Regression with binary outputs

Logit and Probit

Paul Johnson[1]

[1] CRMDA, University of Kansas

2018

Outline

KU

Outline ...

- Small Sample with Separation

8. Testing Statistical Significance

9. Model Goodness
   - Percent Correctly Predicted and ROC
   - LLR equivalent of an F test
   - Deviance
   - Why no R square
   - Hosmer and Lemeshow test

Overview

- Dependent variable is categorical.
  - For discussion, we refer to $Y_i \in \{0, 1\}$
  - But the numbers have no substantive meaning. They are just labels, could write No, Yes or $Fail, Pass$
  - Temptation to treat the labels $\{0, 1\}$ as numbers
- Estimated by Maximum Likelihood

  ML idea: Choose estimates to make the observed outcomes most likely

  - proposed by Ronald Fisher 1905-1922
- Alternative models we consider–logit and probit–are members of the "family" of models in the Generalized Linear Model (McCullagh & Nelder, 1989)

Outline

KU

Outline ...

- Small Sample with Separation

8. Testing Statistical Significance

9. Model Goodness
   - Percent Correctly Predicted and ROC
   - LLR equivalent of an F test
   - Deviance
   - Why no R square
   - Hosmer and Lemeshow test

Terminology

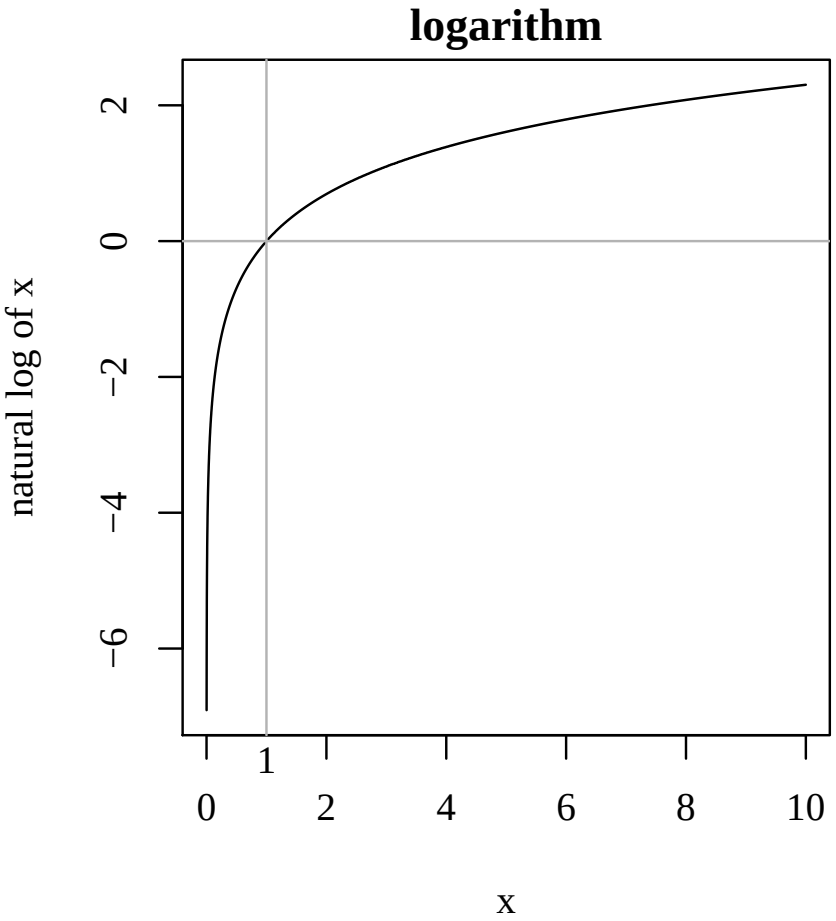
- This is a terminology section, possibly a refresher
- We need to be comfortable with $\log$, $\exp$, PDF and CDF in order to make progress in what follows

log

$\log_b(x)$ answers question "to what power must b be raised in order to equal x?"

Facts

1. $\log_b(1) = 0$, no matter what b is
2. $\log_b(x)$ is undefined if $x \leq 0$
3. e is Euler's constant. It is the most widely used value of b.
   1. the slope of $\log_e(x)$ has slope $1/x$
   2. called the "natural log", often written $\ln(x)$
   3. if I omit b, assume it is e
4. Handy facts:
   1. $\log(x/y) = \log(x) - \log(y)$
   2. $\log(x \cdot y) = \log(x) + \log(y)$
   3. $\log(x^k) = k \cdot \log(x)$

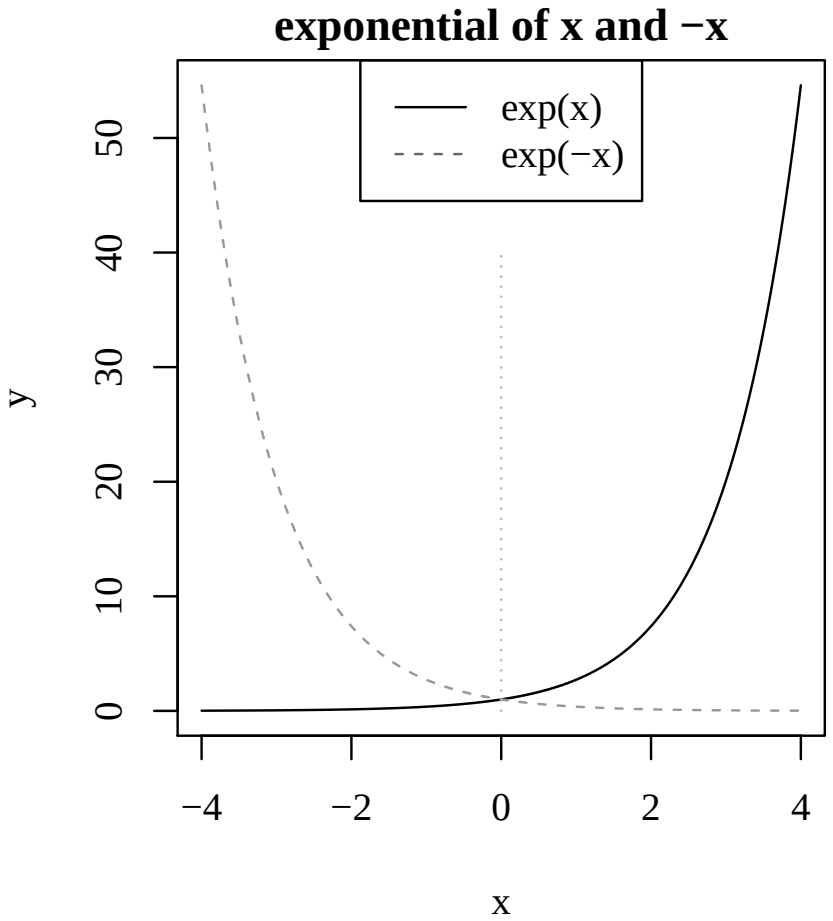


exponential function

$\exp(x)$ means e^x. Why $\exp(x)$? Typesetters prefer to avoid superscripts.

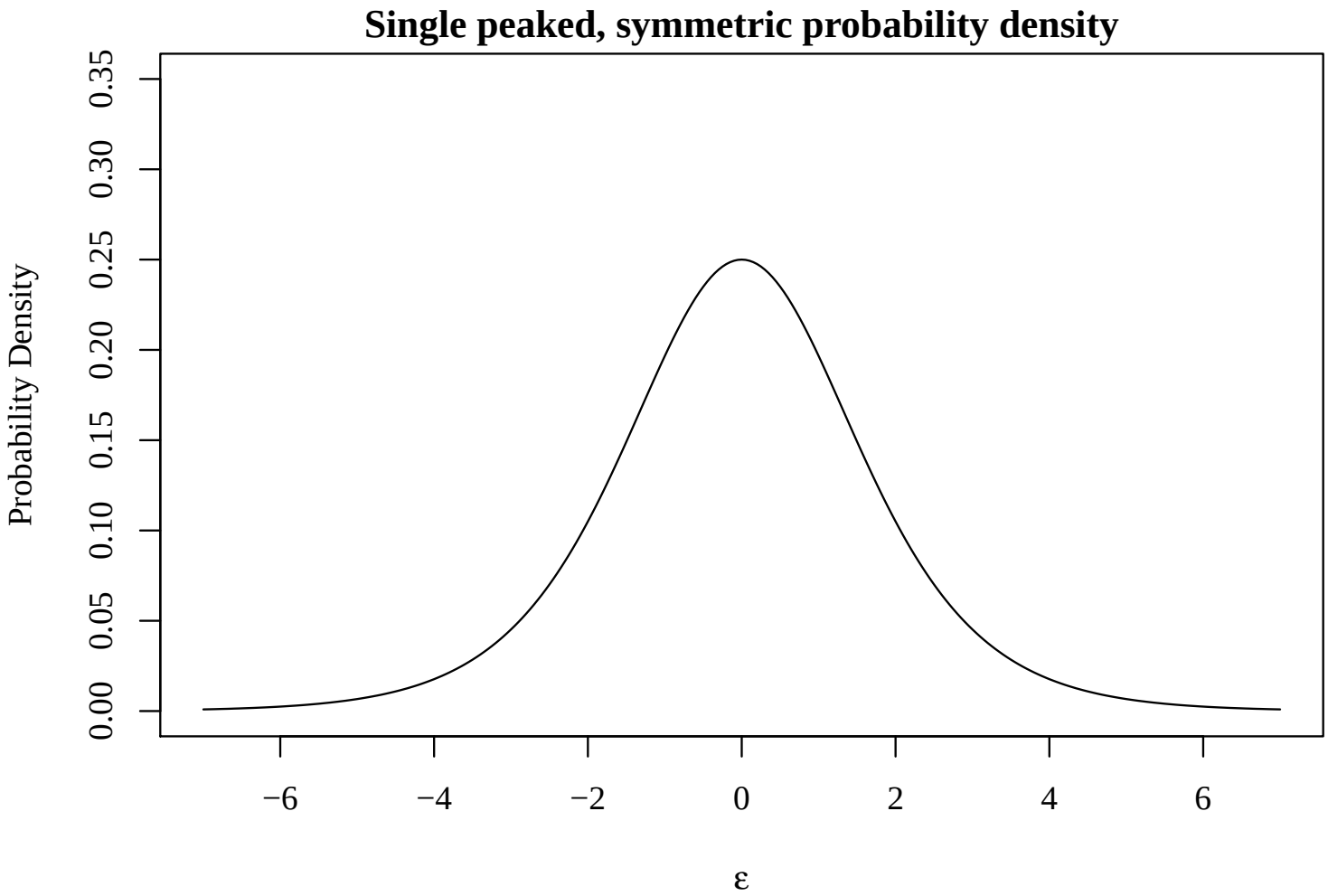
Handy facts:

1. $\exp(x)$ is
   1. $\exp(0) = 1$
   2. $\exp(x) > 0$, its always positive,
   3. $\exp(x + y) = \exp(x) * \exp(y)$
2. $\exp(-x) = \frac{1}{\exp(x)}$
   1. general fact about exponents: $x^{-k} = \frac{1}{x^k}$
3. $\exp$ and $\log$ are inverses of each other
   1. $x = \exp(\log(x))$
   2. $x = \log(\exp(x))$



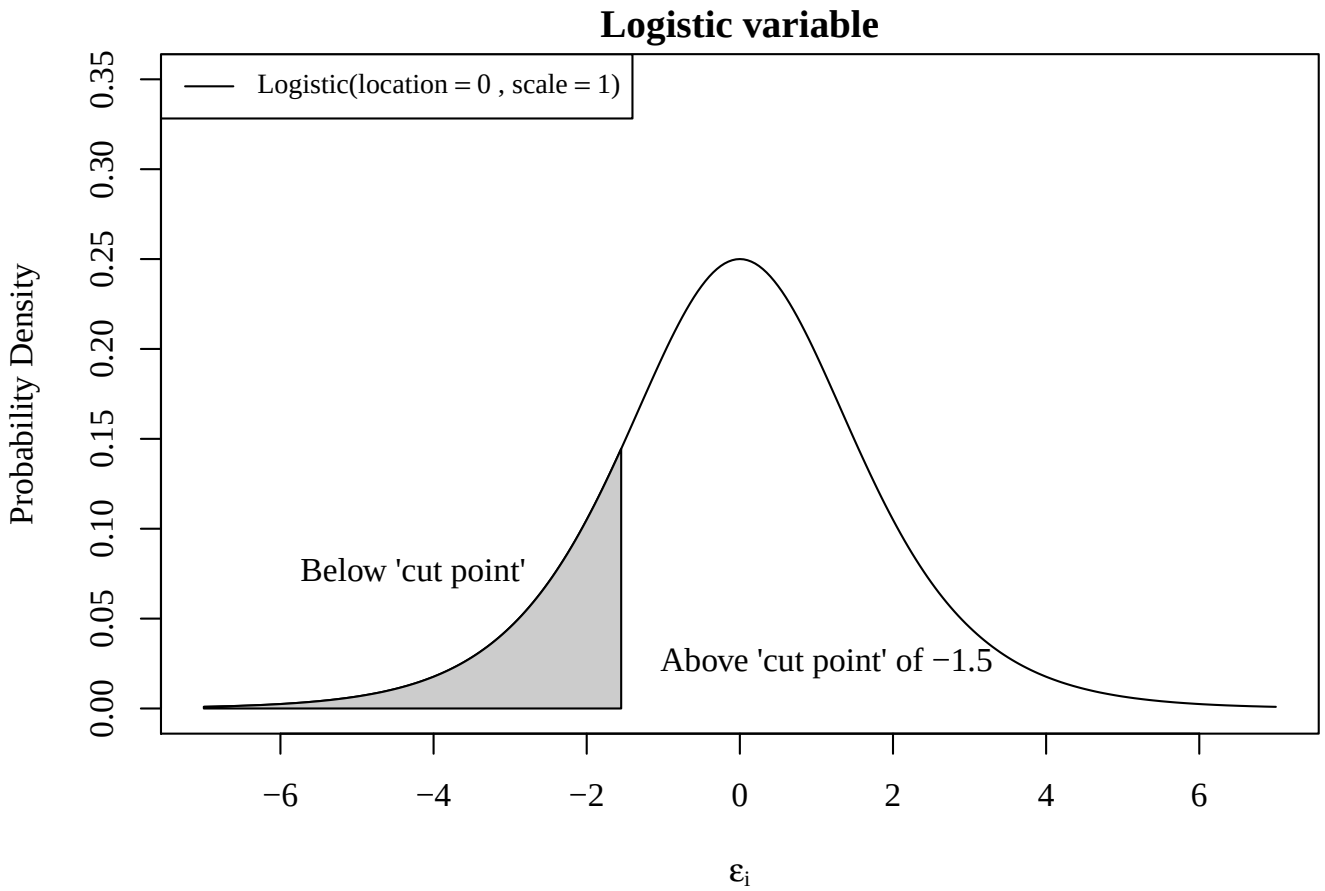
Vital: b/c $\exp$ is always positive and exists for all $+/-$ inputs, it is often used to transform input into a positive value.

A Probability Density Function (PDF)



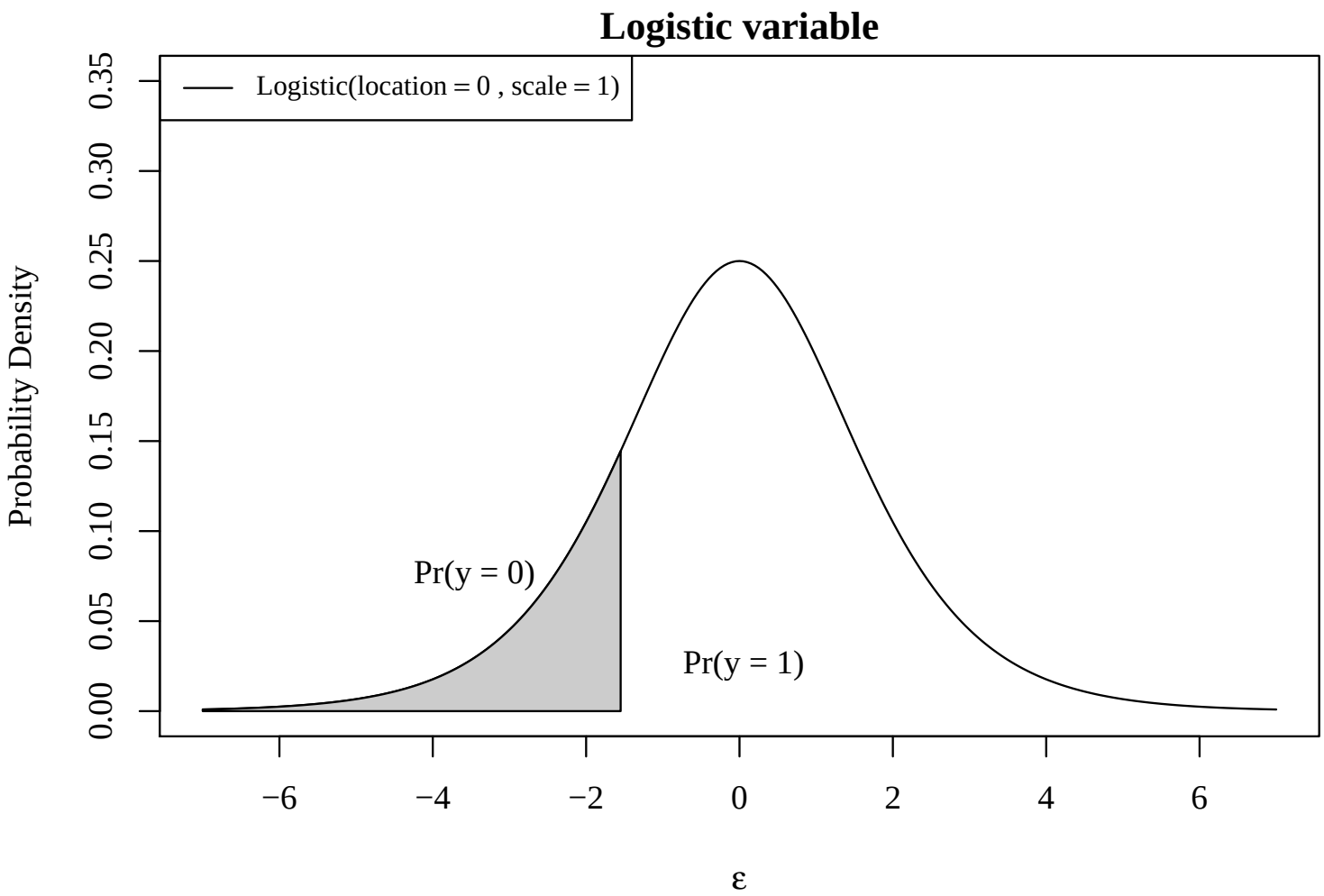
The random variable $\varepsilon \in (-\infty, \infty)$ has PDF $f(\varepsilon)$

