Binary Models – Logit and Probit

Philipp Broniecki

University College London

6 June 2016

Binary Dependent Variables

Binary dependent variables are frequent in social science research...

- ... why does someone vote for or against Brexit?
- ... why do dictators hold elections?
- ... why do UN peacekeeping missions succeed/fail?

Linear Regression?

LPM

The linear probability model relies on linear regression to analyze binary variables.

$$\begin{aligned} Y &= \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + ... + \varepsilon \\ Pr(Y_i = 1|X_1, X_2, ...) &= \beta_0 + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + ... \end{aligned}$$

Advantages:

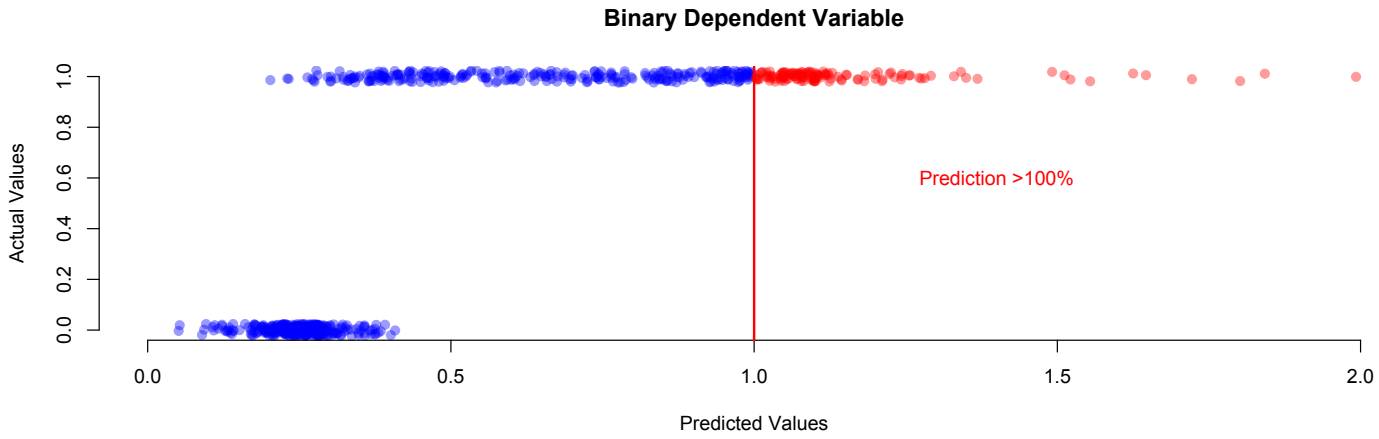
- We can use a well-known model for a new class of phenomena
- Easy to interpret the marginal effects of X

Problems with Linear Probability Model

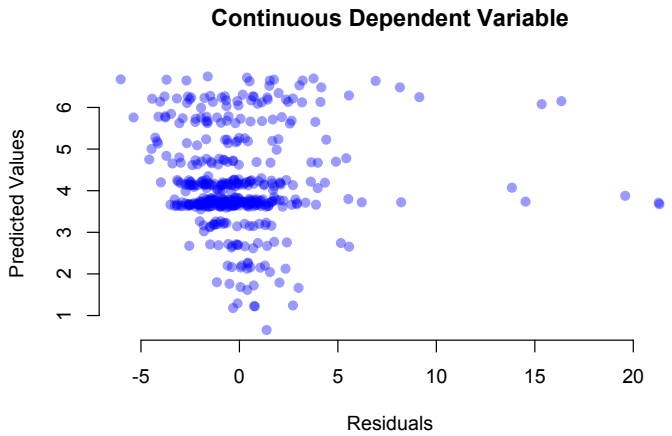
The linear model needs a continuous dependent variable, if the dependent variable is binary we run into problems:

