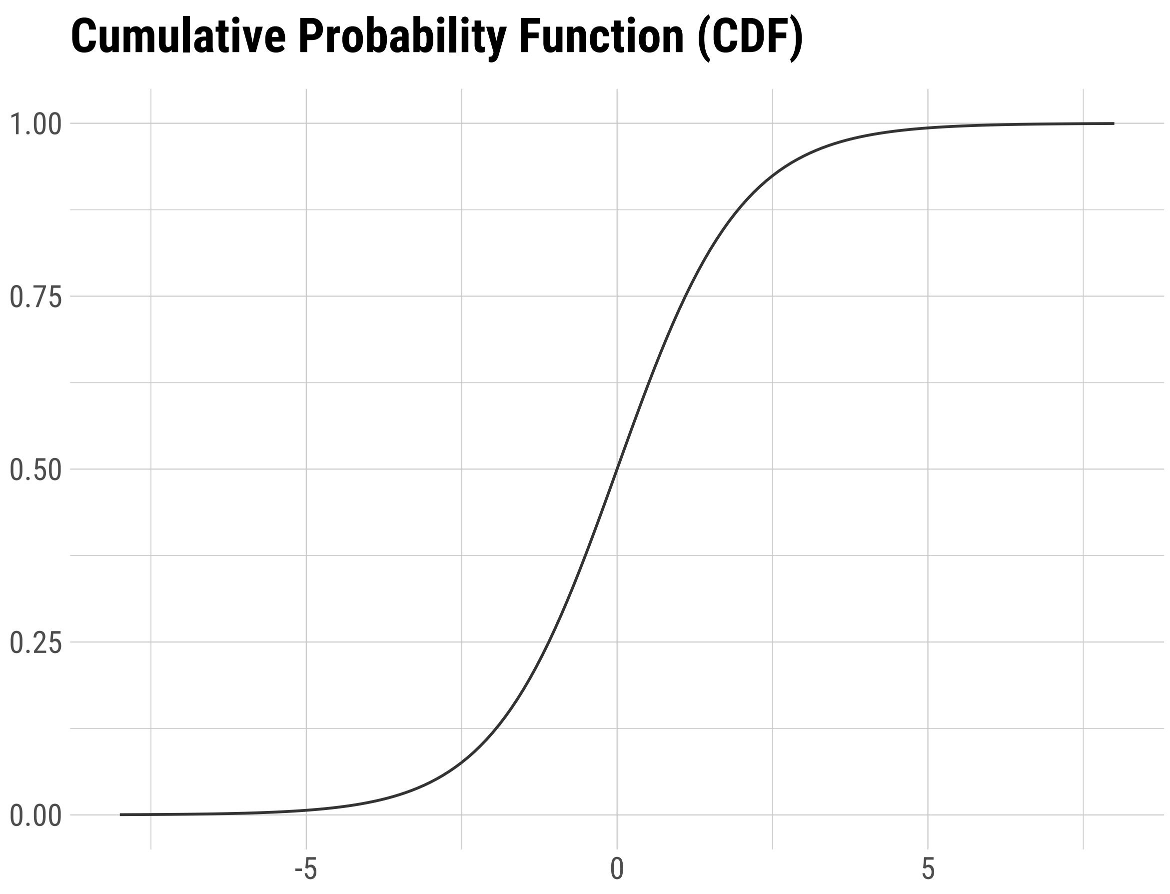
And its **Cumulative Density Function** (CDF) is given by

$$\Lambda(l) = p\left[L \leq l\right] = \frac{1}{1+e^{-l}}$$

Logit Models



Logit Models



Interpreting `Logit` Models

Interpreting `Logit` Models

`Logit` and `Probit` models use **maximum likelihood** to estimate model coefficients.

This implies a **completely different** coefficient interpretation from these models.