Example: The Logistic Distribution

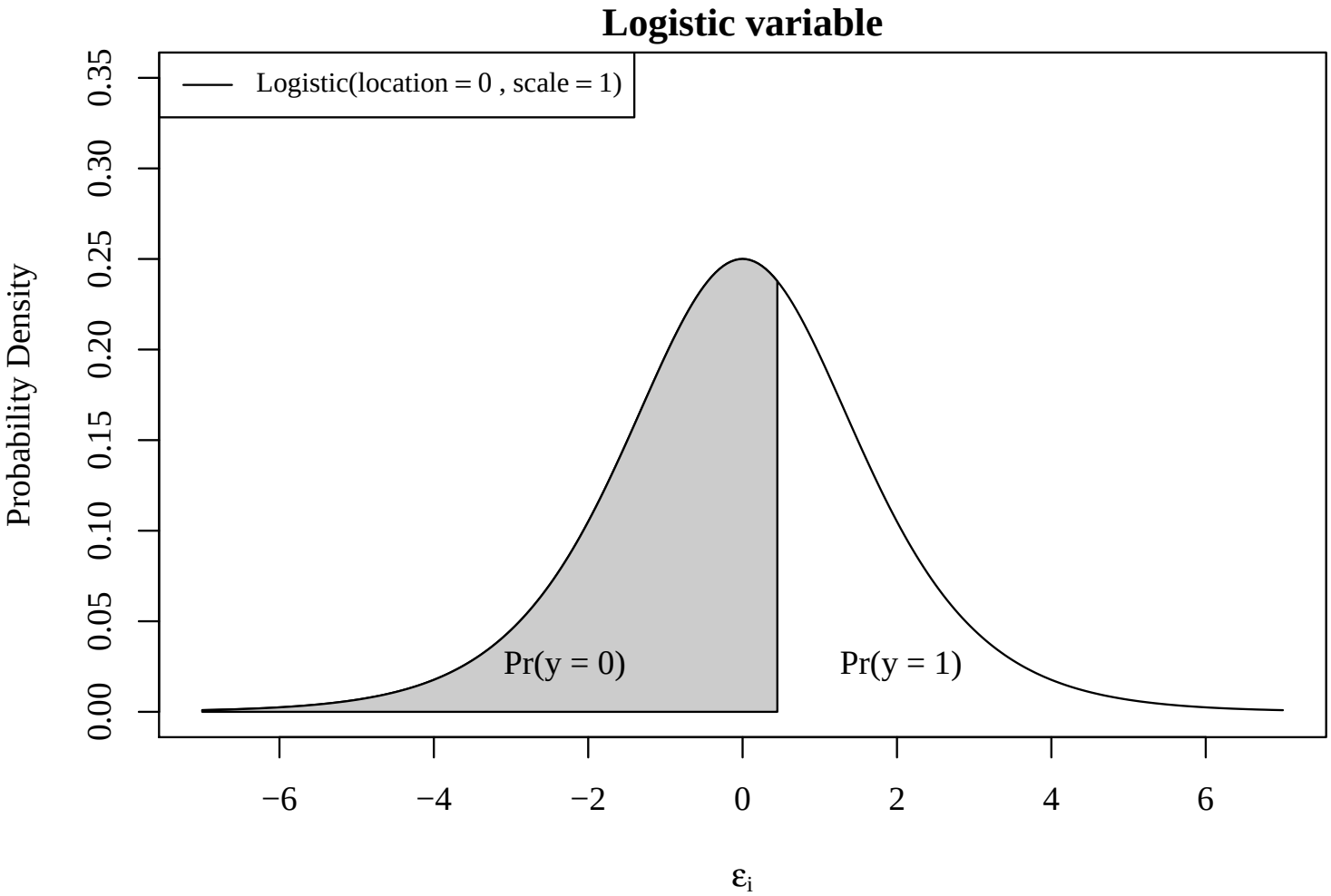


Can calculate area under curve after setting dividing point

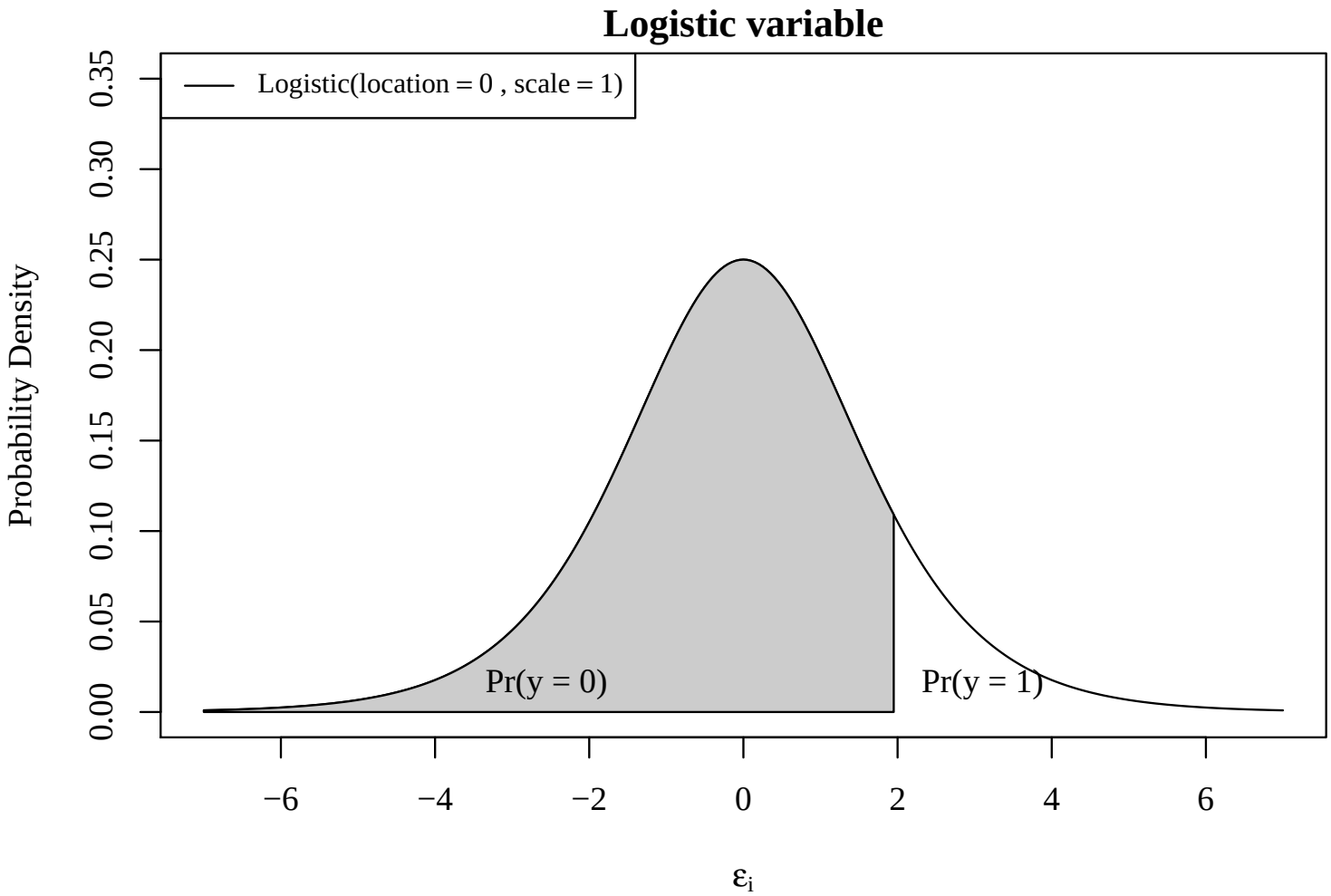
Use the areas to represent chance $y = 0$ or $y = 1$



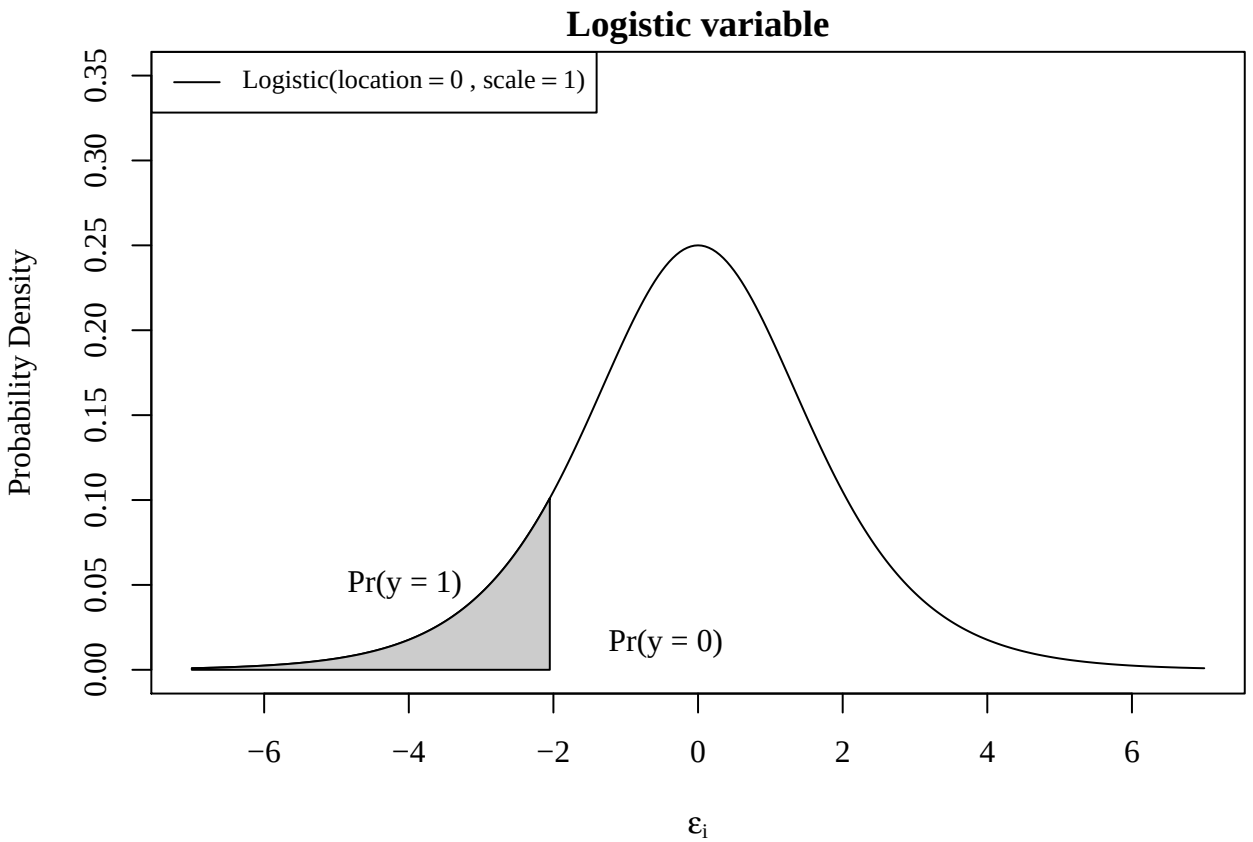
Increase the cut point, change the probabilities



Increase the cut point



Easy to swap 0 and 1



- The PDF of ε is symmetric, can swap the sides in the graph
  - i.e., change divider from 2 to -2, then area below becomes area above.

The area on the left is Cumulative Probability

- About the dividing point. Lets don't worry too much about notation, just call it τ ("tau").
- Nature's data generating process to create $y_i \in \{0, 1\}$ (our statistical model!)
  - Draw a random ε, then let $\varepsilon \leq \tau$ decide if y_i as "low" or "high"
- The **Cumulative Distribution Function** is customary term for that area.
  - CDF is the chance of an outcome smaller than τ
  - Customary also to refer to it by the capital letter $F(\tau)$ if the PDF was $f(\varepsilon)$

The area on the left is Cumulative Probability ...

- Mathematically, the probability that $y = 0$ is an integral

$$F(\tau) = \int_{-\infty}^{\tau} f(\varepsilon) d\varepsilon$$

- Mathematically, the probability that $y = 1$ is area on the high side.

$$1 - F(\tau) = \int_{\tau}^{\infty} f(\varepsilon) d\varepsilon$$

- This is confusing, I know, but for a dividing point k.

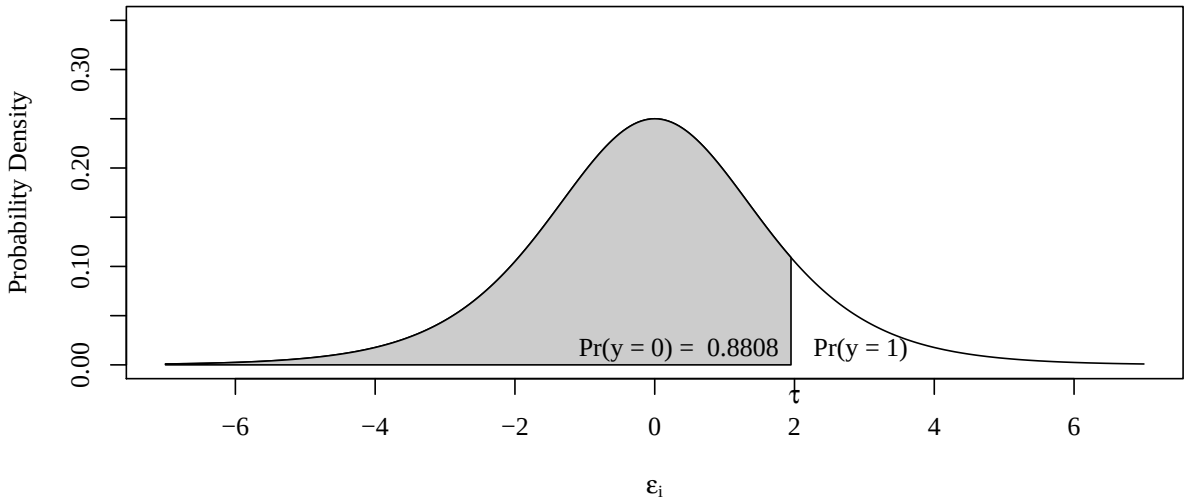
$$1 - F(-k) = F(k) \tag{1}$$

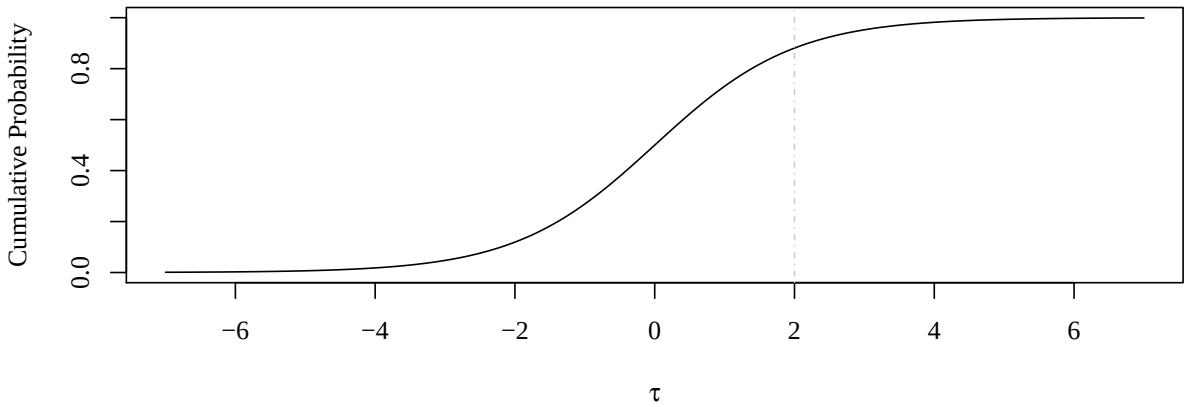
- Because software programs differ in the "choice of sides" for $y_i = 0$, we see different signs reported for regression models (SAS is opposite of Stata, for example)

The area on the left is Cumulative Probability ...

- Clearly,
  - if $\tau \to -\infty$, then $F(\tau) \to 0$
  - if $\tau \to \infty$, then $F(\tau) \to 1$
  - Between those extremes, the graph of the CDF is
    - always increasing (or flat, but we usually have increasing)
    - S-shaped

Compare PDF and CDF





Looking Forward

- Will build a predictive model that sets the dividing line each case.
- All Logistic programs will give technically consistent results, but may have different signs on β because there is "artistic license" in putting $y = 0$ or $y = 1$ on the left side of the graph.
- *I'm getting it wrong on a regular basis, so you are not alone.*

The Linear Predictor

- **Linear Predictor** is the right hand side of a regression, *BUT* with no error term

$$\beta_0 + \beta_1 x1_i + \beta_2 x2_i + \beta_3 x3_i \tag{2}$$

- Shorthand $X_i\beta$.
  - X_i is a "row" of predictor values, usually with 1 in the first position
  - β is a column vector of coefficients

$$[1, x1_i, x2_i, x3_i] \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \beta_0 + \beta_1 x1_i + \beta_2 x2_i + \beta_3 x3_i$$

  - That's called an "**inner product**" or "**dot product**"

The Linear Predictor ...

- An OLS regression

$$y_i = \beta_0 + \beta_1 x1_i + \beta_2 x2_i + \beta_3 x3_i + \varepsilon_i \tag{3}$$

would be

$$y_i = X_i\beta + \varepsilon_i$$

- Often use the Greek η_i ("eta") for the linear predictor, so

$$y_i = \eta_i + \varepsilon_i$$

- The expected value of y_i in ordinary regression equals the linear predictor

$$\begin{aligned} E[y_i|X_i] &= E[\eta_i + \varepsilon_i] \\ &= E[\eta_i] + E[\varepsilon_i] \\ &= \eta_i + 0 = X_i\beta \end{aligned}$$

GLM

- McCullagh & Nelder (1989) worked out a framework called generalized linear models (GLM)
- The Logit & Probit models are examples
- There are 2 twists on the usual regression that we need to watch for.

Generalized Linear Model, step 1

- Rewrite OLS in this way.
- Regression

$$y_i = X_i\beta + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma_\varepsilon^2)$$

- Because error is Normal, and $E[\varepsilon_i] = 0$, the conditional distribution of y_i is normal with mean η_i and variance σ_ε^2:

$$y_i \sim N(\eta_i, \sigma_\varepsilon^2)$$

Generalized Linear Model, step 1 ...

- **Generalization step 1**: Put other distributions in place of the Normal.
  - $y_i \sim SomeOtherDistribution(\eta_i)$ {may have more parameters}
- In the categorical model, we will use a Binomial distribution, a random draw from {0,1}.
- Clarify Binomial versus Bernoulli
  - Binomial answers "how many positives out of X draws with probability p_i", $Bin(X, p_i)$
  - Bernoulli is "take 1 draw with probability p_i", give back 0 or 1
  - Binomial with $X = 1$ is same as Bernoulli.

Generalized Linear Model, step 2

- Transform η_i before putting it into the probability distribution.
- We'll demonstrate it with the "S-shaped" curve below
- Choice of transform is often for practical/computational reasons, but we wish there were substantive reasons.
  - example, suppose a probability distribution requires positive input for a mean parameter.
  - Because $X_i\beta$ can be negative, we are often told to transform that as $\exp(X_i\beta)$
  - Exponentiating creates positive values, but can we tell a meaningful story to justify that decision?

KU

Outline

KU

Outline ...

- Small Sample with Separation

8. Testing Statistical Significance

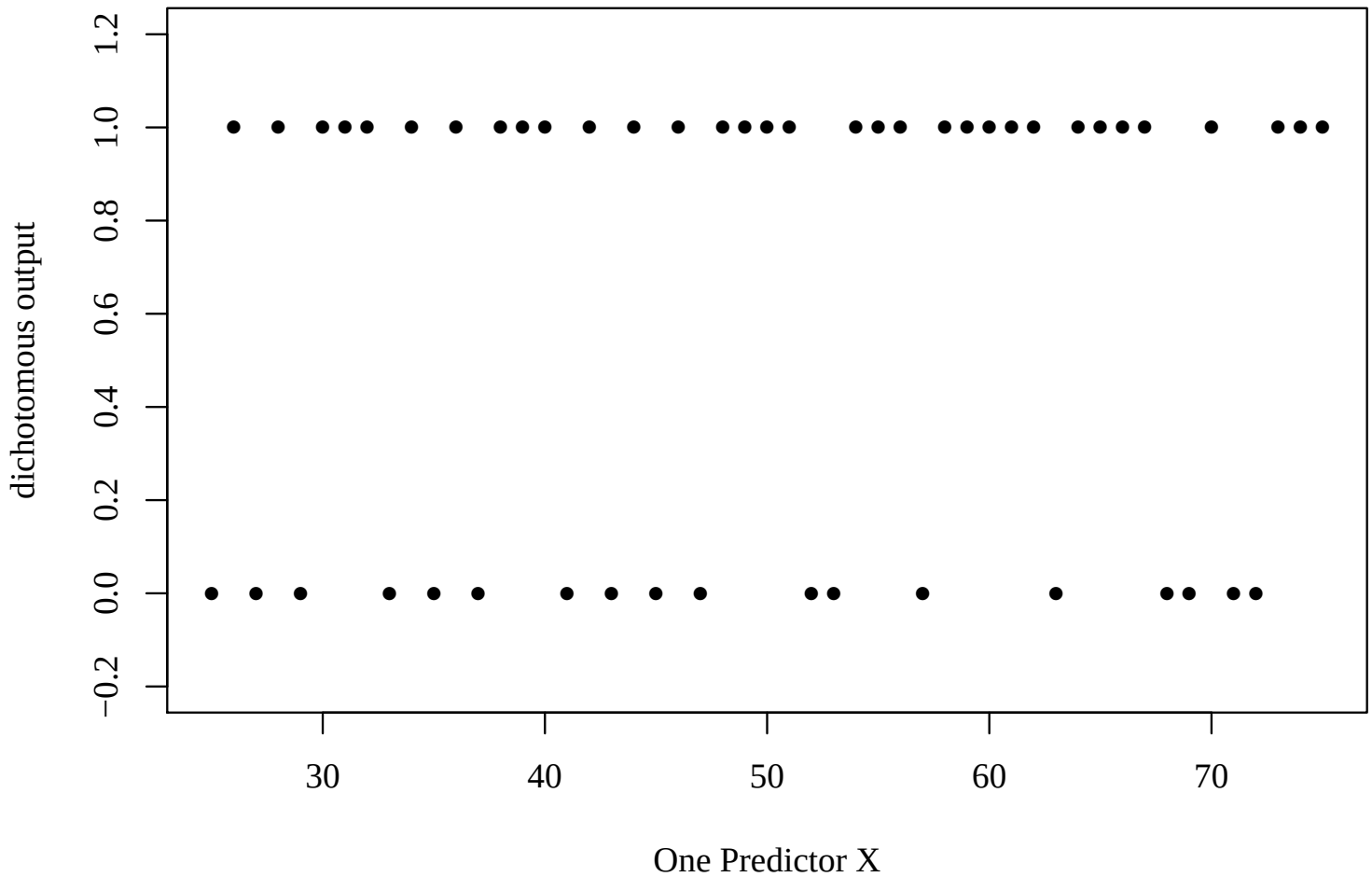
9. Model Goodness
   - Percent Correctly Predicted and ROC
   - LLR equivalent of an F test
   - Deviance
   - Why no R square
   - Hosmer and Lemeshow test