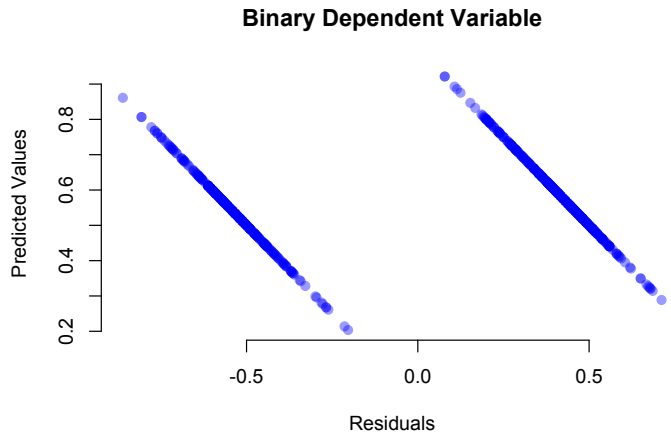
- Predictions, $\hat{y}$, are interpreted as probability for $y = 1$
  $\rightarrow P(y = 1) = \hat{y} = \beta_0 + \beta_1 X$, can be above 1 if X is large enough
  $\rightarrow P(y = 0) = \hat{y} = \beta_0 + \beta_1 X$, can be below 0 if X is small enough
- The errors will not have a constant variance.
  $\rightarrow$ For a given X the residual can be either (1-β_0-$\beta_1 X$) or ($\beta_0 + \beta_1 X$)
- The linear function might be wrong (functional form)
  $\rightarrow$ Imagine you buy a car. Having an additional 1000£ has a very different effect if you are broke or if you already have another 12,000£ for a car.

Predictions can lay outside $I = [0, 1]$



Residuals if the dependent variable is binary:





What is a Function

A function maps values from x on exactly one value of y.

We say that x is the argument of a function $f(\cdot)$.

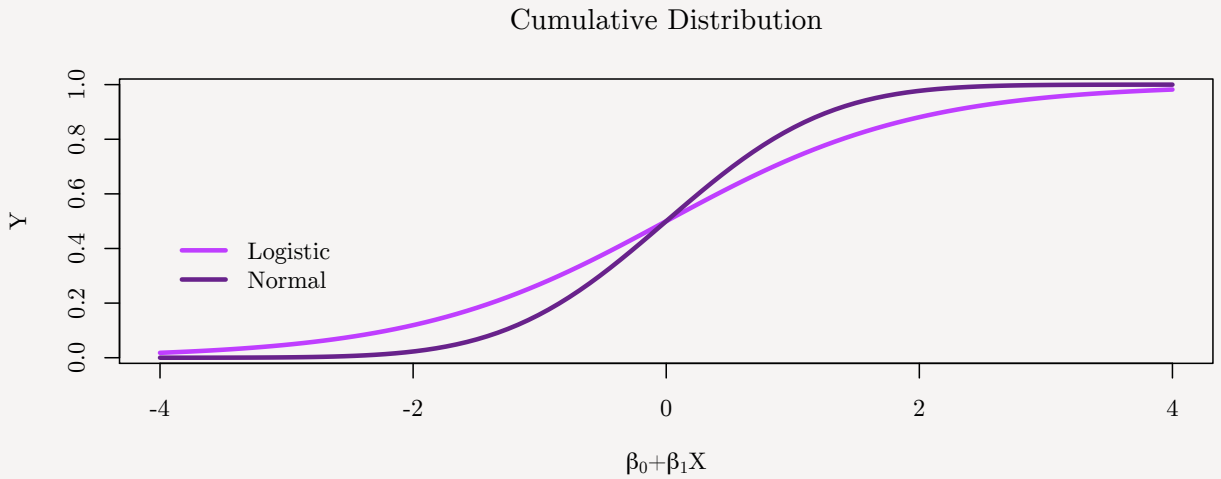
You actually do know functions (many!), e.g.:

- $f(x) = x^2$
- $f(x) = 2 + 3 \cdot x$
- $f(x) = \beta_0 + \beta_1 x$

$\rightarrow$ Think of a function as a rule. It tells you what to do with a generic x once you apply it to a specific x it returns y:
$f(x) = 2 + 3 \cdot x$, for $x = 2 \rightarrow f(x) = 8$

Predictions should only be within $I = [0, 1]$

- We want to make predictions in terms of probability
- We can have a model like this: $P(y_i = 1) = F(\beta_0 + \beta_1 X_i)$ where $F(\cdot)$ should be a function which never returns values below 0 or above 1
- There are two possibilities for $F(\cdot)$: cumulative normal (Φ) or logistic (Δ) distribution



Logit and Probit

- This is good news: We now have a model where $\hat{y} \in [0, 1]$
  $\rightarrow$ All predictions are probabilities
- We have two possible models to use
  $\rightarrow$ The **logit model** is based on the cumulative logistic distribution (Δ)
  $\rightarrow$ The probit model is based on the cumulative normal distribution (Φ)