In case x_k is a **continuous** explanatory variable, its marginal effect on $p(y_i = 1 \mid \mathbf{x}_i)$ is given by

$$\frac{\partial\, p(\mathbf{x}_i)}{\partial\, x_{ik}} = \frac{\partial\, \Lambda(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki})}{\partial\, \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}} \cdot \frac{\partial\, \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}}{\partial\, x_{ik}} =$$

$$\frac{\partial\, p(\mathbf{x}_i)}{\partial\, x_{ik}} = \lambda(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki})\beta_k$$

Interpreting `Logit` Models

In case x_k is a **discrete explanatory variable** (such as a *dummy* variable), its interpretation is a bit different:

$$\Delta p(\mathbf{x}_i) = p(\mathbf{x}_i \mid x_k = 1) - p(\mathbf{x}_i \mid x_k = 0) =$$

$$\Delta p(\mathbf{x}_i) = \Lambda(\beta_0 + \beta_1 x_{1i} + \beta_k) - \Lambda(\beta_0 + \beta_1 x_{1i})$$

Interpreting `Logit` Models

So far, we have talked about model estimation.

But what about **coefficient interpretation**?

`Logit` coefficients are **not** directly interpretable.

Therefore, in order to do that, we have a few strategies.

The one we will focus on here is the **Average Marginal Effect** (AME).

$$\frac{\partial P(y_i = 1 \mid \mathbf{x}_i)}{\partial x_{ij}} = \frac{\partial \Lambda(\cdot)}{\partial x_{ij}} = \frac{\sum_{i=1}^{n} \lambda(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k)}{n} \cdot \hat{\beta}_j$$

The **AME** is the sample average of the ML estimation evaluated at each sample observation.

Interpreting `Logit` Models

For **discrete** explanatory variables, the **AME** is given by

$$\frac{\partial P(y_i = 1 \mid \mathbf{x}_i)}{\partial x_{ij}} = \frac{\sum_{i=1}^{n} \Lambda(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_j)}{n} - \frac{\sum_{i=1}^{n} \Lambda(\hat{\beta}_0 + \hat{\beta}_1 x_1)}{n}$$

A `Logit` example

```
logit_model <- glm(inlf ~ nwifeinc + educ  + exper +
                   I(exper^2) + age + kidslt6 + kidsge6, data = mroz,
              family = binomial(link='logit'))
logit_model %>% tidy()
```

```
#> # A tibble: 8 × 5
#>   term        estimate std.error statistic  p.value
#>   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)  0.425     0.860       0.495 6.21e- 1
#> 2 nwifeinc    -0.0213    0.00842    -2.53  1.13e- 2
#> 3 educ         0.221     0.0434      5.09  3.55e- 7
#> 4 exper        0.206     0.0321      6.42  1.34e-10
#> 5 I(exper^2)  -0.00315   0.00102    -3.10  1.91e- 3
#> 6 age         -0.0880    0.0146     -6.04  1.54e- 9
#> 7 kidslt6     -1.44      0.204      -7.09  1.34e-12
#> 8 kidsge6      0.0601    0.0748      0.804 4.22e- 1
```

From this output, we cannot directly interpret the model's **coefficients**.

However, we can interpret the coefficient's **signs**.