OLS versus Logit

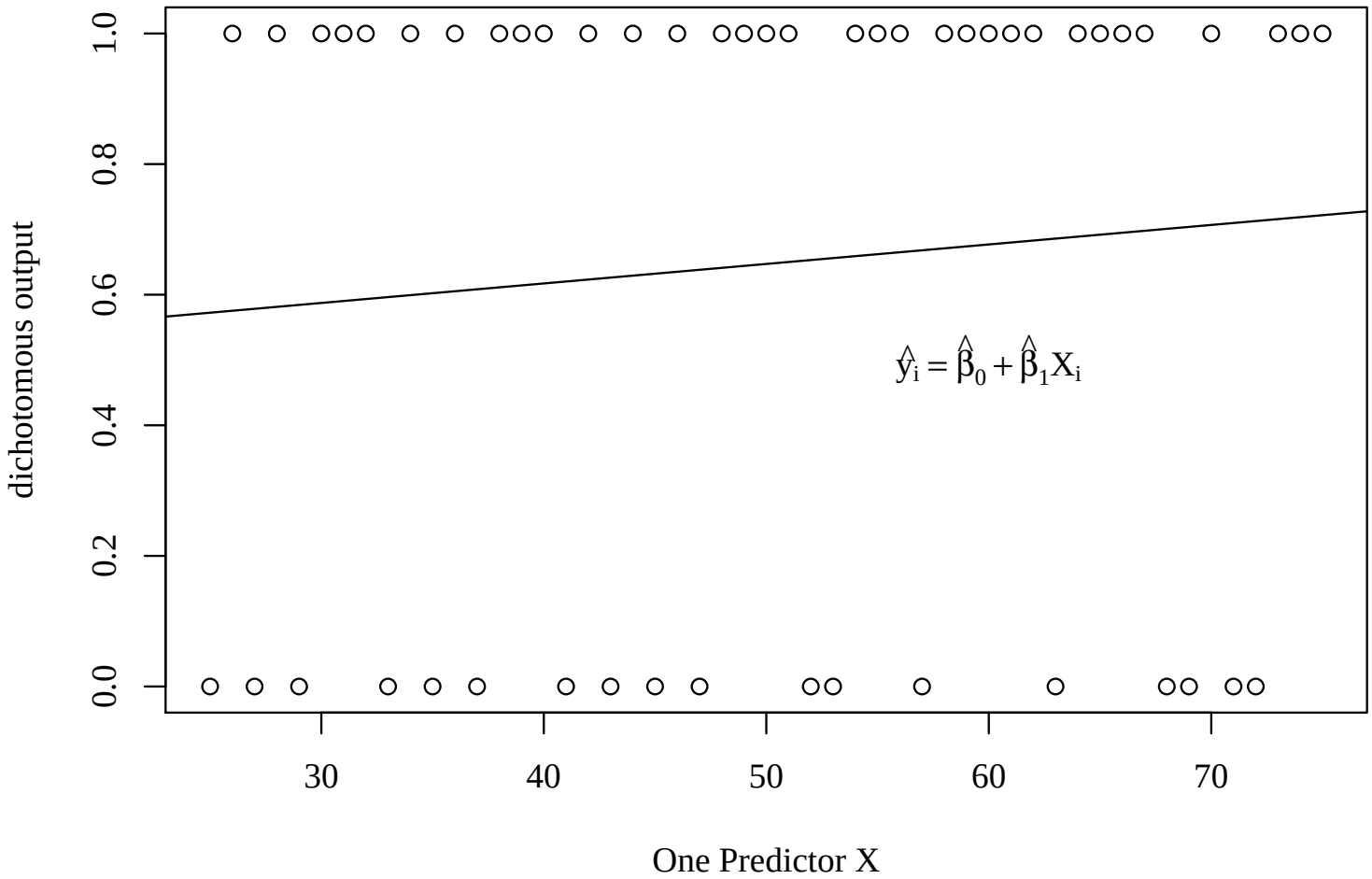
- Ordinary Least Squares is popular
  - researchers frequently treated the $\{0,1\}$ data as numbers
- Logit (or other "categorical regression models") grew rapidly in popularity in the late 1970s.
- The "big data" ("data mining") field has put new life into the linear model, mainly because it is easy to calculate and most often results are similar to other, more elaborate models

y_i is dichotomous

Suppose y_i is coded 0 and 1, representing answers to a Yes or No question.

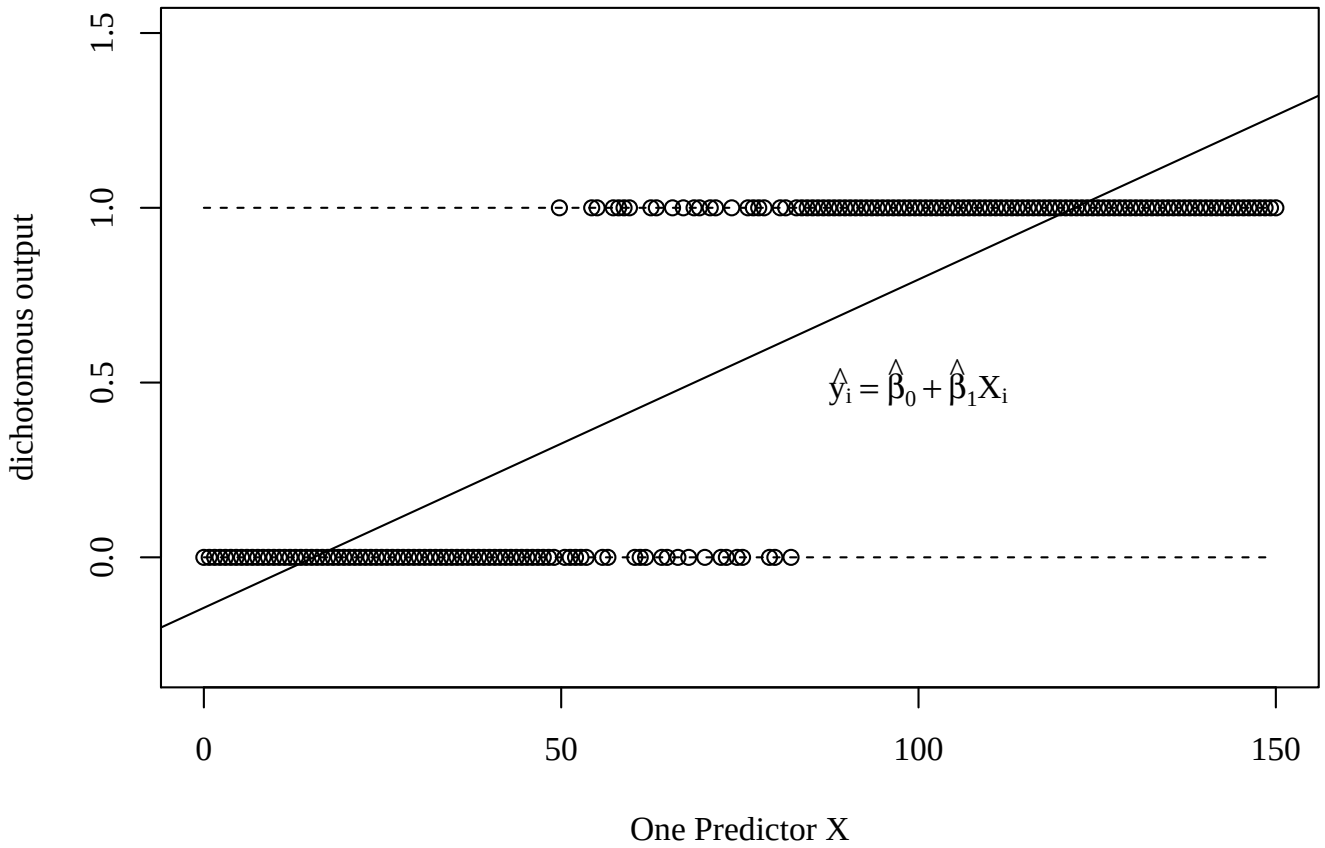


Fit OLS: interpret the line



Problem: OLS predicts out of range

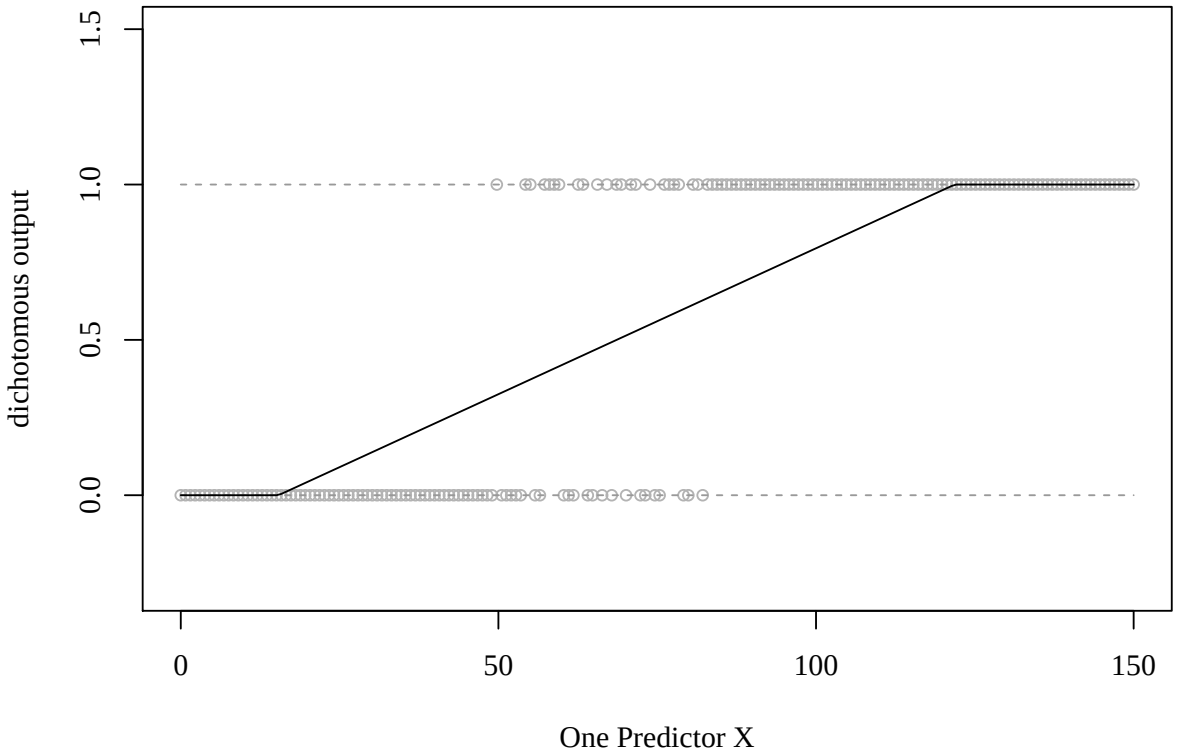
A straight line will eventually go above 1 and below 0.



That means $\hat{y}$ can't represent probabilities? (I'm asking, not telling)

Might try 'truncation' of the predicted values

To prevent the OLS model from going out of bounds, one approach is to insert "kinks" in the fitted line.



But I've never seen anybody carry through on that in a serious way.

Might try 'truncation' of the predicted values ...

Shortcomings

1. Theoretically unappealing. Effect of X is 0, then β_1, then 0 again? Really?
2. Difficult to interpret $\hat{y}_i = 0$. Something is actually impossible?
3. $\hat{y}_i = 1$? Something is certain to happen. Suppose observed $y_i = 0$. Does that mean whole model is "impossible"?
4. How to estimate the "kink" points coherently?

Maybe data is such that predictions will stay in bounds. Whew. Lucky!

Linear Probability Model

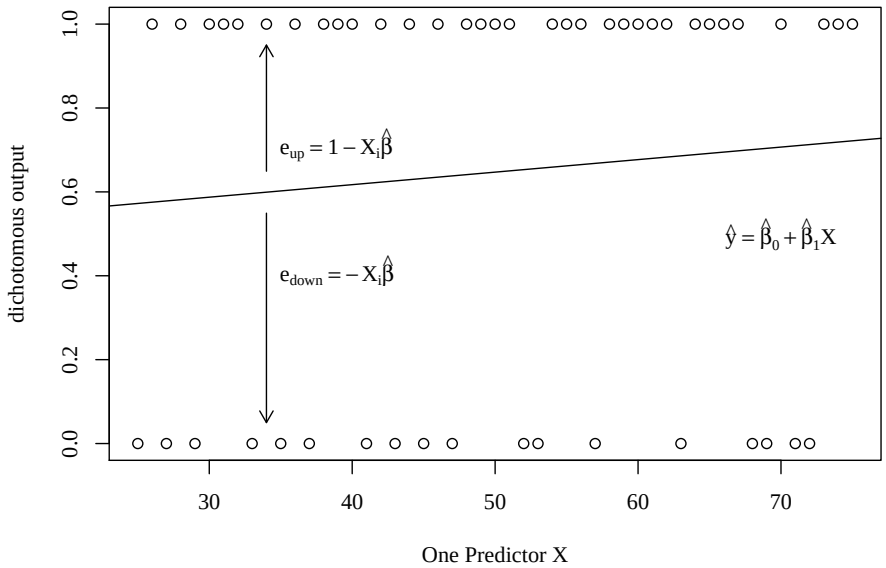
- Linear probability model says

$$y_i = X_i\beta + \varepsilon_i \tag{4}$$

- The predicted value $\hat{y}_i = X_i\hat{\beta}$ is interpreted as a probability estimate
- Problem: The error term ε_i can't be normally distributed. (Explain next slide)

Error term must be one of 2 values for any particular observation

- error term is defined as difference between
  - $X_i\beta$, and
  - the observed value, 1 or 0.
- ε_i must be either
  - $1 - X_i\beta$ if $y_i = 1$, or
  - $-X_i\beta$ if $y_i = 0$



- Repeat, error term can have only 2 values. It goes Up $(1 - X_i\beta)$ or Down $(-X_i\beta)$.
- That can't be Normal!

You insist that $E[\varepsilon_i] = 0$, then find $Var(\varepsilon)$

- Let P_i be the probability of a 1 for case i.
  - Then probability that $y_i = 0$ is $(1 - P_i)$.
- The expected value of error term:

$$\begin{aligned} E[\varepsilon_i] &= P_i(1 - P_i) + (1 - P_i)(-P_i) \\ &= P_i - P_i^2 - P_i + P_i^2 = 0 \end{aligned} \tag{5}$$

- Recall, $E[]$ is sum of probabilities times outcomes.
  - The probabilities are P_i and $(1 - P_i)$.
  - The outcomes are $(1 - P_i)$ and $(-P_i)$

That Implies Heteroskedasticity

- Heteroskedasticity is not fatal, we'd need to do "weighted least squares" to address it. But only would bother if I could convince myself he "mean model" is reasonable
- The variance of the error term is

$$\begin{aligned} Var(\varepsilon_i) &= E([\varepsilon_i - E[\varepsilon_i])^2] \\ &= E[\varepsilon_i^2] \ \{\text{because } \mathrm{E}[\varepsilon_\mathrm{i}] = 0\} \\ &= P_i(1 - P_i)^2 + (1 - P_i)(-P_i)^2 \\ &= P_i(1 - P_i) \end{aligned} \tag{6}$$

- The error variance is biggest when $P_i = 0.5$, smaller (tending toward 0) as $P_i \rightarrow 0$ or $P_i \rightarrow 1$.
- Weighted Least Squares might be used, parameter estimates will have lower variance and the standard errors more accurate.
  - Assumes we believe the linear model for probabilities
  - We ignore the boundary problem.

KU

Outline

Outline ...

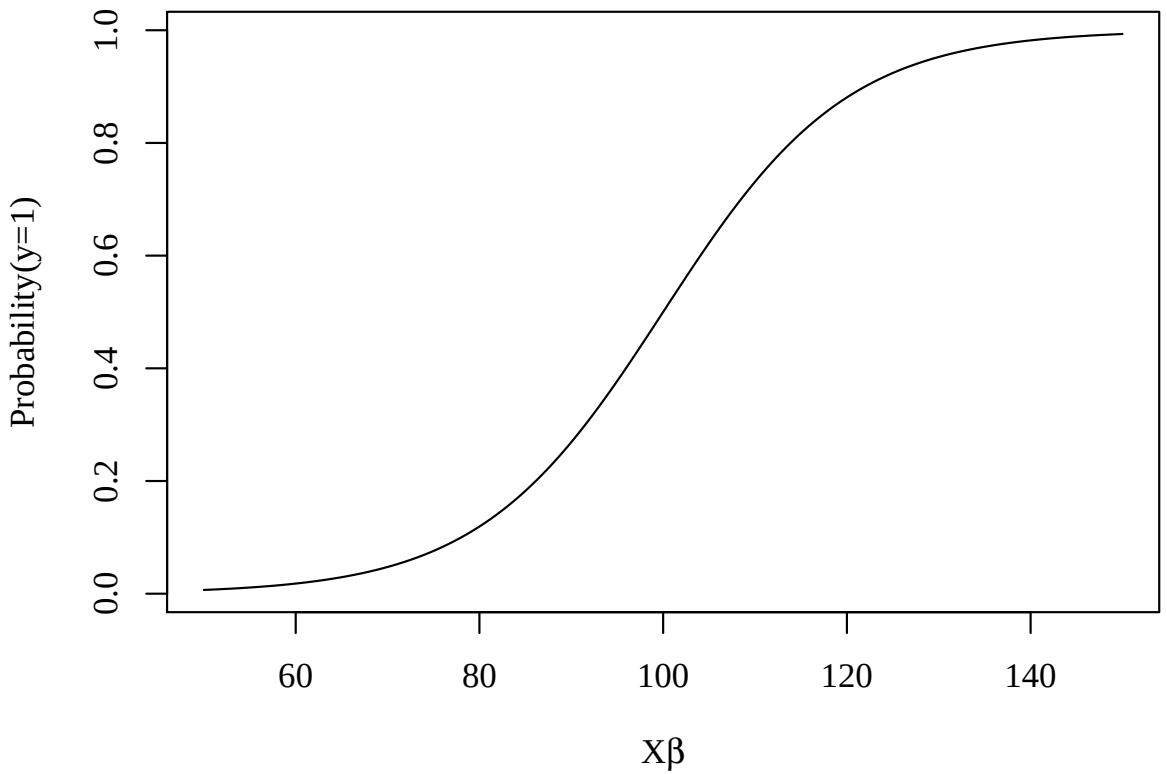
- Small Sample with Separation

8. Testing Statistical Significance

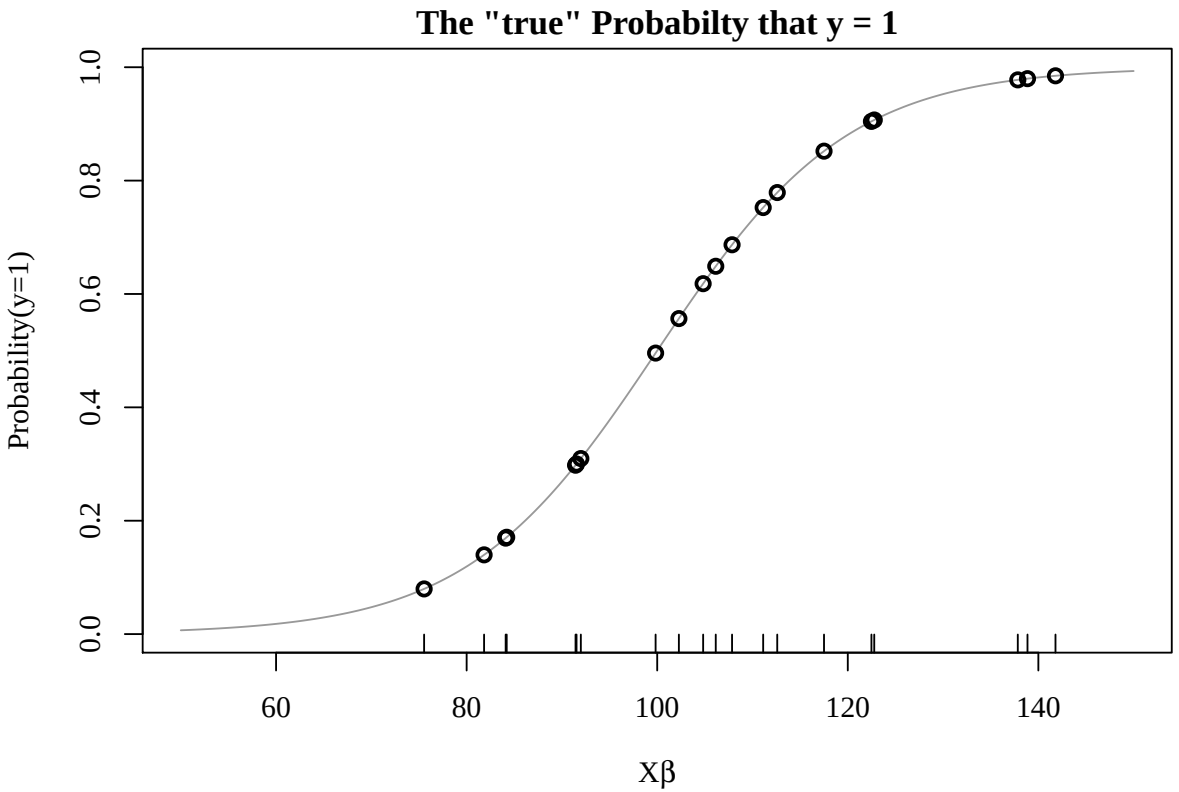
9. Model Goodness
   - Percent Correctly Predicted and ROC
   - LLR equivalent of an F test
   - Deviance
   - Why no R square
   - Hosmer and Lemeshow test

Smooth Curve for Probability

- Theory: probability that $y_i = 1$ changes "smoothly" in response to changes in $X_i\beta$.
- Monotonic relationship implies an S-shaped curve.