Logit or Probit

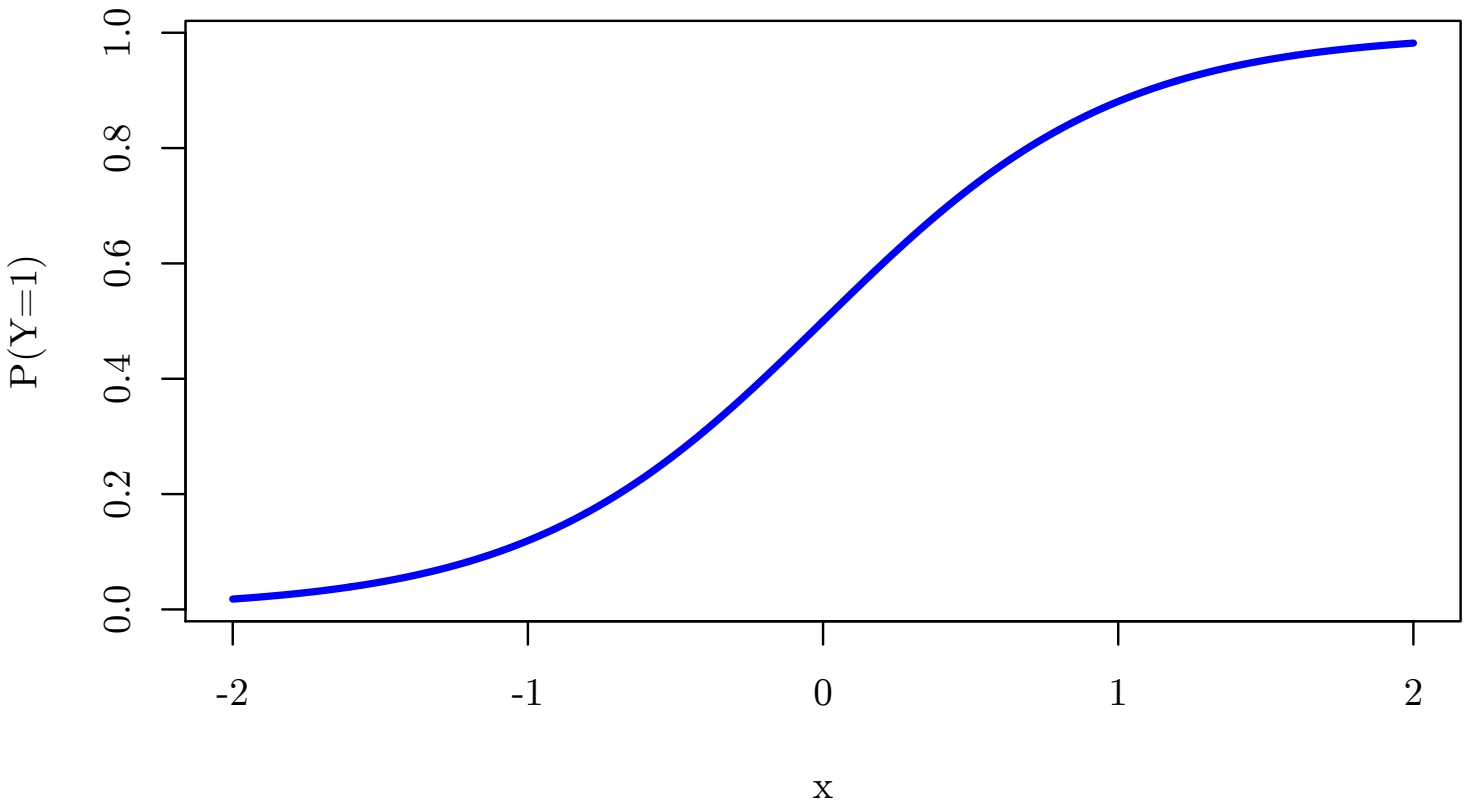
It does not matter which one you use!