A `Logit` example

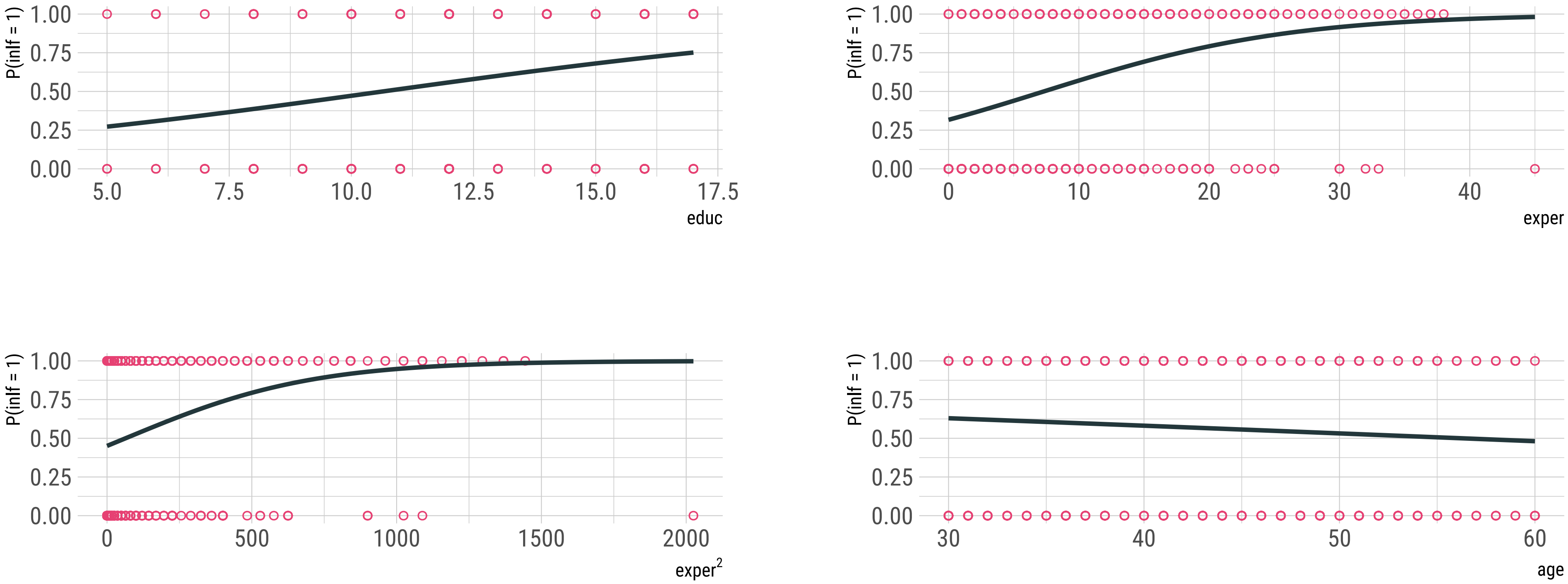
```
. logit inlf nwifeinc educ exper expersq age kidslt6 kidsge6

Iteration 0:   log likelihood =  -514.8732
Iteration 1:   log likelihood = -402.38502
Iteration 2:   log likelihood = -401.76569
Iteration 3:   log likelihood = -401.76515
Iteration 4:   log likelihood = -401.76515

Logistic regression                                     Number of obs =    753
                                                        LR chi2(7)    = 226.22
                                                        Prob > chi2   = 0.0000
Log likelihood = -401.76515                             Pseudo R2     = 0.2197

------------------------------------------------------------------------------
        inlf | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
    nwifeinc |  -.0213452   .0084214    -2.53   0.011    -.0378509   -.0048394
        educ |   .2211704   .0434396     5.09   0.000     .1360303    .3063105
       exper |   .2058695   .0320569     6.42   0.000     .1430391    .2686999
     expersq |  -.0031541   .0010161    -3.10   0.002    -.0051456   -.0011626
         age |  -.0880244    .014573    -6.04   0.000     -.116587   -.0594618
     kidslt6 |  -1.443354   .2035849    -7.09   0.000    -1.842373   -1.044335
     kidsge6 |   .0601122   .0747897     0.80   0.422     -.086473    .2066974
       _cons |   .4254524   .8603697     0.49   0.621    -1.260841    2.111746
------------------------------------------------------------------------------
```

From this output, we cannot directly interpret the model's **coefficients**.

A `Logit` example

The **PDF** for this estimated model looks like this:

A `Logit` example



A `Logit` example

Average Marginal Effects:

```
#>     Variable           AME
#> 1 intercept  0.0759771297
#> 2  nwifeinc -0.0038118135
#> 3      educ  0.0394965238
#> 4     exper  0.0367641056
#> 5   exper^2 -0.0005632587
#> 6       age -0.0157193606
#> 7   kidslt6 -0.2577536551
#> 8   kidsge6  0.0107348186
```

How to **interpret** these coefficients?

`Probit` Models

`Probit` Models

`Probit` models are based on the **standard normal** distribution's **Cumulative Distribution Function** (CDF).

Consider a random variable Z that follows a standard normal distribution.