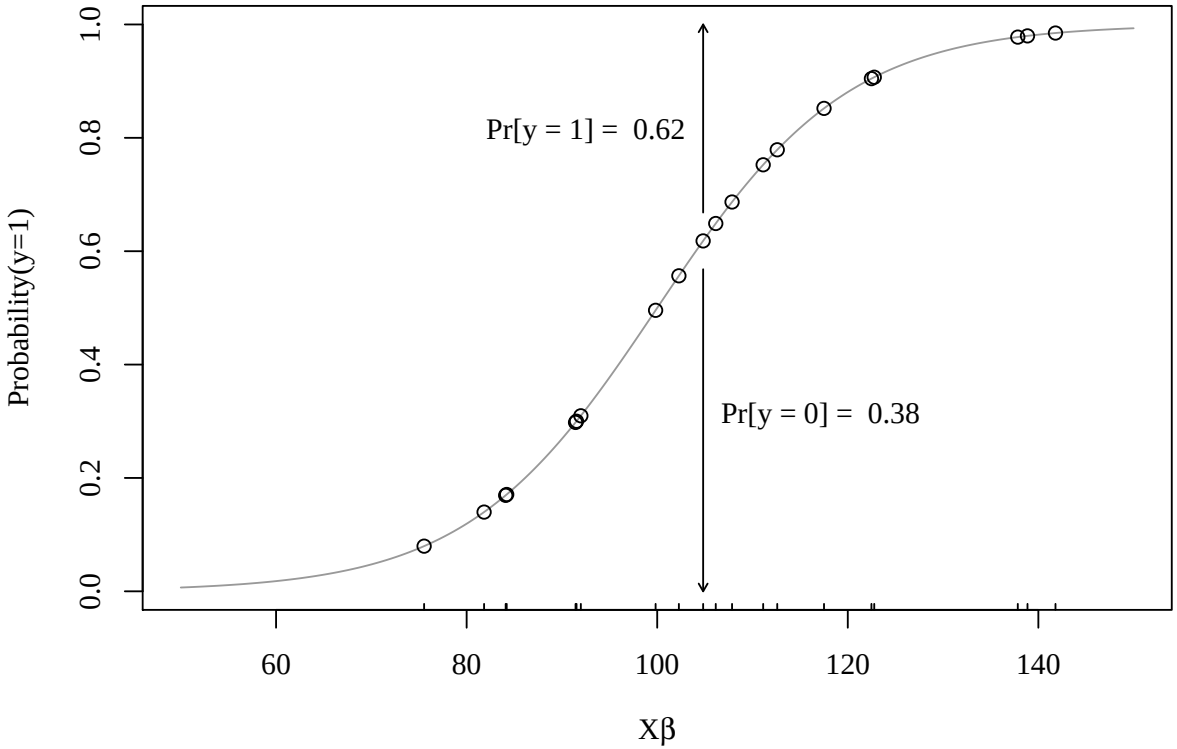


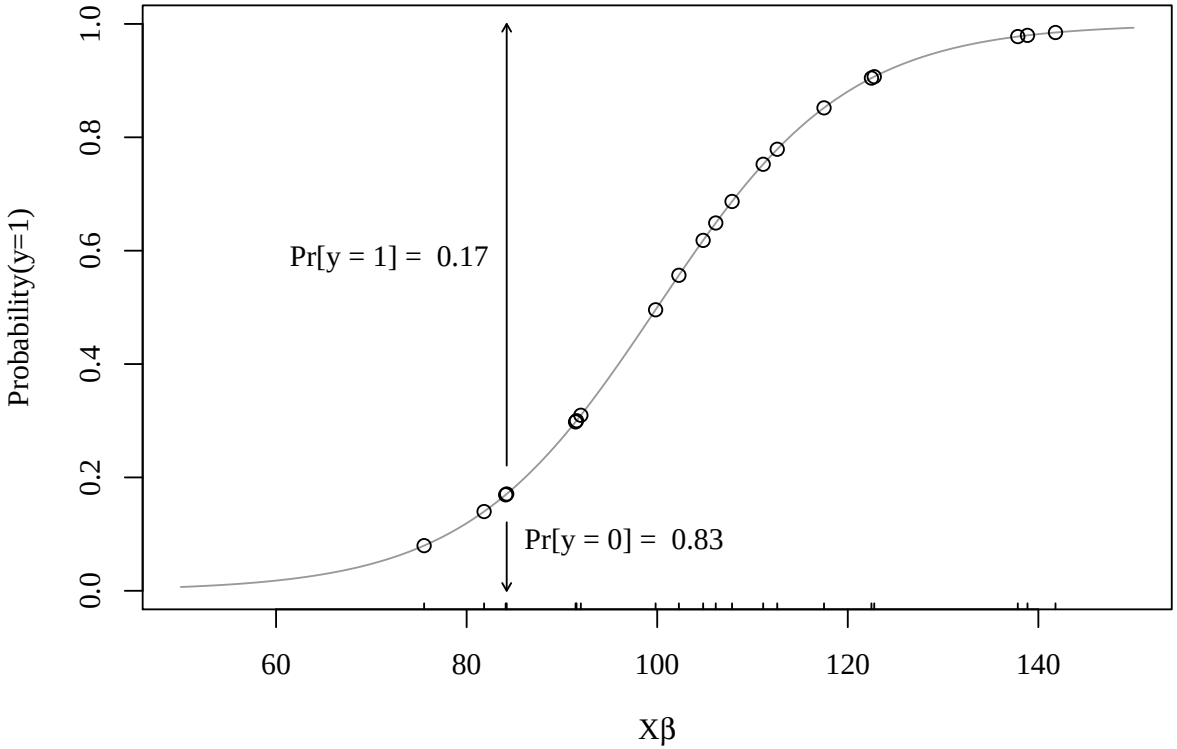
The True Probability for each value of $X_i\beta$

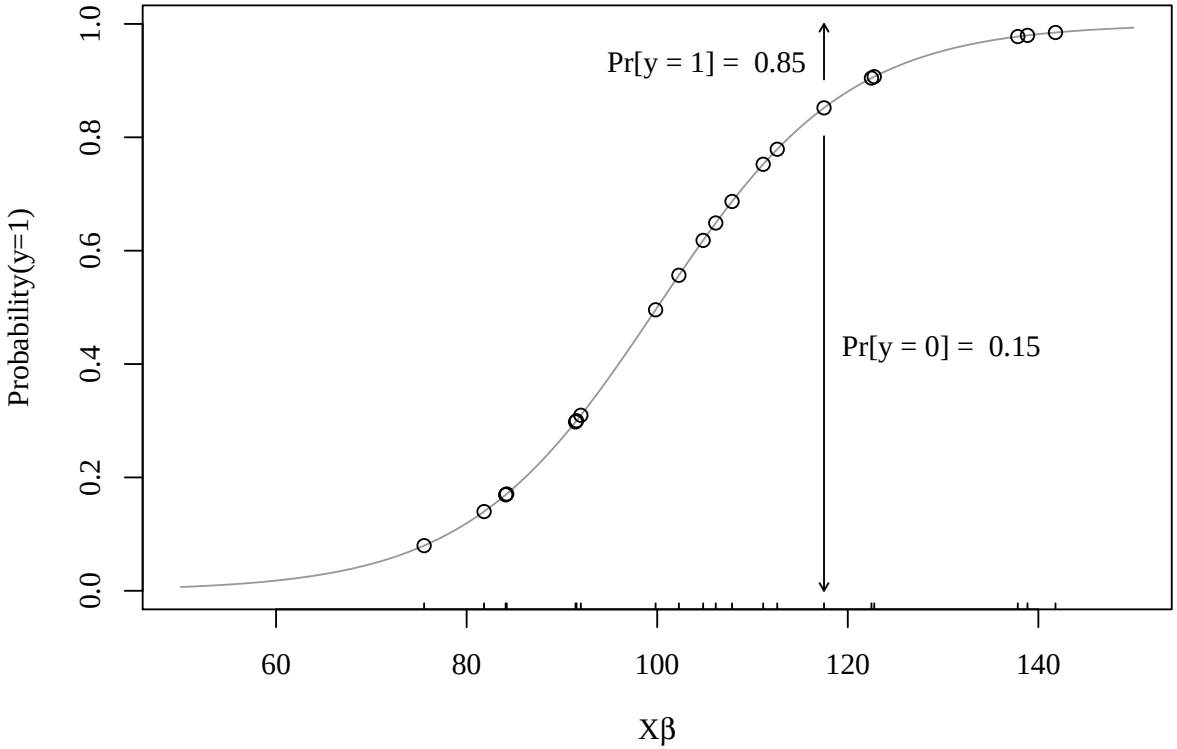


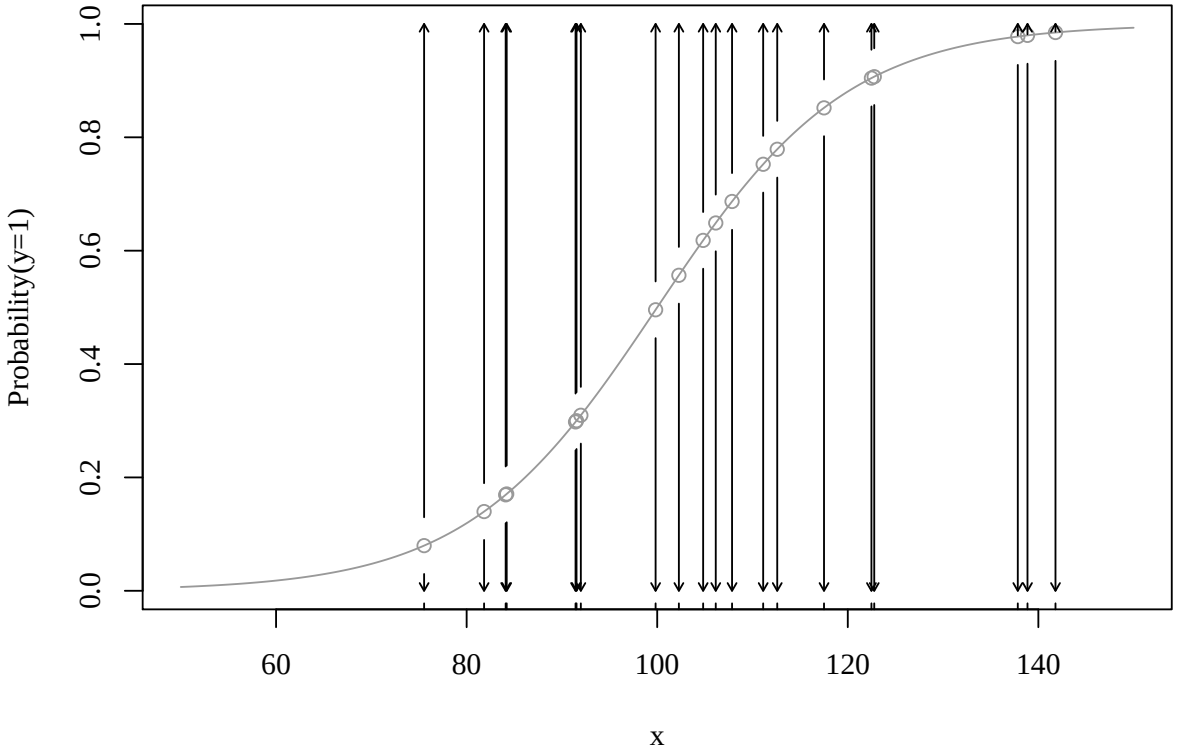
The "true probabilities" are not observed. We see 0 or 1 for each case.

Observed 0's and 1's: Downs and Ups




Probability(y=1)
0.0
0.2
0.4
0.6
0.8
1.0
Pr[y = 1] = 0.17
Pr[y = 0] = 0.83
60
80
100
120
140
Xβ

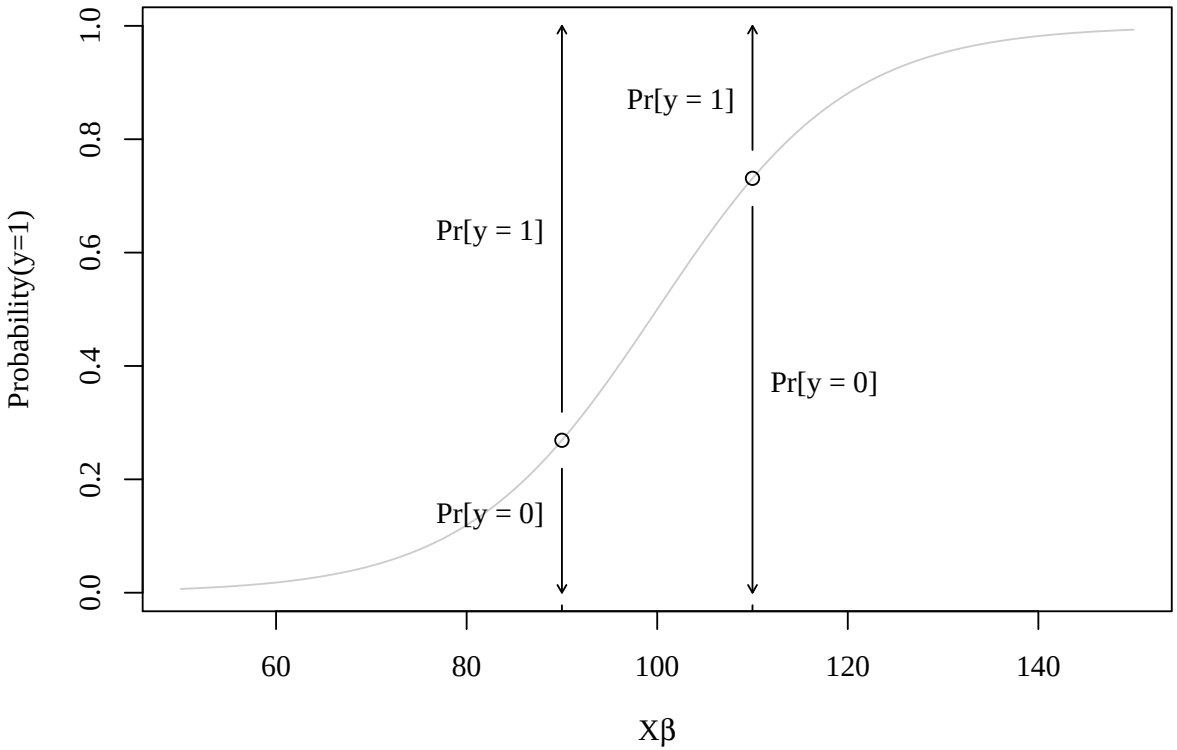

Probability(y=1)
0.0
0.2
0.4
0.6
0.8
1.0
Pr[y = 1] = 0.85
Pr[y = 0] = 0.15
60
80
100
120
140
Xβ


Probability(y=1)
0.0
0.2
0.4
0.6
0.8
1.0
60
80
100
120
140
x

Ways to think about that

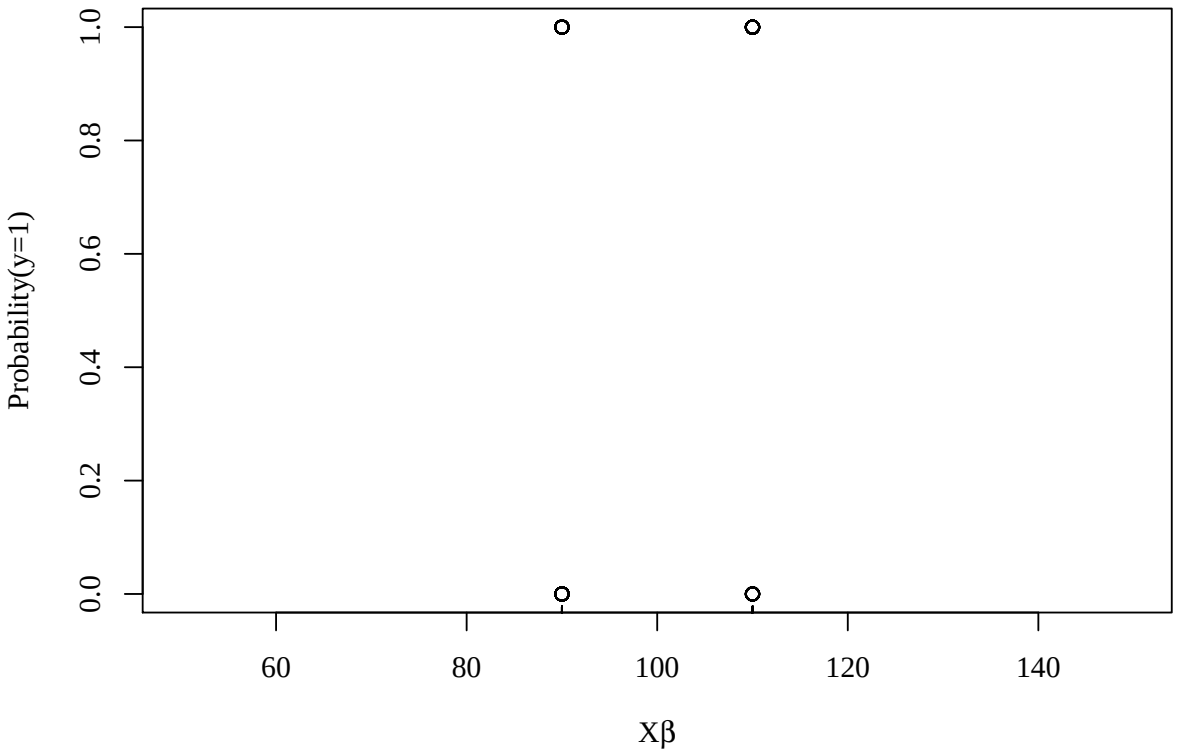
- The linear predictor is transformed to the S-shaped curve
  - Then an "up or down" random draw assigns 0 or 1
  - The "up or down" draw is said to be drawn from a Binomial probability distribution
- The probability values (expected values of y_i) are link-transformed back to values of $X_i\beta$
- Any "S-shaped" curve would work. Many reasonable suggestions exist.

If X is a dichotomy



We still think as though there is an underlying S-curve, but outcomes only observed at a few points

What does a scatterplot look like if X is a dichotomy



Fitting the coefficients for an S-shaped curve seems somewhat "heroic", but that's what we do.