We will use logit more often because we can write
$\Delta(x) = \frac{1}{1+\exp(-x)}$,
while probit models are tricky: $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp(\frac{-(x)^2}{2}) dx$

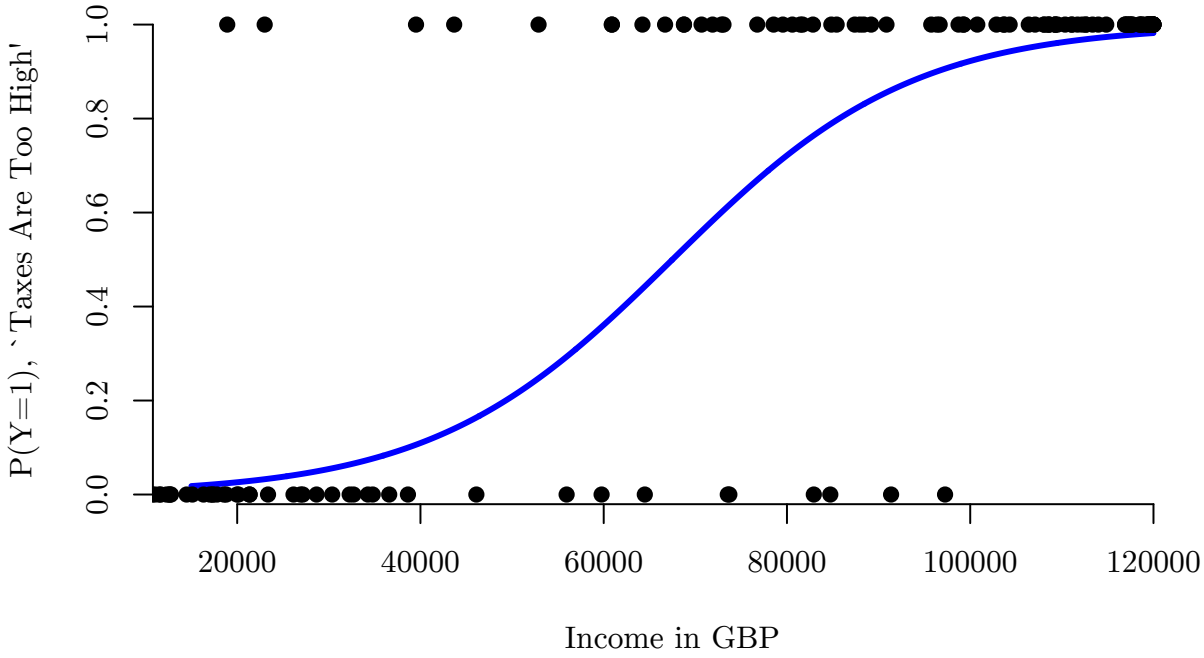
Logit Model

The logit model is then: $P(y_i = 1) = \frac{1}{1+\exp(-\beta_0-\beta_1 X_i)}$

For $\beta_0 = 0$ and $\beta_1 = 2$ we get:

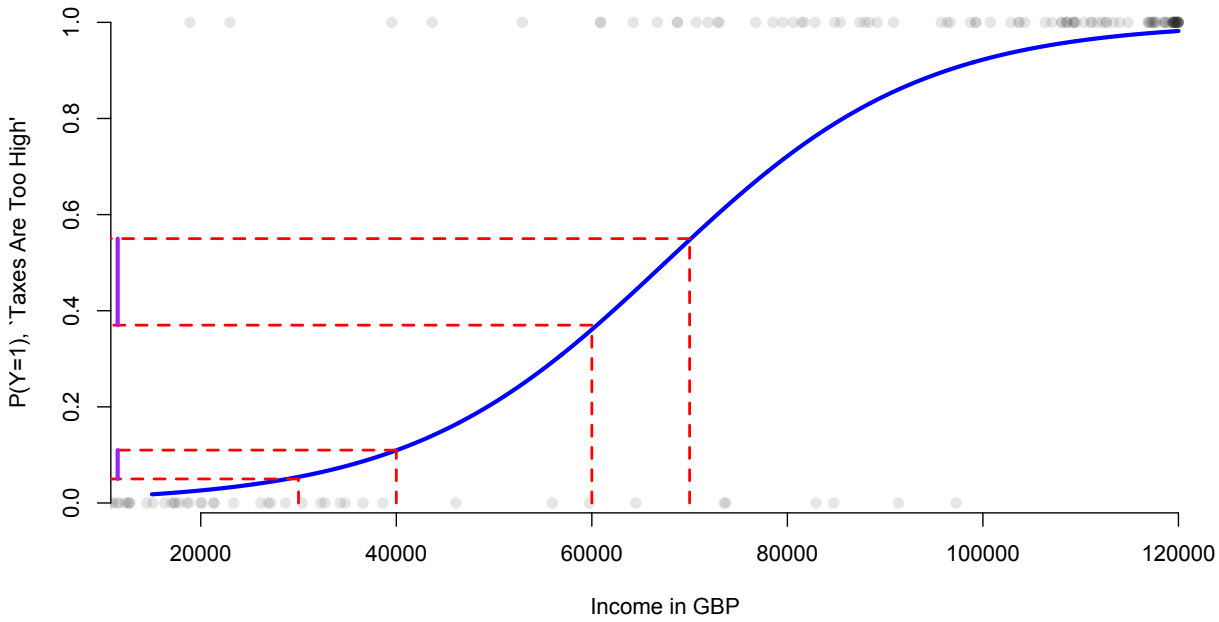


Logit Model: Example



- We can make a prediction by calculating:
  $P(y = 1) = \frac{1}{1+\exp(-\beta_0-\beta_1 \cdot X)}$

Logit Model: Example



- Depending on where we add 10,000£ we get a different marginal effect
  $\rightarrow$ because of our different functional form (s-shaped)

Latent Variable - Where the Error Is Hiding

The Latent Variable: y^*

We can call the part $\beta_0 + \beta_1 X + \beta_2 X_2 + \varepsilon$ the latent variable y^*.

- We never observe $y^* (= \beta_0 + \beta_1 X + \beta_2 X_2 + \varepsilon)$
- Think of y^* as the difference in utility an agent gets from choosing one option over the other option

We never see the latent variable, but we observe y which will be 0 or 1

- On average, we expect that y^* is larger than 0 when $y = 1$ and that $y^* < 0$ when $y = 0$

Example: Women in the 1980s and Labour Market

```
> m1 <- glm(inlf ~ kids + age + educ, dat = my.data, family = binomial(logit))
```

```
             Estimate Std. Error
(Intercept) -0.11436871 0.73459406
kids        -0.50348811 0.19932135
age         -0.03108088 0.01136842
educ         0.16901771 0.03505405
```

Example: Women 1980 (2)

```
Call:
glm(formula = inlf ~ kids + educ + age, family = binomial(logit), data = my.data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.11437    0.73459  -0.156  0.87628
kids        -0.50349    0.19932  -2.526  0.01154 *
educ         0.16902    0.03505   4.822 1.42e-06 ***
age         -0.03108    0.01137  -2.734  0.00626 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

- Only interpret direction and significance of a coefficient
- The test statistic always follows a normal distribution (z)

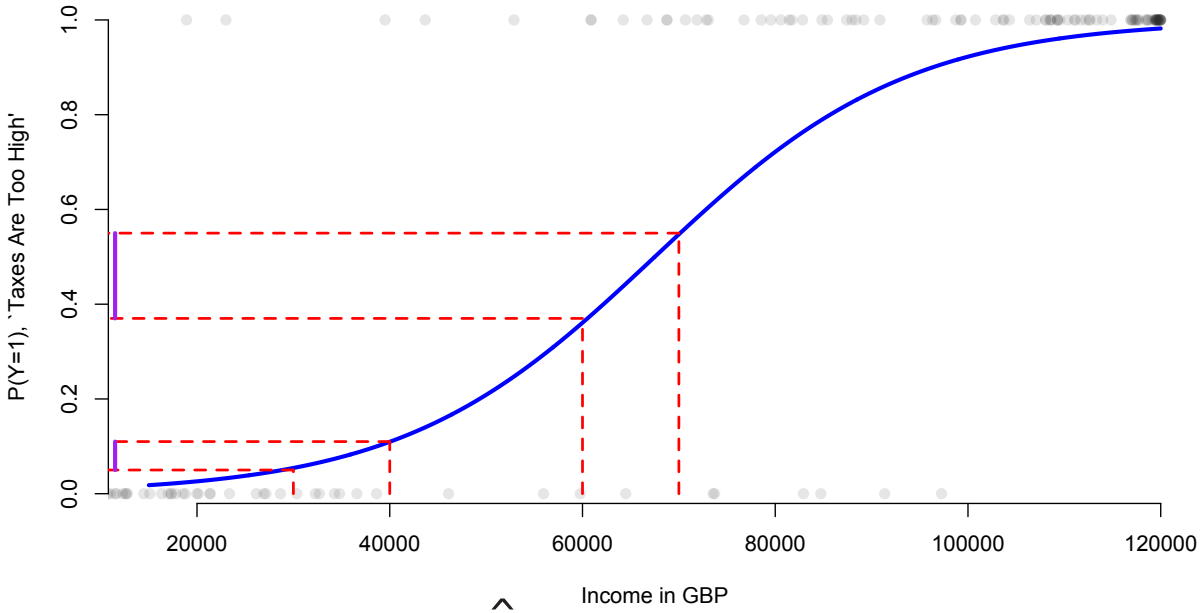
Interpreting Estimation Results