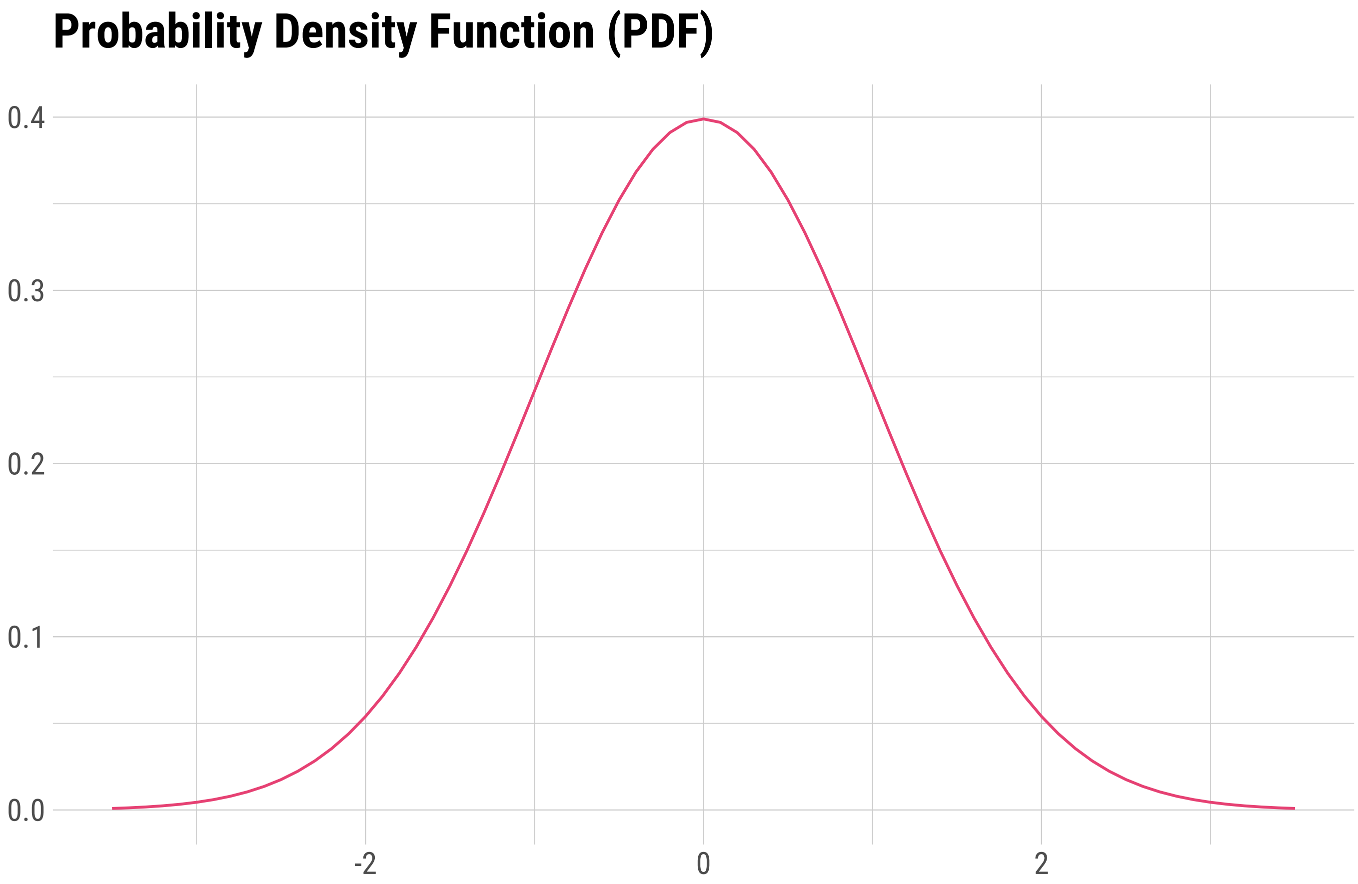
Then, its **Probability Density Function** (PDF) is given by

$$\phi(z) = \frac{1}{\sqrt{2\pi}}\, e^{-s^2/2\ z^2} \qquad -\infty < z < \infty$$