Logistic S-Shape Formula

- One formula for an S-shaped curve, the Logistic Model

$$Prob(y_i = 1|X_i) = \frac{exp(X_i\beta)}{1 + exp(X_i\beta)} \quad (7)$$

$$= \frac{1}{1 + exp(-X_i\beta)} \quad (8)$$

$$\{\text{same as}\} = \frac{1}{1 + e^{-(X_i\beta)}} \quad (9)$$

- Remember that
  - $exp(X_i\beta) \rightarrow \infty$ (rapidly) as $X_i\beta$ grows.
  - $exp(-X_i\beta) \rightarrow 0$ as $X_i\beta$ grows.

Logit Transform

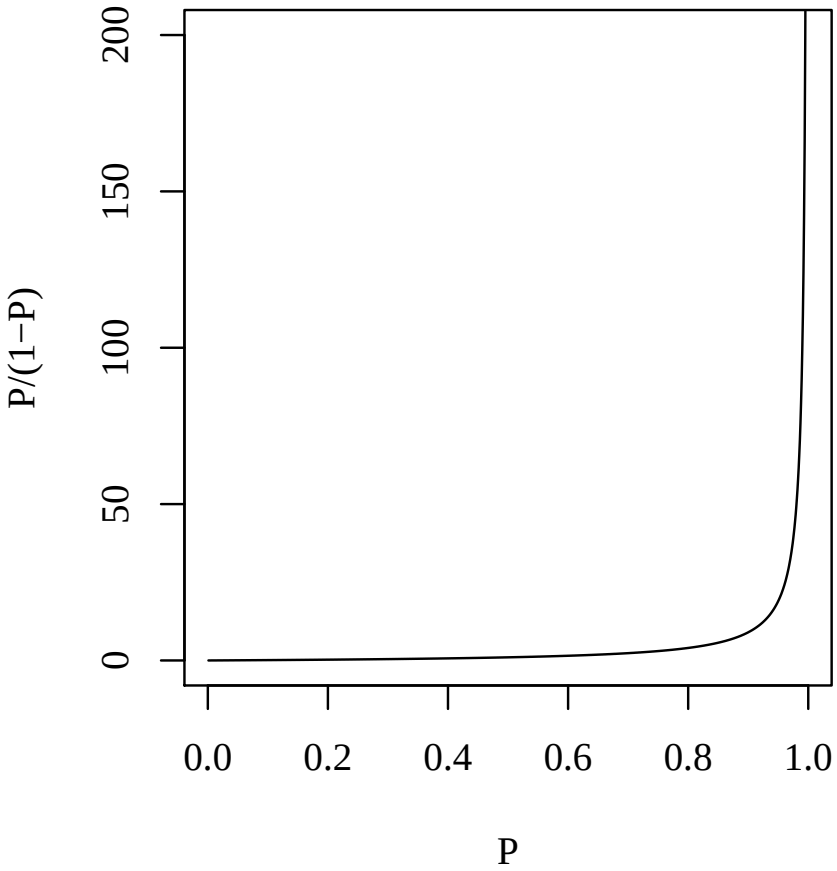
Let $P_i = Prob(y_i = 1|X_i)$.

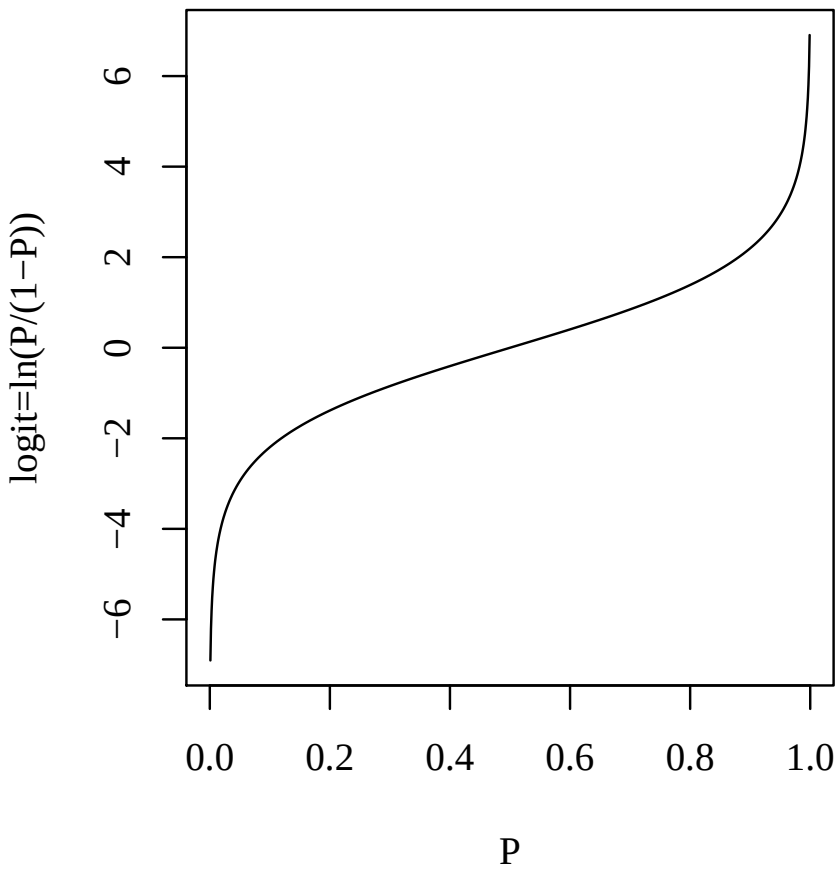
We can re-arrange equation (9) to produce a new formula that has the linear predictor on the right side by itself.

$$\ln\left[\frac{P_i}{1-P_i}\right] = x_i\beta \tag{10}$$

- This is the "logit" transform of P_i. logit = "log of the odds ratio".
- Can somebody interpret the odds for me?
- if $P_i = 0$ or 1, the logit is undefined. The theory does not allow a "sure thing"

Logit Transform ...





Logistic Coefficient Interpretation

Let P_i be the chance $y = 1$

Let x_i be one predictor

1. The slope for one predictor, $\frac{\partial Pi}{\partial x_i} = \beta_1 \cdot P_i \cdot (1 - P_i)$.
   - Logistic population growth: as P_i nears the upper limit of 1, its rate of growth is slower and slower.
   - The effect of 1 unit change in x_i is β_1 weighted by $P_i(1 - P_i)$.
   - At $P_i = 0.5$, the slope is at a maximum. There $P_i(1 - P_i) = 0.25$.
     - $\frac{1}{4}\beta_1$ the slope of the "S curve" at the mid point.
   - If y_i is very likely to be a 1 or a 0, a change in X_i doesn't make much difference.
2. Obtain predicted probabilities, make nice table or plot
3. Explore estimates in the linear predictor space. Because

$$\ln\left[\frac{P_i}{1 - P_i}\right] = x_i\beta$$

then people who can think of $ln[\frac{P}{1-P}]$ as a meaningful value can use $\hat{\beta}$ as a prediction of it.

The odds ratio: One way of "scaling" logit coefficients

- Some like to interpret the odds of the outcomes.
  - $\frac{P}{1-P}$ (graphed that on slide 51)
- Going further, some consider the ratio of the odds for 2 sets of predictors. The OR (Odds Ratio) is widely used.
- Suppose a predictor x_i has only 2 values, 0 and 1. i.e., a "dummy variable"

The odds ratio: One way of "scaling" logit coefficients ...

- Calculate the difference in the log odds ratio.
- Let
  - P_1 be the probability that $y_i = 1$ if $x_i = 1$ and
  - P_0 be the probability $y_i = 1$ if $x_i = 0$
- The difference in the log odds:

$$ln\left[\frac{P_1}{1-P_1}\right] - ln\left[\frac{P_0}{1-P_0}\right] = \{\beta_0 + \beta_1 \cdot 1\} - \{\beta_0 + \beta_1 \cdot 0\}$$

$$ln\left[\frac{\frac{P_1}{1-P_1}}{\frac{P_0}{1-P_0}}\right] = \beta_1$$

The odds ratio: One way of "scaling" logit coefficients ...

- Now apply $\exp$ to both sides get the odds ratio by itself on the left side

$$\frac{\frac{P_1}{1-P_1}}{\frac{P_0}{1-P_0}} = exp(\beta_1)$$

- On the left, we have the ratio of the log odds. Hence,
  - $exp(\beta_1)$ is the "odds ratio".
- Another Interpretation: odds in case 1 are proportional to odds in case 0

$$\frac{P_1}{1-P_1} = \frac{P_0}{1-P_0} exp(\beta_1)$$

- If x_i is a "dummy variable", then this approach may be meaningful. When x_i has a different range, I don't see any value in it.

The Weaknesses in the odds ratio

- I'm not sure who first proposed the OR, but there are many publications that do use it.
- There are also quite a few articles that discourage it (Cummings (2009); Lee et al. (2009); Sainani (2011). The gist of this is as follows
- Many authors actually want to study the "relative risk"

$$RelativeRisk = RR = \frac{P_1}{P_0}$$

- However, from the logistic regression output, we cannot get RR. Additional calculation would be needed.
- Decades ago, it was noticed that if P_1 and P_0 are very large or very small, then the Odds Ratio is a reasonable, quick, "back of the envelope" guess about RR.
- Nevertheless, over time, the roots were forgotten and authors are urged to report OR, mistakenly believing that they are interpreted as relative risk.

KU

Outline

Outline ...

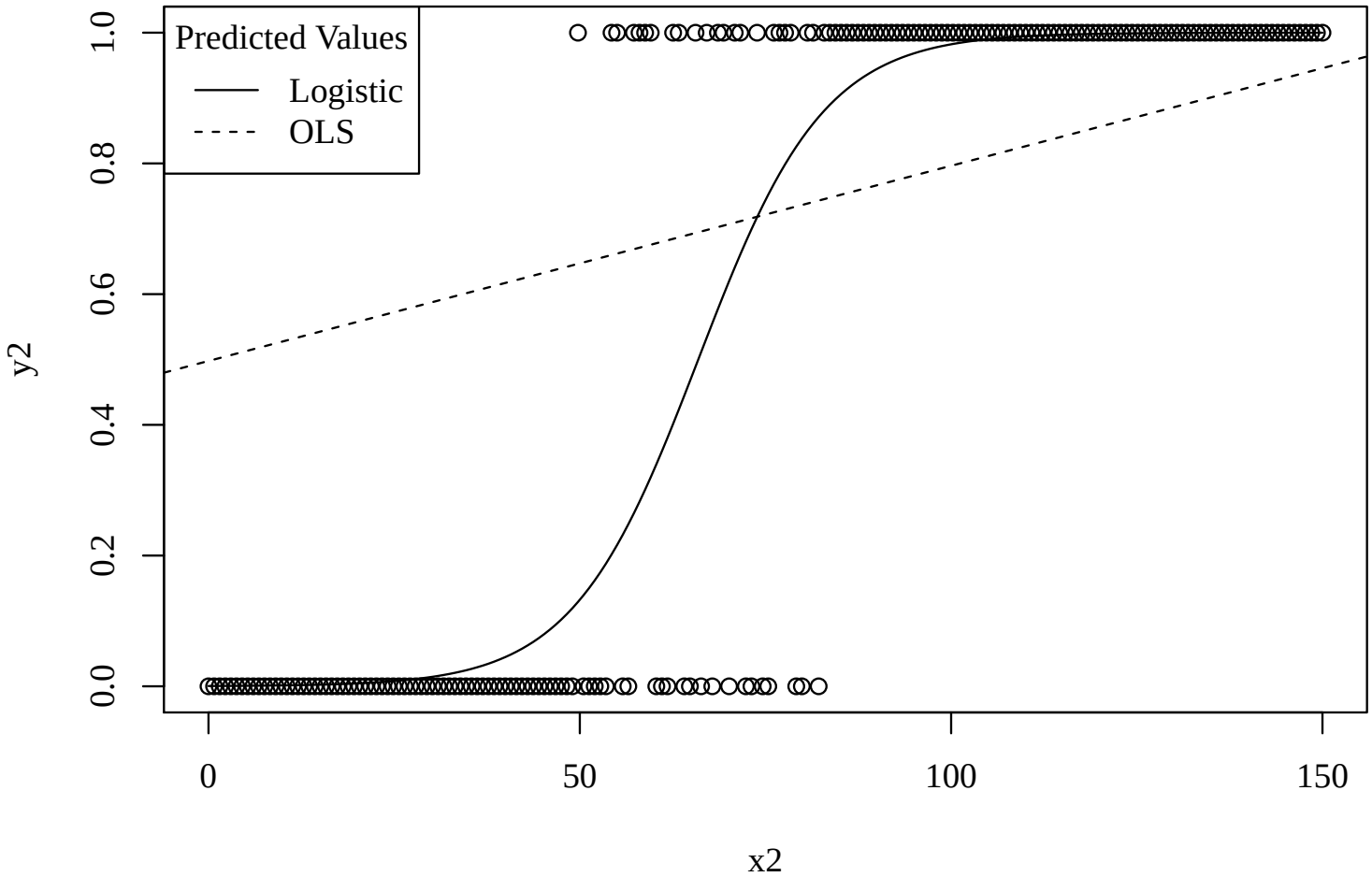
- Small Sample with Separation

8. Testing Statistical Significance

9. Model Goodness
   - Percent Correctly Predicted and ROC
   - LLR equivalent of an F test
   - Deviance
   - Why no R square
   - Hosmer and Lemeshow test

The BIG PICTURE: Overlay OLS and Logit

My test data has outcome "y2" and predictor "x2". Was graphed above.



Logistic fitted with glm in R

- In R, it is fitted as a generalized linear model.
  - Family is binomial
  - link is logit

```
glm1 <- glm(y2~x2, family=binomial(logit))
summary(glm1)
```

```
Call:
glm(formula = y2 ~ x2, family = binomial(logit))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.02397  -0.15142   0.01666   0.15136   2.02366

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.77205    1.31902  -5.892 3.81e-09 ***
x2           0.11784    0.01947   6.051 1.44e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

Logistic fitted with glm in R ...

```
    Null deviance: 274.372  on 199  degrees of freedom
Residual deviance:  73.971  on 198  degrees of freedom
AIC: 77.971

Number of Fisher Scoring iterations: 7
```

Features worth noting

- Familiar "parameter summary table" with columns for $\hat{\beta}$ and $std.err.(\hat{\beta})$
- 3rd column is not t, but rather z.
- There's no R^2
- New statistics "Null deviance", "Residual deviance" and AIC.

rockchalk outreg table: Compare OLS and Logistic

	OLS Estimate (S.E.)	Logit Estimate (S.E.)
(Intercept)	-0.144***	-7.772***
	(0.040)	(1.319)
x2	0.009***	0.118***
	(0.000)	(0.019)
N	200	200
RMSE	0.283	
R^2	0.678	
F(df_{num},df_{denom})	417(1,198)***	
Deviance	73.971	
$-2LLR(Model\chi^2)$		200.401***
AIC	77.971	77.971

$*p \leq 0.05 ** p \leq 0.01 *** p \leq 0.001$

Consider the Standard Error and z stat columns

- No T test
- Technically, the ratio $\frac{\hat{\beta}-0}{std.err.(\hat{\beta})}$ is not compared against t distribution because that ratio has an unknown distribution in small samples.
- If N were infinite, the ratio $\frac{\hat{\beta}-0}{std.err.(\hat{\beta})}$ would be distributed Normally, that's why column in output is labeled z
  - That's what "asymptotically distributed as" refers to.
- Some software reports a squared ratio $\left(\frac{\hat{\beta}}{std.err.(\hat{\beta})}\right)^2$ which is labeled as a Wald χ^2 statistic.

Predicted values

- OLS: estimates the β's in $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Those are "linear probability" estimates
- Logistic: predicted values may be transformed, or not.
  - NOT: In linear predictor scale $\hat{\eta}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.
  - transform in order to get probabilities

$$\frac{e^{\hat{\eta}_i}}{1 + e^{\hat{\eta}_i}} \text{ same as } \frac{1}{1 + e^{-\hat{\eta}_i}}$$

About Predicted Values

- The "`predict`" function calculates $\hat{\eta}$ values from the fitted glm:

```
glm1 <- glm(y ~ x1 + x2 + x3, data=dat,
    family=binomial(logit))
predy1 <- predict(glm1)
```

- The default predict output is on the linear predictor scale, a number that can range from $-\infty$ to ∞.
- To shrink that back to $(0, 1)$ interval (representing probability scale of observed scores), transform back to the "response" scale.

```
predy2 <- predict(glm1, type="response")
```

- Works well with the `newdata` parameter to get predicted probabilities for particular cases

About Predicted Values ...

```
nd <- rockchalk::newdata(glm1,  <whatever you
    need>)
predy2 <- predict(glm1, newdata = nd,
    type="response")
```

- **HOWEVER**, `predict.glm` does not provide confidence intervals (long, controversial story). Various "improvised" CIs possible

rockchalk::predictOMatic gives fitted values on the response scale

```
predictOMatic(glm1, predVals = c(x2 = "seq"), n =
    10)
```

```
           x2           fit
1     0.00000 0.0004211702
2    16.66667 0.0029945053
3    33.33333 0.0209611159
4    50.00000 0.1324087151
5    66.66667 0.5210494822
6    83.33333 0.8857780288
7   100.00000 0.9822314635
8   116.66667 0.9974686620
9   133.33333 0.9996441153
10  150.00000 0.9999500592
```

Show off some more predictOMatic

HOWEVERHOWEVER

Here the values of the predictor are selected so they equal the mean, +/- standard deviation, +/- 2 standard deviations

```
predictOMatic(glm1, predVals = c(x2 =
    "std.dev."), n = 5, interval = "confidence")
```

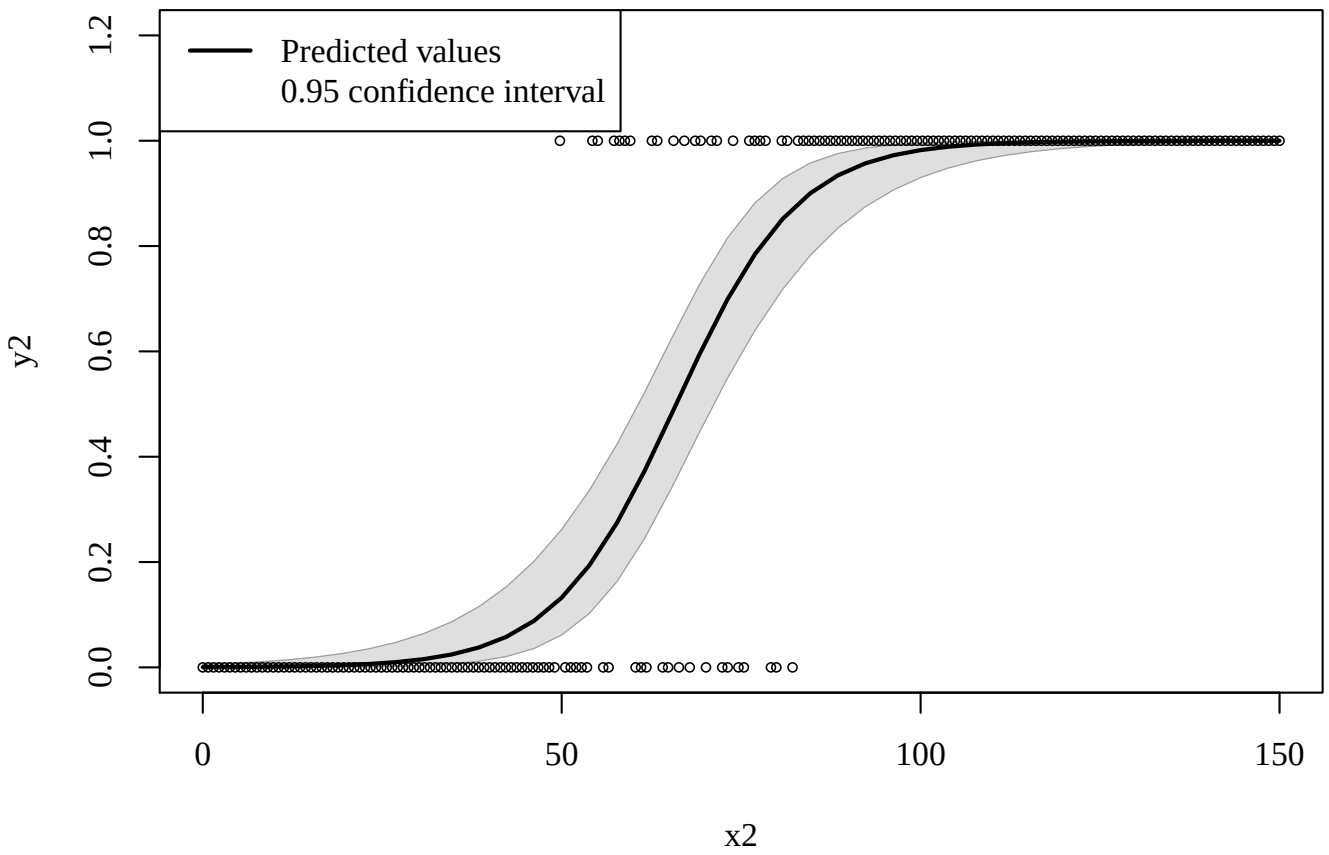
```
rockchalk:::predCI: model's predict method does not return an interval.
 We will improvize with a Wald type approximation to the confidence
     interval
      x2           fit           lwr          upr
1 -12.26 9.934368e-05 4.738676e-06 0.002078759
2  31.37 1.670386e-02 3.991158e-03 0.067178164
3  75.00 7.438898e-01 5.958178e-01 0.851257890
4 118.63 9.979905e-01 9.839348e-01 0.999751739
5 162.26 9.999882e-01 9.995137e-01 0.999999715
```

plotCurves in rockchalk