```
glm(formula = inlf ~ kids + educ + age, family = binomial(logit),
    data = data1)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.11437    0.73459  -0.156  0.87628
kids        -0.50349    0.19932  -2.526  0.01154 *
educ         0.16902    0.03505   4.822 1.42e-06 ***
age         -0.03108    0.01137  -2.734  0.00626 **
```

- How can we generate a prediction for a woman with no kids, 13 years of education, who is 32?
  - Compute first the prediction on y^*, i.e. just compute $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
  - $P(y = 1) = \frac{1}{1+exp(0.11+.50 \cdot 0-0.17 \cdot 13+0.03 \cdot 32)} = \frac{1}{1+exp(-1.09)} = 0.75$

Interpreting Estimation Results



- You cannot just report $\hat{\beta}$ and say that this is the marginal effect
- You can calculate the marginal effect for a specific profile

  Profile: Is a specific hypothetical observation, a profile is a set of values for each X
- You can also say what the change in predicted probabilities is

  $\rightarrow$ As a person with 60k £ gains another 10k we expect them to be 8% more likely to agree that taxes are too high

Interpreting Estimation Results

Change in Predicted Probability

An easier way, the one your grandpa understands, is to compute the predicted probabilities for two different profiles where you only change the value on one variable

- P(woman works) $= F(age, kids)$
- Do it for a young woman (age=30) and change kids from 0 to 1, could yield $p_{k=0} = 80\%$ and $p_{k=1} = 65\%$.

Example Interpretation

```
library(Zelig)
# step 1: estimate model
m1 <- zelig(inlf ~ kids + age + educ + exper + huseduc + huswage,
            model = "logit", data = df, cite = FALSE)

# step 2: set X variables to desired values (e.g. average)
avg.women <- setx(m1, kids = median(df$kids), age = mean(df$age), educ = mean(df$educ),
                  exper = mean(df$exper), huseduc = mean(df$huseduc), huswage = mean(df$huswage))

# step 3: simulate
sim.out <- sim(m1, x = avg.women)

# finally check outcome:
summary(sim.out)

 sim x :
 -----
ev
          mean         sd       50%      2.5%     97.5%
[1,] 0.5745597 0.02561208 0.5746364 0.5232509 0.6244216
pv
         0     1
[1,] 0.402 0.598
```

Interpreting Model Quality

- A good model makes a lot of correct predictions
  $\rightarrow$ If $\hat{p}_i$ is larger than 0.5 we predict a 1 otherwise a 0
- A good model improves upon the naive guess
  $\rightarrow$ The naive guess is to predict the modal category for all outcomes (equivalent to not having a model)
  - Modal category: central tendency of categorical variable. The category with most observations.
  - Naive model: A model with no explanatory variables in it, i.e. $P(Y_i = 1) = \frac{1}{1+exp(-\beta_0)}$

Interpreting Model Quality

Correctly Predicted Cases

We will count how many cases are predicted correctly:

```
           -------- PREDICTED --------
Observed |         0            1     |      Total
-----------+--------------------------+-----------
     0     |       100            225   |        325
     1     |        76            352   |        428
-----------+--------------------------+-----------
   Total   |       176            577   |        753
```

- We have here $\frac{100+352}{753} = 0.60$
- Modal category (*inlf*) had 56.8% ($= \frac{428}{428+325}$)
- $\rightarrow$ This is a tiny improvement given that we added so much information.

Model Quality (2)