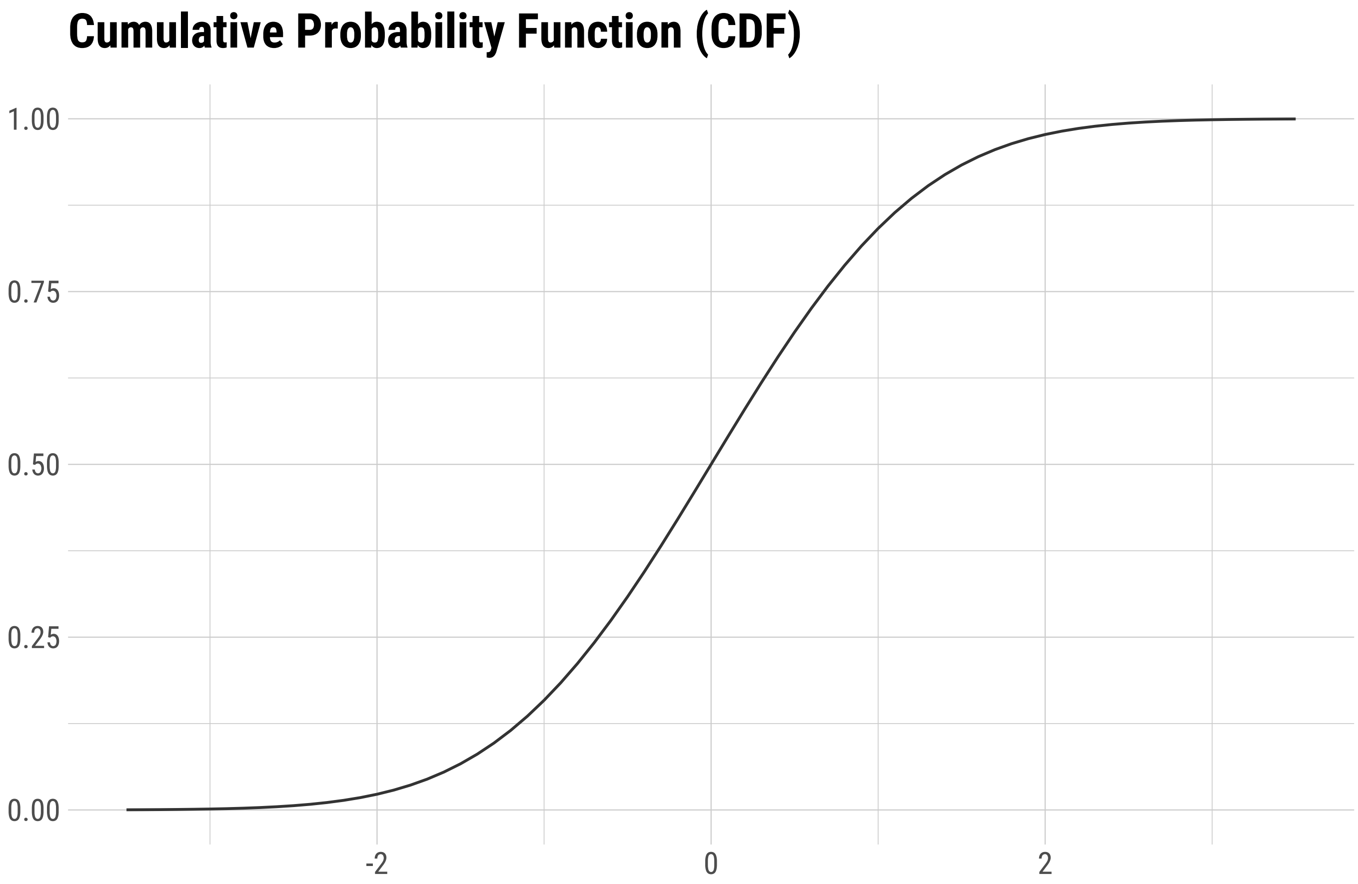
And its **Cumulative Density Function** (CDF) is given by

$$\Phi(z) = P\big[Z \leq z\big] = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-s^2/2\ u^2}\ du$$

`Probit` Models



Probit Models



Interpreting `Probit` Models

Interpreting `Probit` Models

In case x_k is a **continuous** explanatory variable, its marginal effect on $p(y_i = 1 \mid \mathbf{x}_i)$ is given by

$$\frac{\partial\, p(\mathbf{x}_i)}{\partial\, x_{ik}} = \frac{\partial\, \Phi(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki})}{\partial\, \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}} \cdot \frac{\partial\, \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}}{\partial\, x_{ik}}$$

$$\frac{\partial\, p(\mathbf{x}_i)}{\partial\, x_{ik}} = \phi(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki})\beta_k$$