```
plotCurves(glm1, plotx="x2",
    interval="confidence")
```

```
rockchalk:::predCI: model's predict method does not return an interval.
 We will improvize with a Wald type approximation to the confidence
     interval
```

plotCurves in rockchalk ...



Outline

Outline ...

- Small Sample with Separation

8. Testing Statistical Significance

9. Model Goodness
   - Percent Correctly Predicted and ROC
   - LLR equivalent of an F test
   - Deviance
   - Why no R square
   - Hosmer and Lemeshow test

The Probability of Observing a Whole Sample

Assume the observations are statistically independent, meaning the probability of the sample equals the individual probabilities multiplied together. Hence,

$$
\begin{aligned}
Likelihood\,of\,Sample: \; L(\beta_0, \beta_1) = \\
= P(y_1 = 0, y_2 = 0, ..., y_m = 0, y_{m+1} = 1, y_{m+2} = 1, ..., y_N = 1) \\
= P(y_1 = 0)P(y_2 = 0) \cdots P(y_m = 0) \\
\times P(y_{m+1} = 1)P(y_{m+2} = 1) \cdots P(y_N = 1)
\end{aligned}
$$

Simplify That

- Remember that $P(y_i = 0) = 1 - P(y_i = 1)$, so

$$L(\beta_0, \beta_1) = (1 - P(y_1 = 1))(1 - P(y_2 = 1)) \cdots (1 - P(y_m = 1)) \times P(y_{m+1} = 1) P(y_{m+2} = 1) \cdots P(y_N = 1) \quad (11)$$

- Simplify notation: $P_i = P(y_i = 1)$

$$L(\beta_0, \beta_1) = (1 - P_1)(1 - P_2) \cdots (1 - P_m)) \times P_{m+1} P_{m+2} \cdots P_N \quad (12)$$

Log That To Simplify Further

- The Log of the Likelihood Function is a sum of logs:

$$\ln L(\beta_0, \beta_1) = \ln(1-P_1)+\ln(1-P_2)+\cdots+\ln(1-P_m)+\ln(P_{m+1})+\cdots+\ln(P_N) \tag{13}$$

- In a Logistic case, we'd fill in $P_i = \frac{1}{1+e^{-(\beta_0+\beta_1 X_i)}}$
- MLE, short for Maximum Likelihood Estimate, is the choice of estimators β_0, β_1 that maximize the log of the likelihood function. This solution is also a maximizer of L.

Quick summary of MLE properties

1. NOT unbiased. (suspicious, or even poor, small sample properties)
2. Large sample ("asymptotic") properties. MLE's are
   1. consistent,
   2. asymptotically efficient,
   3. asymptotically Normal (the central limit theorem).
3. Asymptotic standard errors. Fisher showed a way to create a variance-covariance matrix of $\hat{\beta}$. The square root of the diagonal has standard errors that appear in output.
   1. Asymptotic: if you had an infinite sample with this var-covar matrix, then we could do hypothesis tests that are accurate.
   2. Problem: In real life data, samples are not infinite, and thus we have approximate standard errors.

KU

Estimation requires iteration

- The "score equations" set the first derivatives of the log likelihood function equal to 0

$$\frac{\partial \ln L}{\partial \beta_j} = 0$$

- The numerical method for finding the β's goes roughly like this
  - make guess
  - choose a direction going "uphill" on the likelihood surface
  - repeat until guesses don't change
- The clearest description I've found is in Hastie et al. (2009, p. 121)

Bias is a realistic concern

- Because estimates are known to be biased in finite samples, there is increasing pressure to consider non-ML estimates that trim off some of the known bias
- While I don't delve into that here, the place to start is the article by Firth Firth (1993)
- Very recently, an R package to estimate the "bias reduced" logistic regression was introduced (brglm2).

Outline

KU

Outline ...

- Small Sample with Separation

8. Testing Statistical Significance

9. Model Goodness
   - Percent Correctly Predicted and ROC
   - LLR equivalent of an F test
   - Deviance
   - Why no R square
   - Hosmer and Lemeshow test

Where did the error term go?

- In OLS, we were constantly going on about ε_i and its variance. OLS says

$$y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- Logit says

$$P_i = \frac{1}{1 + exp(-(\beta_0 + \beta_1 X_i))}$$

- Where is the error term?
- Along the way, we answer "what's the difference between probit and logit?" and "what is probit, anyway?"

Restate the theory in a different way

- Write a predictive statement about a latent variable Z_i.

$$Z_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- I frequently make mistakes with signs, etc, but maybe I have this correct now

Restate the theory in a different way ...

- Theory: If Z_i is greater than 0, then $y_i = 1$.

$$\beta_0 + \beta_1 X_i + \varepsilon_i > 0$$

$$\beta_0 + \beta_1 X_i > -\varepsilon_i$$

$$-\beta_0 - \beta_1 X_i < \varepsilon_i$$

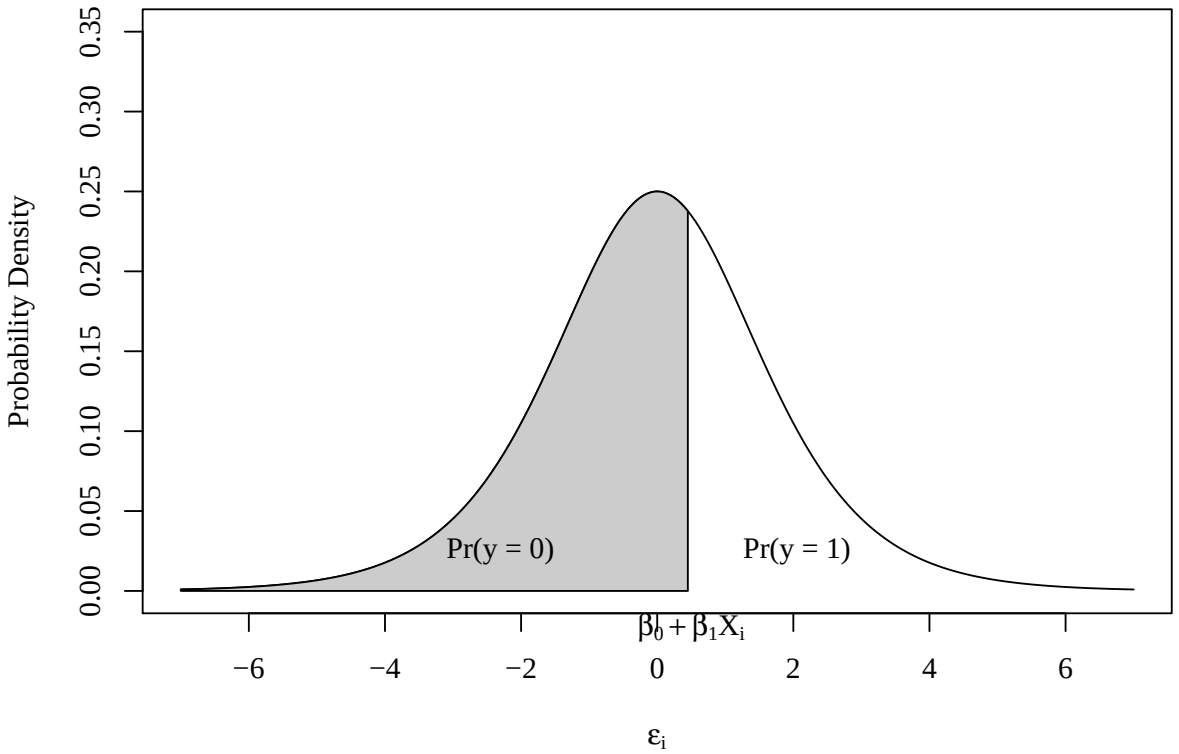
- That's the upper right side of a CDF. If $f(\varepsilon_i)$ is the PDF, then the chance of a 1 is $1 - F(-\beta_0 - \beta_1 X_i)$.
- And, as mentioned above, because $f(\varepsilon)$ is symmetric

$$F(\beta_0 + \beta_1 X_i) = 1 - F(-\beta_0 - \beta_1 X_i)$$

- Now recall my CDF discussion above. The "dividing point" in the data is $\beta_0 + \beta_1 X_i$.

KU

Restate the theory in a different way ...



Because I want to think of the chance of a 1 as a CDF, I'm graphing $y = 1$ on the LEFT side.

Logit

- The logistic probability distribution is single peaked and symmetric.
- Its formulae are especially simple and it is easy to calculate the "area under the curve" up to a point.
- We already did the CDF graph
- All you should need is proof that
  1. the $f(\varepsilon_i)$ is a formula that exists in a book somewhere, with parameters like any probability model
  2. the CDF of the logistic has the formula $exp(X_i\beta)/(1 + exp(X_i\beta))$.
- Maybe you are willing to accept those claims and we skip 2 slides.

Logistic Distribution Details (skip this slide)

- Logistic PDF for a variable ε_i:

$$f(\varepsilon_i) = \frac{exp(-(\varepsilon_i - \mu)/\sigma)}{\sigma(1 + exp(-(\varepsilon_i - \mu)/\sigma))} \tag{14}$$

- The expected value is μ.
- σ is a scale parameter, NOT the standard deviation
- The variance is $Var(\varepsilon_i) = \frac{1}{3}(\pi\sigma)^2$ and the standard deviation is $\frac{\pi\sigma}{\sqrt{3}}$
- To simplify, it is usual to assume $\mu = 0$ and $\sigma = 1$.
  - $\mu = 0$ because any non-zero amount of ε_i would show up as β_o.
  - $\sigma = 1.0$ because that "works" well enough.
- σ is an "unidentified" (unestimable) coefficient.
  - Theory says

$$y_i = 1 \; if \; \varepsilon \leq \beta_0 + \beta_1 X_i$$

Logistic Distribution Details (skip this slide) ...

- Note that the inequality is not altered if we divide both sides by any value

$$y_i = 1 \; if \; \frac{\varepsilon}{\sigma} \leq \frac{\beta_0}{\sigma} + \frac{\beta_1}{\sigma} X_i \tag{15}$$

- Because the observable result is the same, no matter how σ is set, we suppose it is 1 (and simplify our formulas)

- If we assume $\mu = 0$ and $\sigma = 1$, then the PDF simplifies to

$$f(\varepsilon_i) = \frac{e^{-\varepsilon_i}}{(1 + e^{-\varepsilon_i})^2} \tag{16}$$

- Through the magic of integral calculus, the solution is

$$P(y_i = 0 | X_i, \beta_i) = \frac{e^{\beta_0 + \beta_1 X_i}}{1 + e^{\beta_0 + \beta_1 X_i}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_i)}} \tag{17}$$

Show My Work: Derivation

- The indefinite integral is very simple.

$$
\begin{aligned}
F(\tau) &= \int_{-\infty}^{\tau} \frac{e^{-\varepsilon_i}}{(1+e^{-\varepsilon_i})^2} de_i \\
&= \frac{e^{\tau}}{1+e^{\tau}} = \frac{1}{1+e^{-\tau}}
\end{aligned}
$$

Probit

- Let ε_i be Normally distributed, $N(0, \sigma_\varepsilon^2)$.
- The Normal is a "good" distribution in many respects, especially because of
  - the central limit theorem
  - the (relatively) easy extension of the Normal to a multi-dimensional outcome variable.
- However, it is more computationally intensive.
- The PDF is

$$f(\varepsilon_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\varepsilon_i - \mu)^2}{2\sigma^2}}$$

- But the CDF is an integral for which there is no simple formula.

$$\int_{-\infty}^{\beta_0 + \beta_1 X_i} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(e_i - \mu)^2}{2\sigma^2}} de_i \tag{18}$$

This must be numerically approximated

Probit ...

- Simplify by assuming
  - $\mu = 0$.
  - $\sigma^2 = 1$. σ^2 is called the "scale parameter", it is unidentified (can't be estimated).

Probit Notation

- Because the probit equation 18 is complicated, refer to it as, Φ, representing the CDF as.

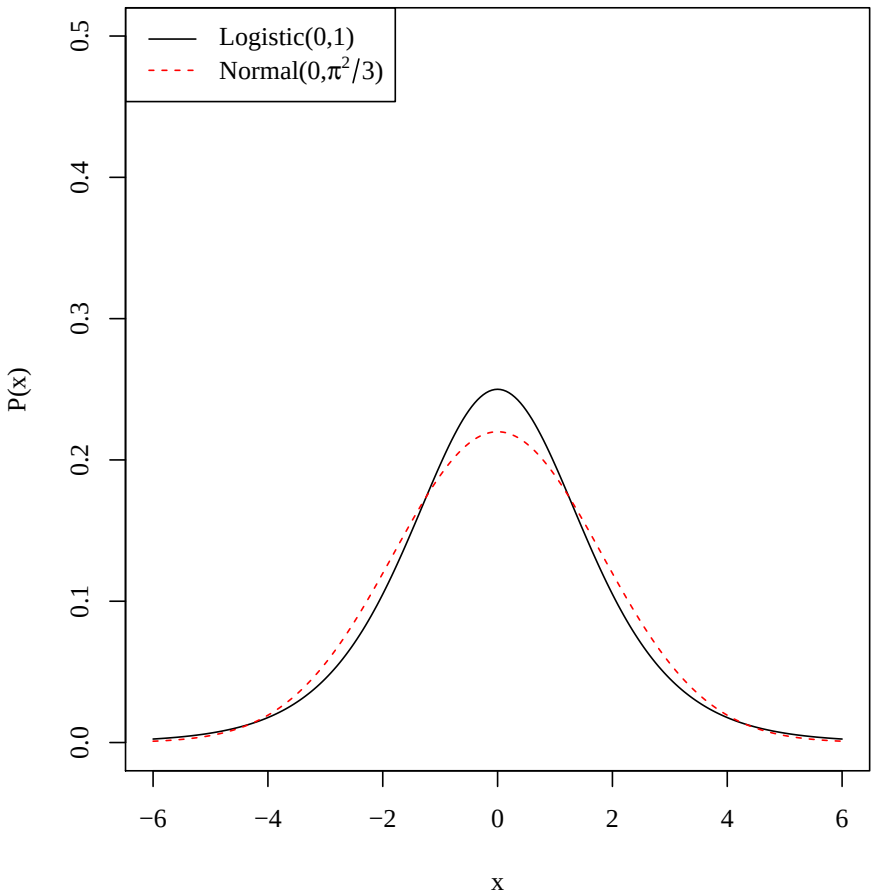
$$P(y_i = 1|X_i, \beta_i) = \Phi(\beta_0 + \beta_1 X_i) \tag{19}$$

- The probability of observing a 0 is

$$P(y_i = 0|X_i, \beta_i) = 1 - \Phi(\beta_0 + \beta_1 X_i) \tag{20}$$

Does it matter if you use Logit or Probit?

- Not very much.
- I adjust the parameter σ so that the variances of the 2 variables are the same
- If Logistic scale param$=1$, then Normal std.dev.$=\pi/\sqrt{3}$.



Fit a probit model

```
glm2 <- glm(y2~x2, family=binomial(probit))
summary(glm2)
```

```
Call:
glm(formula = y2 ~ x2, family = binomial(probit))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.00226  -0.09749   0.00051   0.09864   2.00871

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.51470    0.71031  -6.356 2.07e-10 ***
x2           0.06839    0.01045   6.544 6.00e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 274.372  on 199  degrees of freedom
Residual deviance:  72.216  on 198  degrees of freedom
AIC: 76.216

Number of Fisher Scoring iterations: 8
```

Coefficients differ

	Logit Estimate (S.E.)	Probit Estimate (S.E.)
(Intercept)	-7.772***	-4.515***
	(1.319)	(0.710)
x2	0.118***	0.068***
	(0.019)	(0.010)
N	200	200
Deviance	73.971	72.216
$-2LLR(Model\chi^2)$	200.401***	202.155***
AIC	76.216	76.216

$*p \leq 0.05** p \leq 0.01***p \leq 0.001$

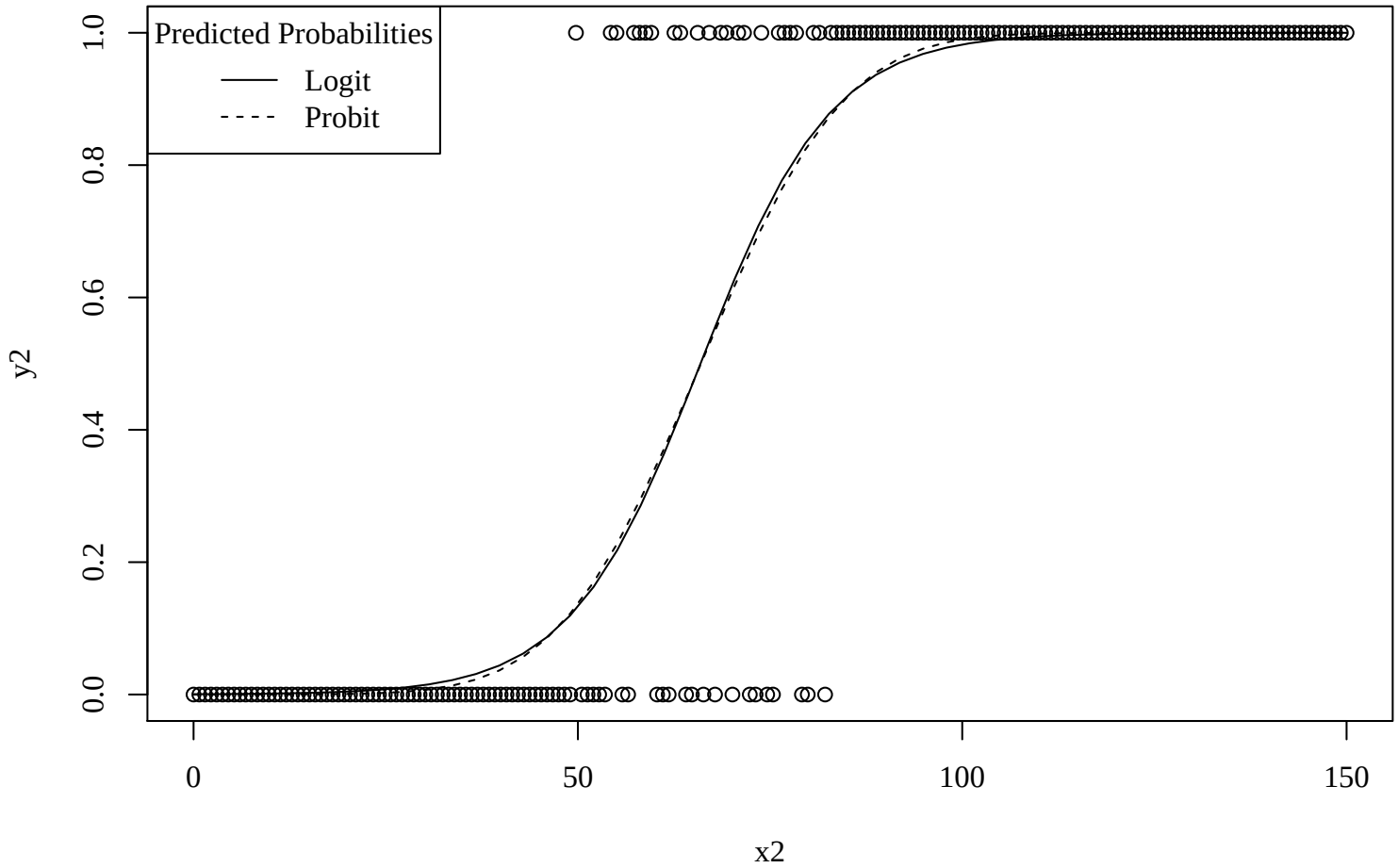
- Coefficients generally proportional to each other, approximately $\hat{\beta}_{logit} = \frac{\pi}{\sqrt{3}}\hat{\beta}_{probit}$

And you see now why people don't care if you use Logit or Probit