```
> observed <- m1$model$inlf
> pred.probs <- predict(mod1, type="response")
> exp.vals <- ifelse(pred.probs > 0.5, yes = 1, no = 0)
> qual.pred <- table(observed,exp.vals)
> qual.pred
        exp.vals
observed   0   1
       0 100 225
       1  76 352
> summary(observed)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  1.0000  0.5684  1.0000  1.0000
> sum(diag(qual.pred)) / sum(qual.pred)
[1] 0.6002656
```

Female Employment: Model Quality

```
> ## Model quality
> observed <- data1$inlf
> exp.vals <- rep(0,length(observed))
> pred.probs <- predict(m1, type="response")
> threshold <- .5
> exp.vals[which(pred.probs > threshold)] <- 1
> qual.pred <- table(observed,exp.vals)
> qual.pred
        exp.vals
observed   0   1
       0 192 133
       1 103 325
>
> (qual.pred[1,1] + qual.pred[2,2])/sum(qual.pred)
[1] 0.686587
> table(observed)/sum(table(observed))
observed
        0         1
0.4316069 0.5683931
```

- With no model we would guess the modal category for everybody and correctly predict 56.8%
- With our model we actually predict 68.7% of all observations correctly
- Nice improvement, we reduce the error rate by about 27% (error: 43% with no model, 31.3% with a model)

Example

```
> screenreg(list(m1,m2), stars=c(0.01,0.05,0.1))

========================================
                Model 1      Model 2
----------------------------------------
(Intercept)       -0.14         0.06
                  (0.79)       (0.81)
kids              -0.17        -0.17
                  (0.22)       (0.22)
age               -0.06 ***    -0.06 ***
                  (0.01)       (0.01)
educ               0.15 ***     0.21 ***
                  (0.04)       (0.05)
exper              0.12 ***     0.12 ***
                  (0.01)       (0.01)
huseduc                        -0.06
                               (0.04)
huswage                        -0.03
                               (0.02)
----------------------------------------
AIC              891.43       888.89
BIC              914.55       921.25
Log Likelihood  -440.72      -437.44
Deviance         881.43       874.89
Num. obs.        753          753
========================================
*** p < 0.01, ** p < 0.05, * p < 0.1
```

Testing Joint Hypotheses: The LR Test

Log-Likelihood Value

The model is not anymore estimated by minimizing the sum of squared residuals.
We maximize the log-likelihood function ($\ell\ell$) and the values of β_0 and β_1 for which the $\ell\ell$ is maximal are then our estimates. In our example on slide 18 $\ell\ell = -496.01418$.

$$\ell\ell = \sum_{i=1}^{n} y_i \cdot \left[\frac{1}{1+\exp(-\beta_0-\beta_1 X)}\right] + \sum_{i=1}^{n}(1-y_i)\left[1-\left(\frac{1}{1+\exp(-\beta_0-\beta_1 X)}\right)\right]$$

Log-Likelihood Ratio Test

The F-test for linear models could be expressed as a test statistic which was a function of the two R^2 values.
The equivalent is the LR-test: $\chi^2_k = 2 \cdot (\ell_{ur} - \ell\ell_r)$

Testing Joint Hypotheses: The LR Test (2)

```
                Model 1      Model 2
-----------------------------------------
(Intercept)       -0.14         0.06
                  (0.79)       (0.81)
kids              -0.17        -0.17
                  (0.22)       (0.22)
age               -0.06 ***    -0.06 ***
                  (0.01)       (0.01)
educ               0.15 ***     0.21 ***
                  (0.04)       (0.05)
exper              0.12 ***     0.12 ***
                  (0.01)       (0.01)
huseduc                        -0.06
                               (0.04)
huswage                        -0.03
                               (0.02)
-----------------------------------------
Log Likelihood  -440.72      -437.44
Num. obs.        753          753
=========================================
*** p < 0.01, ** p < 0.05, * p < 0.1
> anova(m1,m2, test="Chisq")
Model 1: inlf ~ kids + age + educ + exper
Model 2: inlf ~ kids + age + educ + exper +
huseduc + huswage
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       748     881.43
2       746     874.89  2    6.545  0.03791 *
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Let's test whether the husband even matters to understand why a woman works or not:

- We have $\ell\ell_{ur} = -437.44$ and $\ell\ell_r = -440.72$
- $\chi^2_2 = 6.55$
- $p-$value$= 0.038$
- Despite neither variable being significant, we cannot rule out that (education and wage of) husbands matter