In case x_k is a **discrete explanatory variable** (such as a *dummy* variable):

$$\Delta p(\mathbf{x}_i) = p(\mathbf{x}_i \mid x_k = 1) - p(\mathbf{x}_i \mid x_k = 0) =$$

$$\Delta p(\mathbf{x}_i) = \Phi(\beta_0 + \beta_1 x_{1i} + \beta_k) - \Phi(\beta_0 + \beta_1 x_{1i})$$

Interpreting `Probit` Models

For **Average Marginal Effects** (AME), the procedure is the same as with `Logit` coefficients.

The only **change** is in the CDF/PDF portions.

$$\frac{\partial P(y_i = 1 \mid \mathbf{x}_i)}{\partial x_{ij}} = \frac{\partial \Phi(\cdot)}{\partial x_{ij}} = \frac{\sum_{i=1}^{n} \phi(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k)}{n} \cdot \hat{\beta}_j$$

A `Probit` example

```
probit_model ← glm(inlf ~ nwifeinc + educ  + exper +
                  I(exper^2) + age + kidslt6 + kidsge6, data = mroz,
             family = binomial(link='probit'))
probit_model %>% tidy()
```

```
#> # A tibble: 8 × 5
#>   term        estimate std.error statistic  p.value
#>   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)  0.270    0.508        0.532 5.95e- 1
#> 2 nwifeinc    -0.0120   0.00494     -2.43  1.49e- 2
#> 3 educ         0.131    0.0254       5.15  2.55e- 7
#> 4 exper        0.123    0.0188       6.58  4.85e-11
#> 5 I(exper^2)  -0.00189  0.000600    -3.15  1.66e- 3
#> 6 age         -0.0529   0.00846     -6.25  4.22e-10
#> 7 kidslt6     -0.868    0.118       -7.34  2.21e-13
#> 8 kidsge6      0.0360   0.0440       0.818 4.14e- 1
```

As with the `Logit` case, these coefficients are **not** directly interpretable. Only their **signs**.

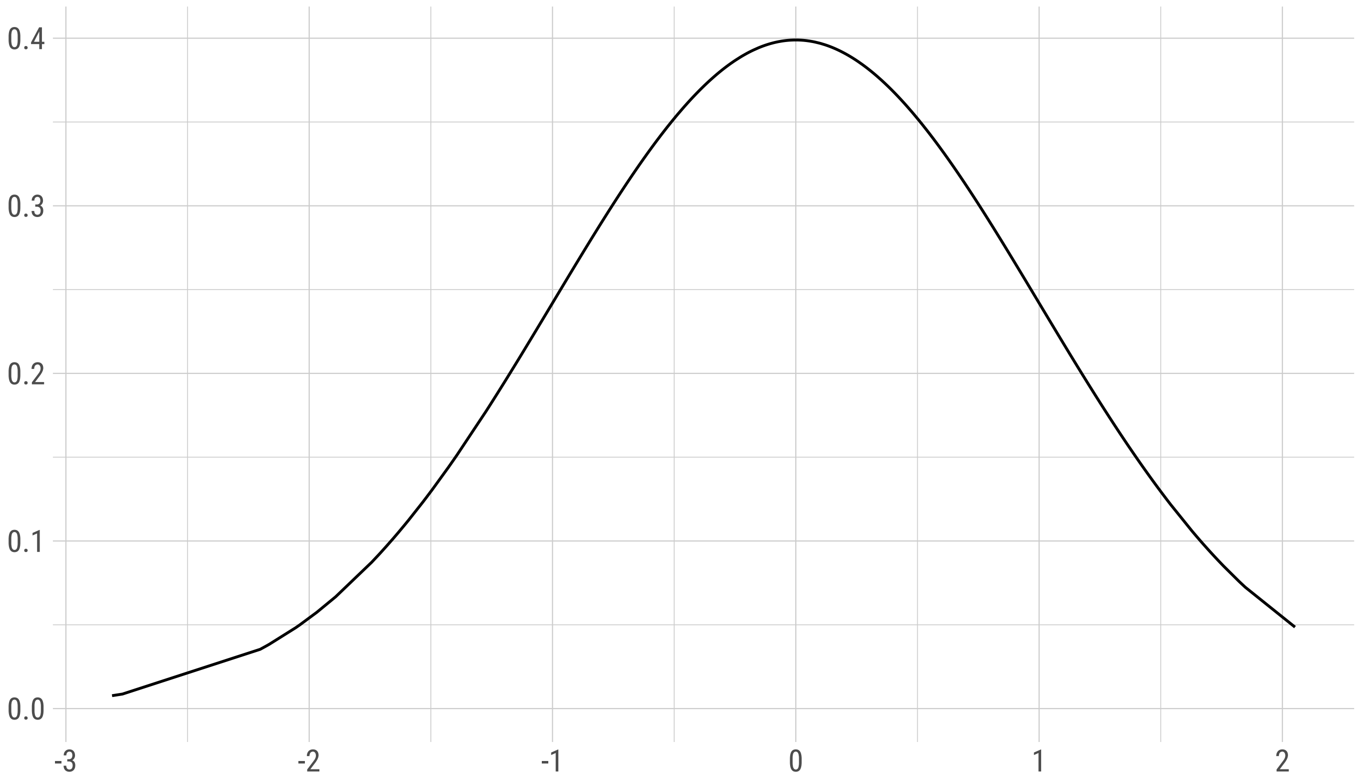
A Probit example

```
. probit inlf nwifeinc educ exper expersq age kidslt6 kidsge6

Iteration 0:   log likelihood =  -514.8732
Iteration 1:   log likelihood = -402.06651
Iteration 2:   log likelihood = -401.30273
Iteration 3:   log likelihood = -401.30219
Iteration 4:   log likelihood = -401.30219

Probit regression                                       Number of obs =    753
                                                        LR chi2(7)    = 227.14
                                                        Prob > chi2   = 0.0000
Log likelihood = -401.30219                             Pseudo R2     = 0.2206

------------------------------------------------------------------------------
        inlf | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
    nwifeinc |  -.0120237   .0048398    -2.48   0.013    -.0215096   -.0025378
        educ |   .1309047   .0252542     5.18   0.000     .0814074     .180402
       exper |   .1233476   .0187164     6.59   0.000     .0866641    .1600311
     expersq |  -.0018871      .0006    -3.15   0.002     -.003063   -.0007111
         age |  -.0528527   .0084772    -6.23   0.000    -.0694678   -.0362376
     kidslt6 |  -.8683285   .1185223    -7.33   0.000    -1.100628    -.636029
     kidsge6 |    .036005   .0434768     0.83   0.408     -.049208    .1212179
       _cons |   .2700768    .508593     0.53   0.595    -.7267473    1.266901
------------------------------------------------------------------------------
```

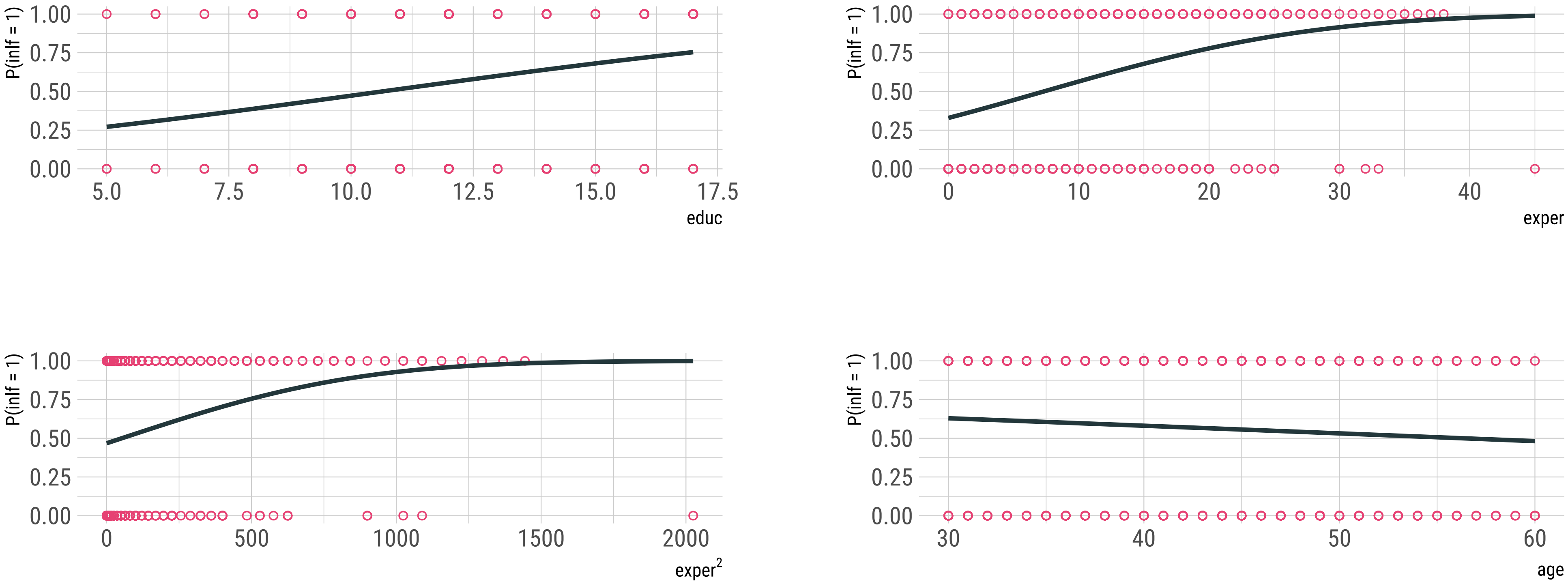
As with the `Logit` case, these coefficients are **not** directly interpretable. Only their **signs**.

A `Probit` example

The **PDF** for this estimated model looks like this:



A `Probit` example



A `Probit` example

Average Marginal Effects:

```
#>    Variable          AME
#> 1 intercept  0.081226125
#> 2  nwifeinc -0.003616176
#> 3      educ  0.039370095
#> 4     exper  0.037097345
#> 5   exper^2 -0.000567546
#> 6       age -0.015895665
#> 7   kidslt6 -0.261153464
#> 8   kidsge6  0.010828887
```

How to **interpret** these coefficients?

Model comparison

In terms of **coefficients**:

```
#>   Coefficient           LPM        Logit       Probit
#> 1 (Intercept)  0.5855192249  0.425452376  0.270073573
#> 2    nwifeinc -0.0034051689 -0.021345174 -0.012023637
#> 3        educ  0.0379953030  0.221170370  0.130903969
#> 4       exper  0.0394923895  0.205869531  0.123347168
#> 5  I(exper^2) -0.0005963119 -0.003154104 -0.001887067
#> 6         age -0.0160908061 -0.088024375 -0.052852442
#> 7     kidslt6 -0.2618104667 -1.443354143 -0.868324680
#> 8     kidsge6  0.0130122346  0.060112222  0.036005611
```

Model comparison

In terms of **Average Marginal Effects**:

```
#>    Variable         Logit       Probit
#> 1 intercept  0.0759771297  0.081226125
#> 2  nwifeinc -0.0038118135 -0.003616176
#> 3      educ  0.0394965238  0.039370095
#> 4     exper  0.0367641056  0.037097345
#> 5   exper^2 -0.0005632587 -0.000567546
#> 6       age -0.0157193606 -0.015895665
#> 7   kidslt6 -0.2577536551 -0.261153464
#> 8   kidsge6  0.0107348186  0.010828887
```

Goodness-of-fit

Goodness-of-fit

The usual R^2 and adjusted R^2 measures are not **satisfactory** for binary dependent variable models.

However, in case **goodness-of-fit** is of interest, we can use the **McFadden's pseudo R^2** measure.

$$R^2 = 1 - \frac{\ell(\hat{\beta})}{\ell(\bar{y})}$$

where $\ell(\hat{\beta})$ is the log-likelihood of the fitted model, and $\ell(\bar{y})$ is the log-likelihood of a restricted model, only containing an intercept term.

For our estimated `Logit` and `Probit` models, the pseudo-R^2 measures are **0.219** and **0.2205**, respectively.

We will calculate these next time.

Next time: Binary models in practice