```
options(scipen = 10)
p1 <- predictOMatic(glm1, predVals = c(x2 =
    "seq"), n = 10)
p2 <- predictOMatic(glm2, predVals=c(x2 = "seq"),
    n = 10)
cbind(p1, p2)
```

```
          x2           fit         x2             fit
1    0.00000 0.0004211702   0.00000 0.000003170302
2   16.66667 0.0029945053  16.66667 0.000369271178
3   33.33333 0.0209611159  33.33333 0.012708304943
4   50.00000 0.1324087151  50.00000 0.136721920172
5   66.66667 0.5210494822  66.66667 0.517818373508
6   83.33333 0.8857780288  83.33333 0.881897144412
7  100.00000 0.9822314635 100.00000 0.989947138628
8  116.66667 0.9974686620 116.66667 0.999734108489
9  133.33333 0.9996441153 133.33333 0.999997928327
10 150.00000 0.9999500592 150.00000 0.999999995374
```

Compare the S-shaped curves



Summary: Comparison of Logit/Probit?

- What is the difference?
  - Logit is based on the Logistic distribution, the $\hat{\eta}_i$ is converted into probabilities by the logistic CDF
  - Probit is based on the Normal distribution, the $\hat{\eta}_i$ is converted into probabilities by the normal CDF
- In practice, Is there a big difference?
  - Not if your dependent variable is dichotomous
  - Otherwise, there can be big differences, more caution needed
- Field dependent
  - Psychologists pretty strongly prefer logistic
  - Economists pretty strongly prefer probit
  - Political scientists indifferent

KU

Outline

KU

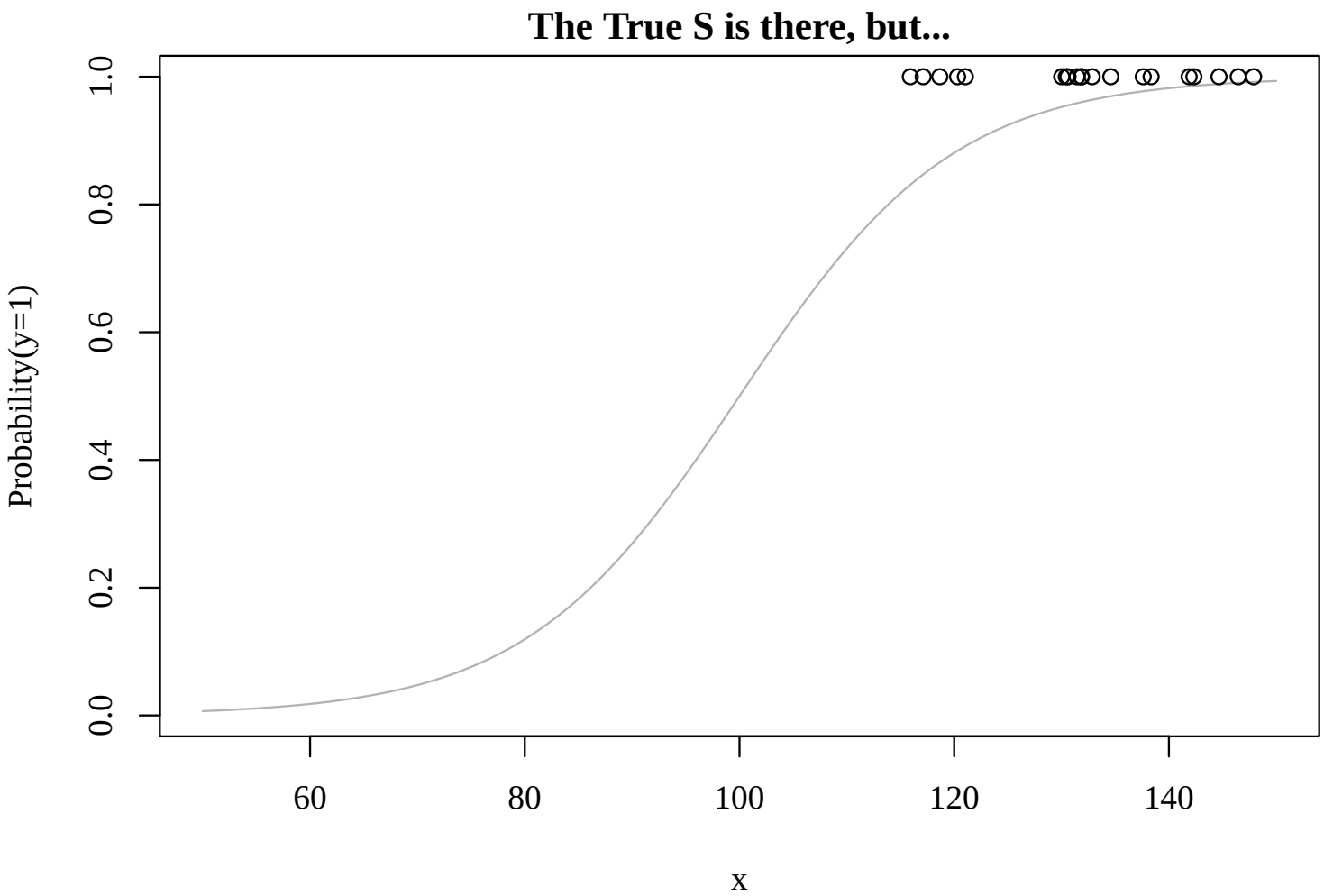
Outline ...

- Small Sample with Separation

8. Testing Statistical Significance

9. Model Goodness
   - Percent Correctly Predicted and ROC
   - LLR equivalent of an F test
   - Deviance
   - Why no R square
   - Hosmer and Lemeshow test

Can't estimate if all $y_i = 1$



Rare Events

- In most data, we don't find all y's are 0 or 1
  but we do find most y's are one or the other
- In political science, this was brought to our attention by King & Zeng (2001), who described it in a "rare events" jargon

Suggested:

- Find more cases in the minority outcome
- Correct the logit estimates, mainly the intercept needs fixing

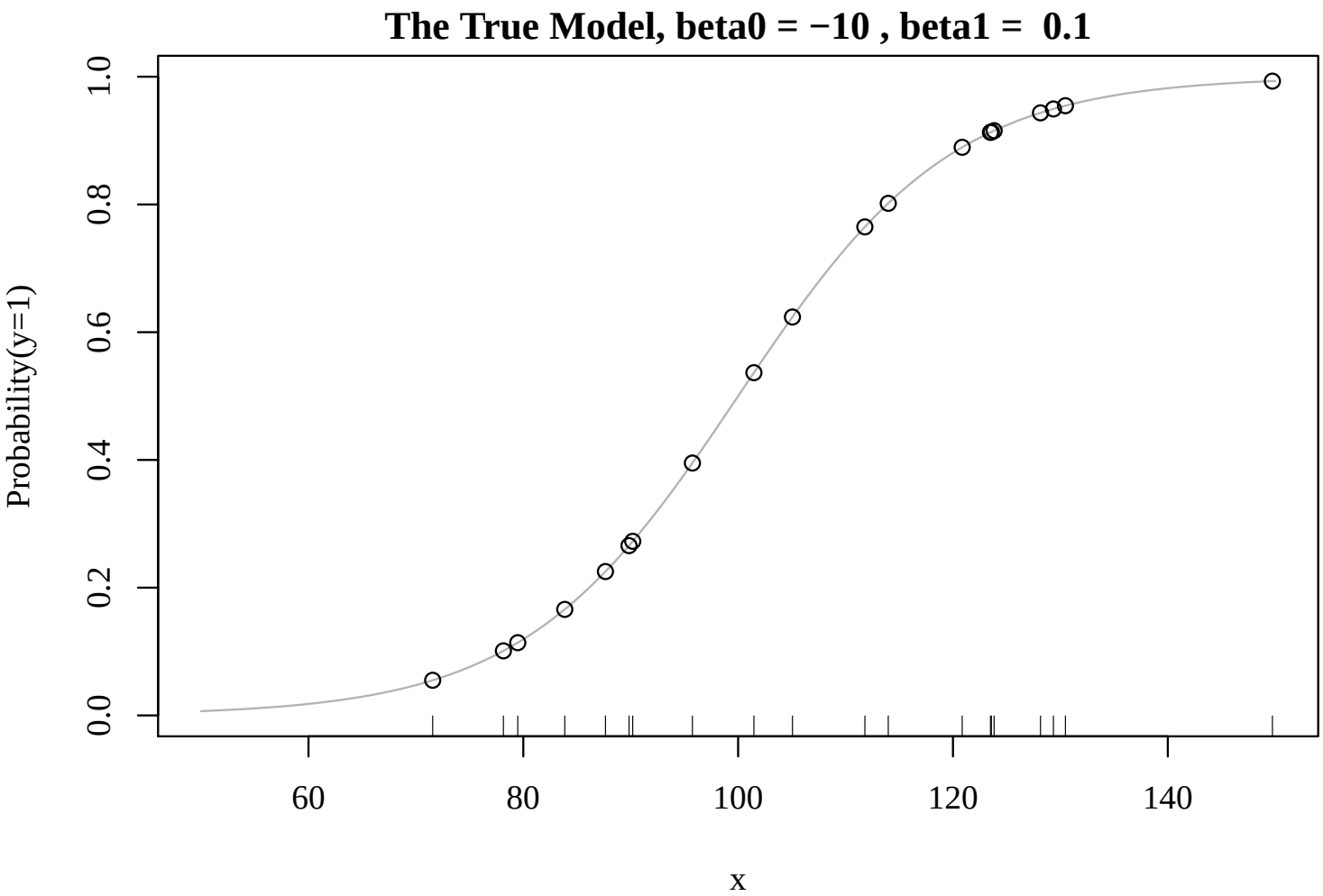
Troubles with categorical predictors

- When estimating dummy variables, estimator must calculate $P/(1-P)$ for each subgroup of observations
- If a group's outcomes are all 0 or all 1, then $P/(1-P)$ can't be calculated
- This is called "separation".
- Next I have an example of a similar problem of "separation" with a numeric predictor.

Small sample example

- Will use small samples, just 20 observations
- My "true" model will have
  - Just 1 numeric predictor, which ranges from 50 to 150
  - Logistic coefficients are $\beta_0 = -10$ and $\beta_1 = 0.1$

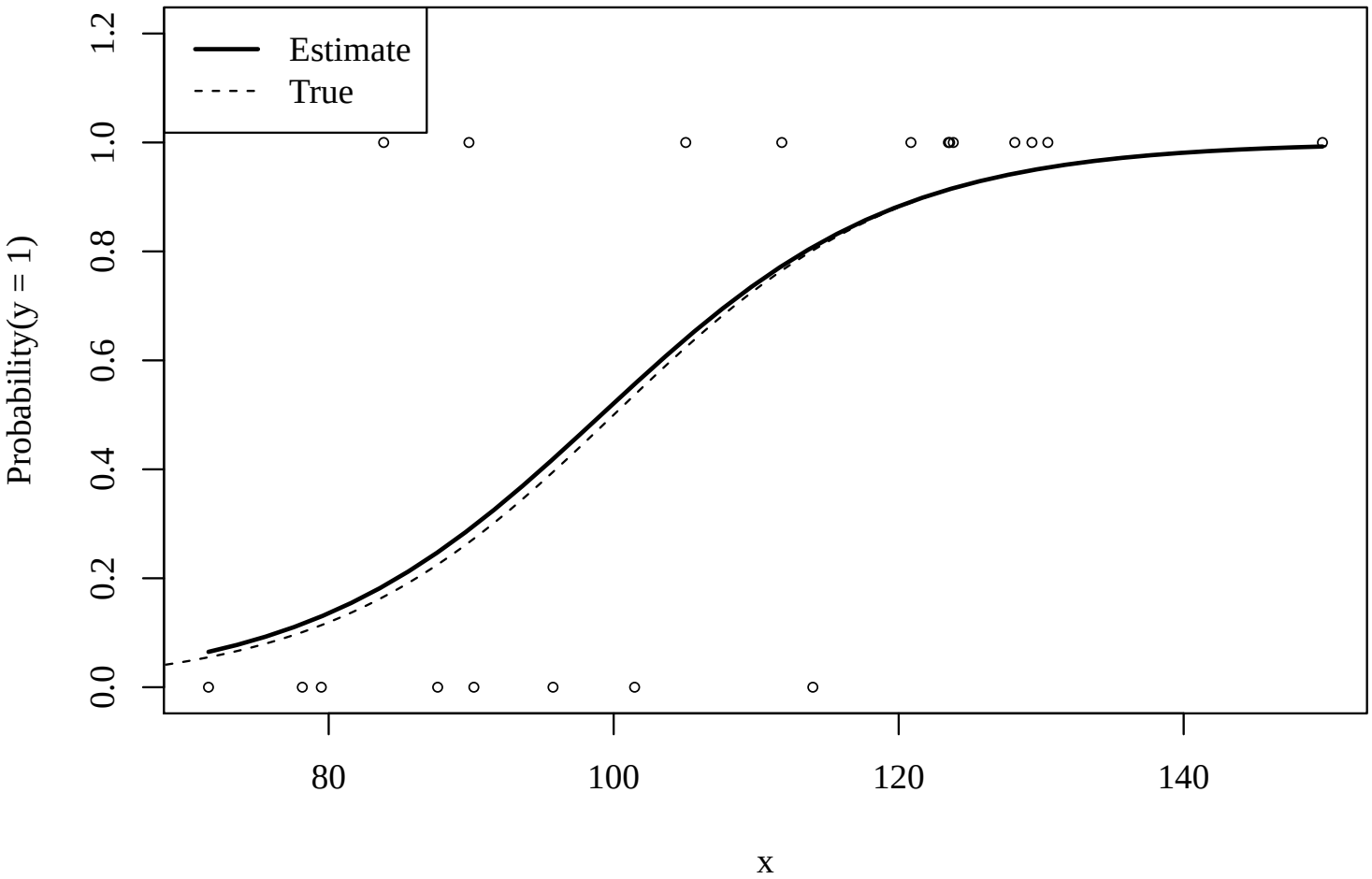
The "true" model



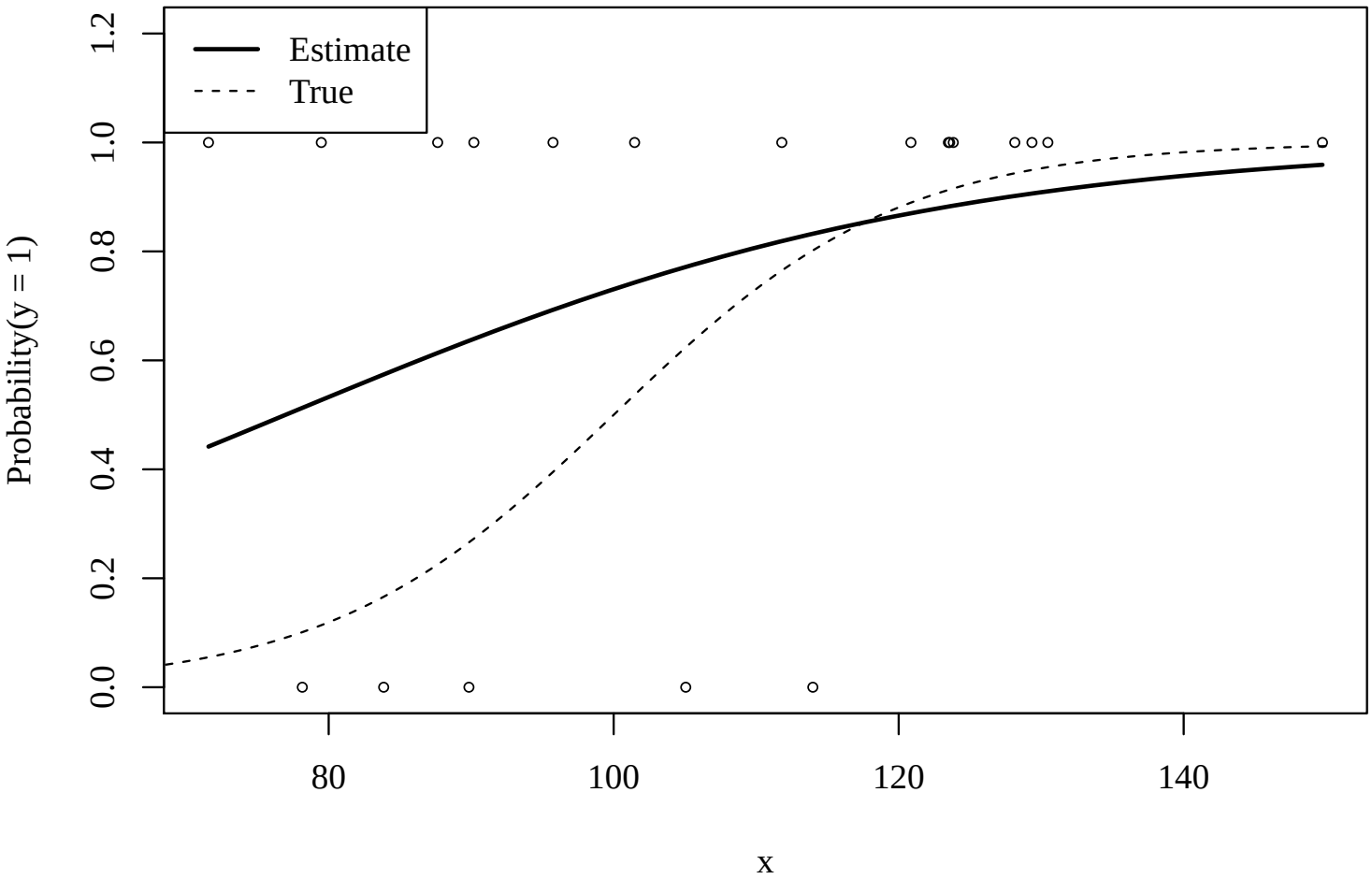
I'll draw 10 samples

- I'll keep the values of x fixed, and I'll draw fresh samples of the observed y from a binomial distribution
- I drew enough samples to make the "funny thing" happen.
- Look for sample 8

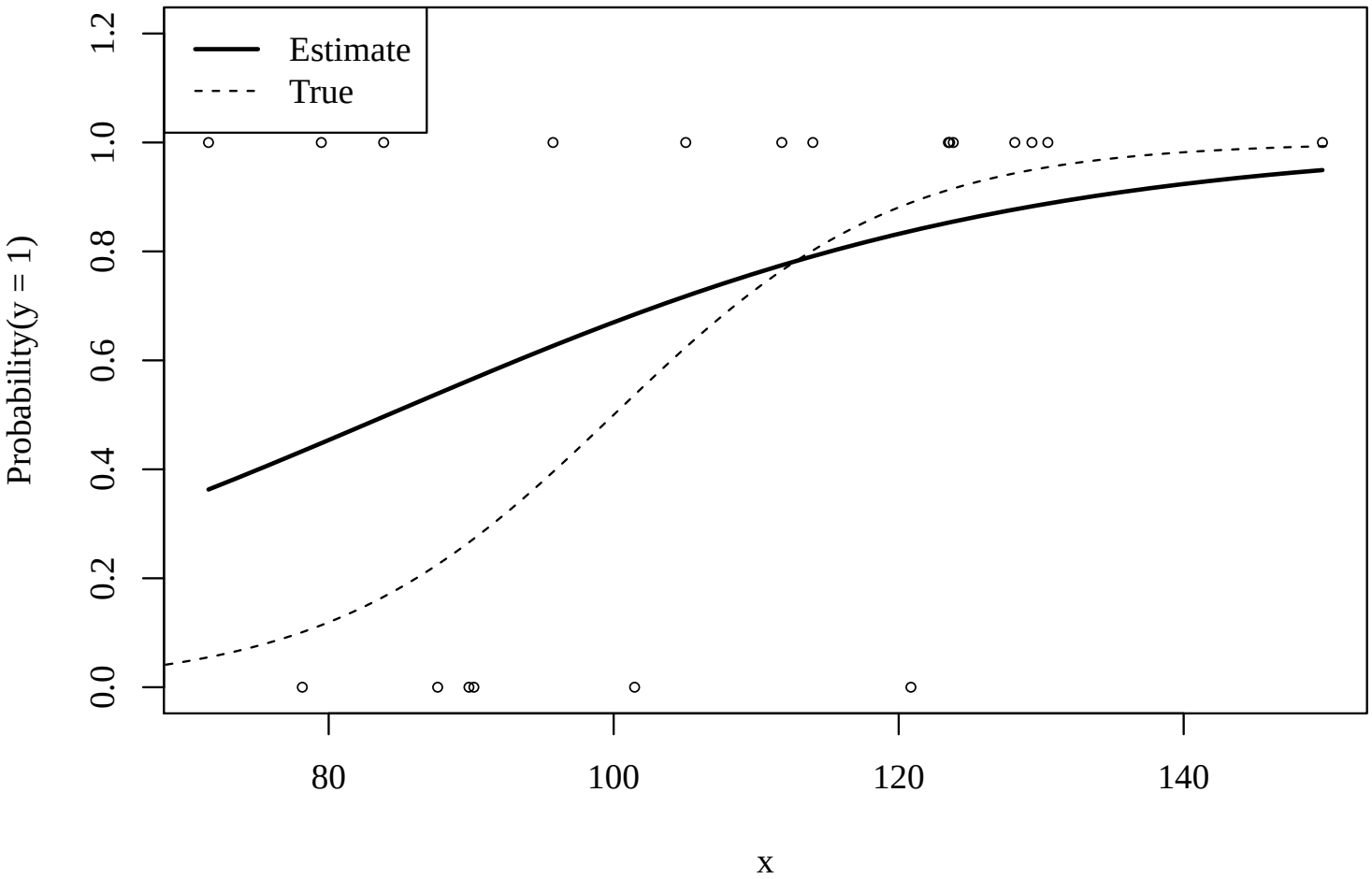
Sample 1


Estimate
True
Probability(y = 1)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
80
100
120
140
x

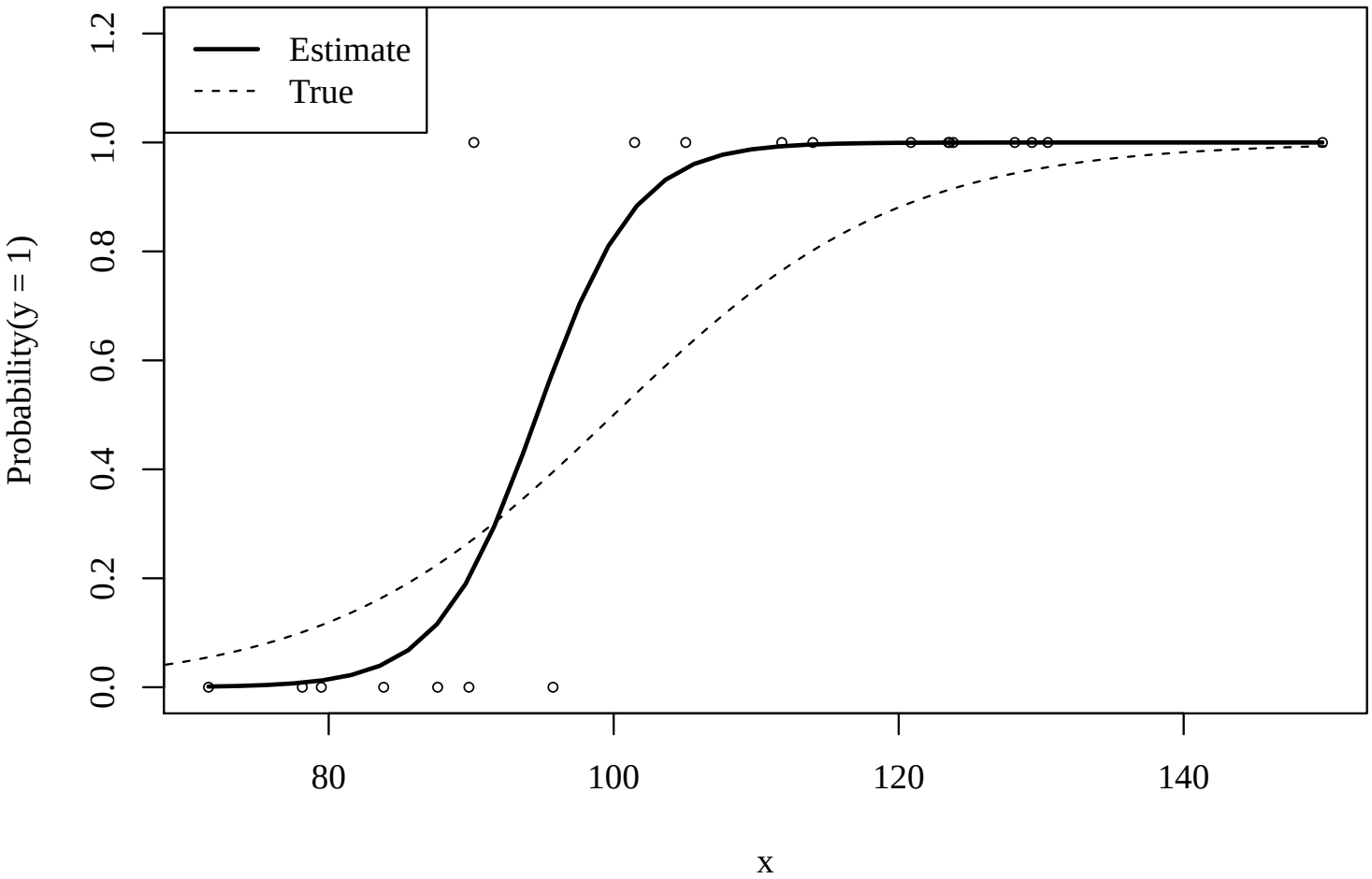
Sample 2


Estimate
True
Probability(y = 1)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
80
100
120
140
x

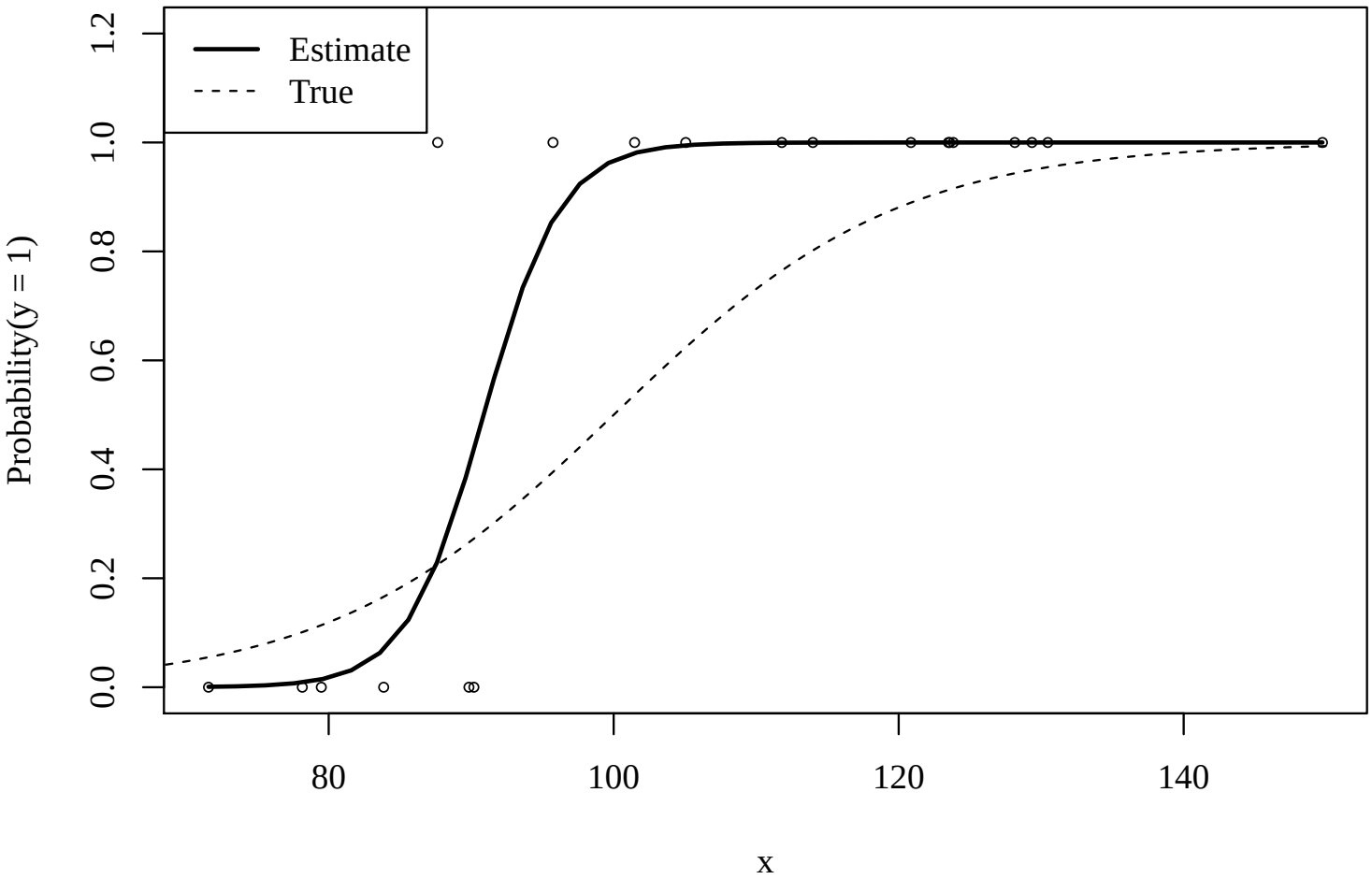
Sample 3


Estimate
True
Probability(y = 1)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
80
100
120
140
x

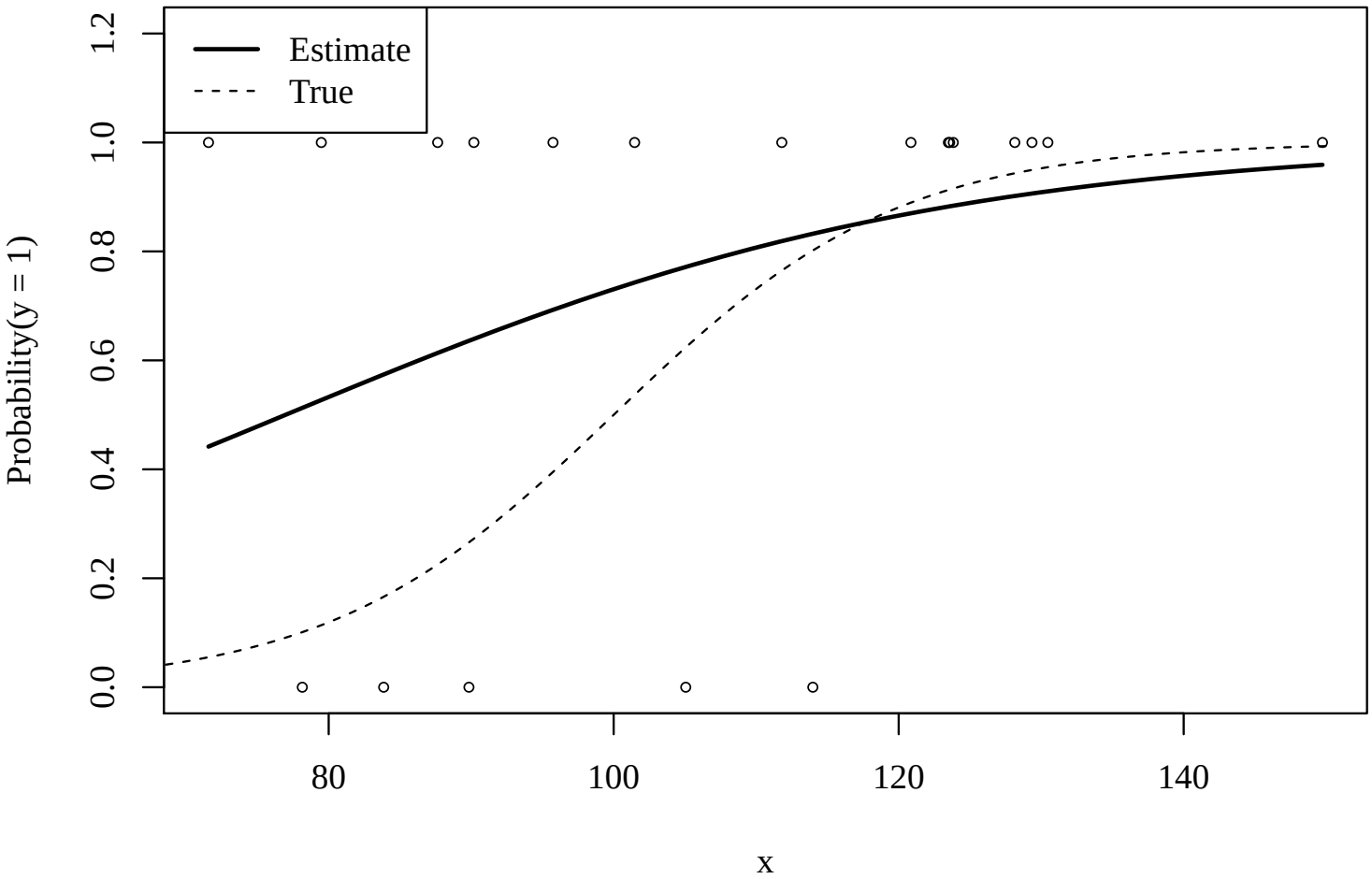
Sample 4


Estimate
True
Probability(y = 1)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
80
100
120
140
x

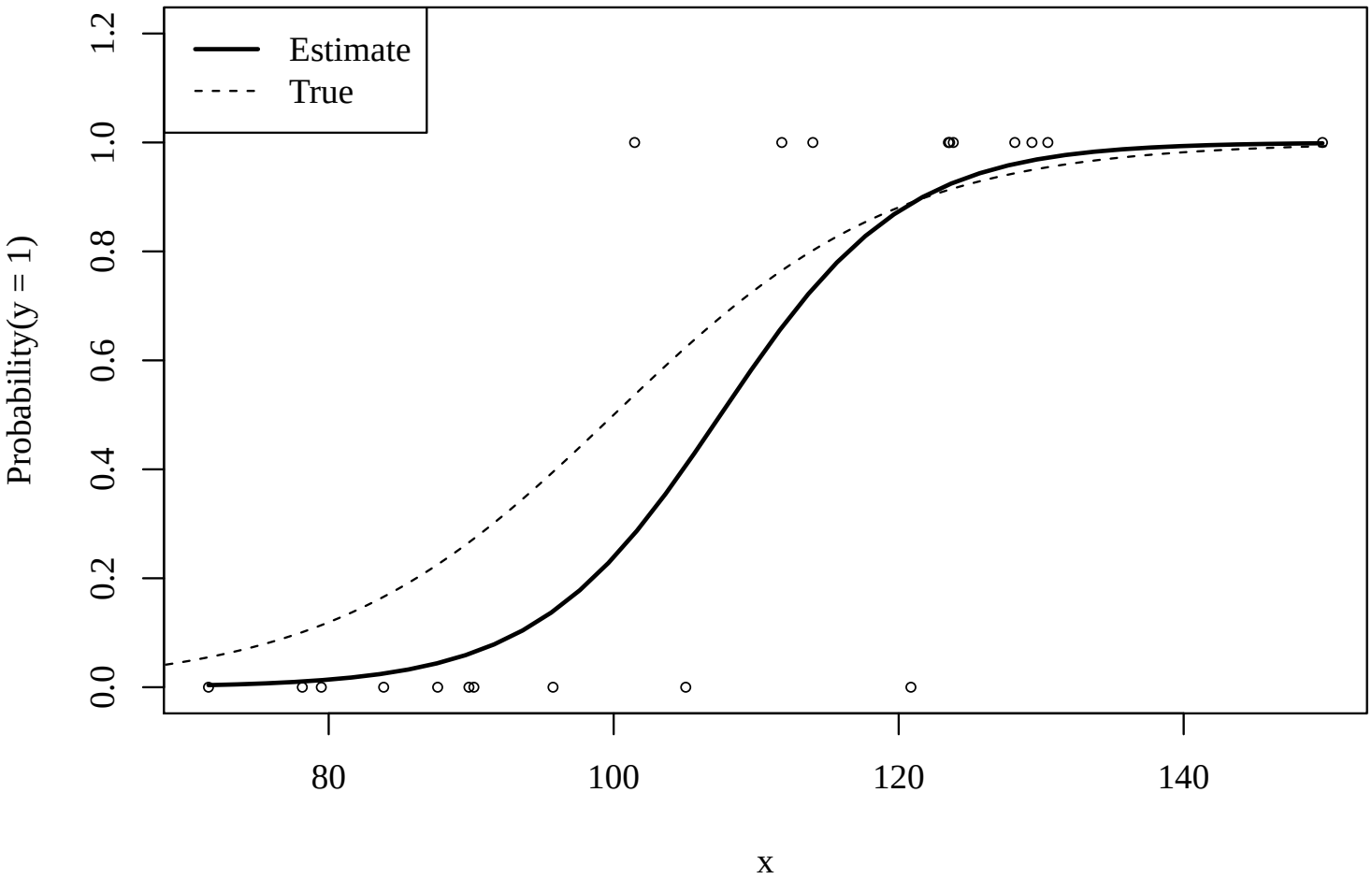
Sample 5


Estimate
True
Probability(y = 1)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
80
100
120
140
x

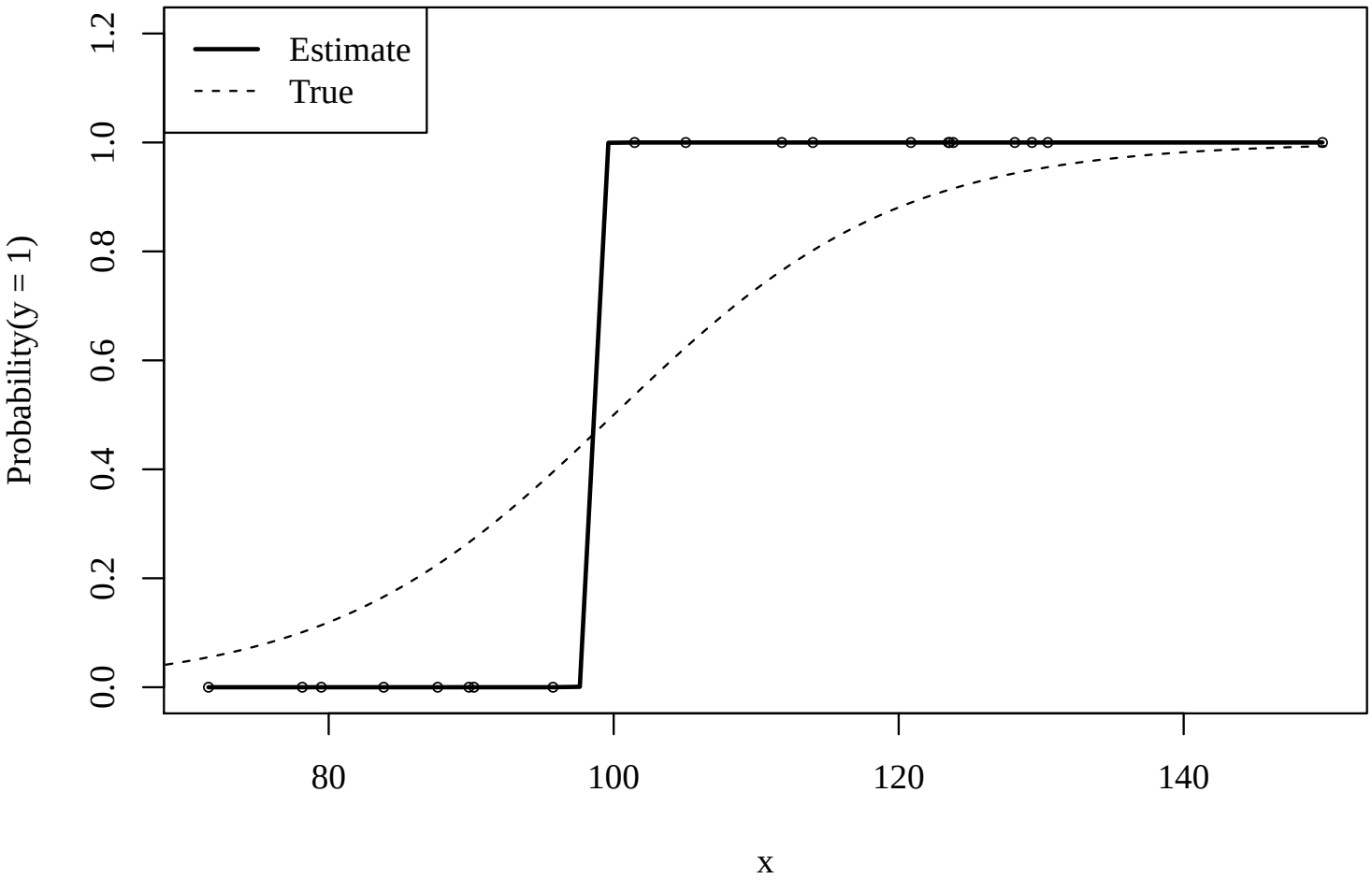
Sample 6


Estimate
True
Probability(y = 1)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
80
100
120
140
x

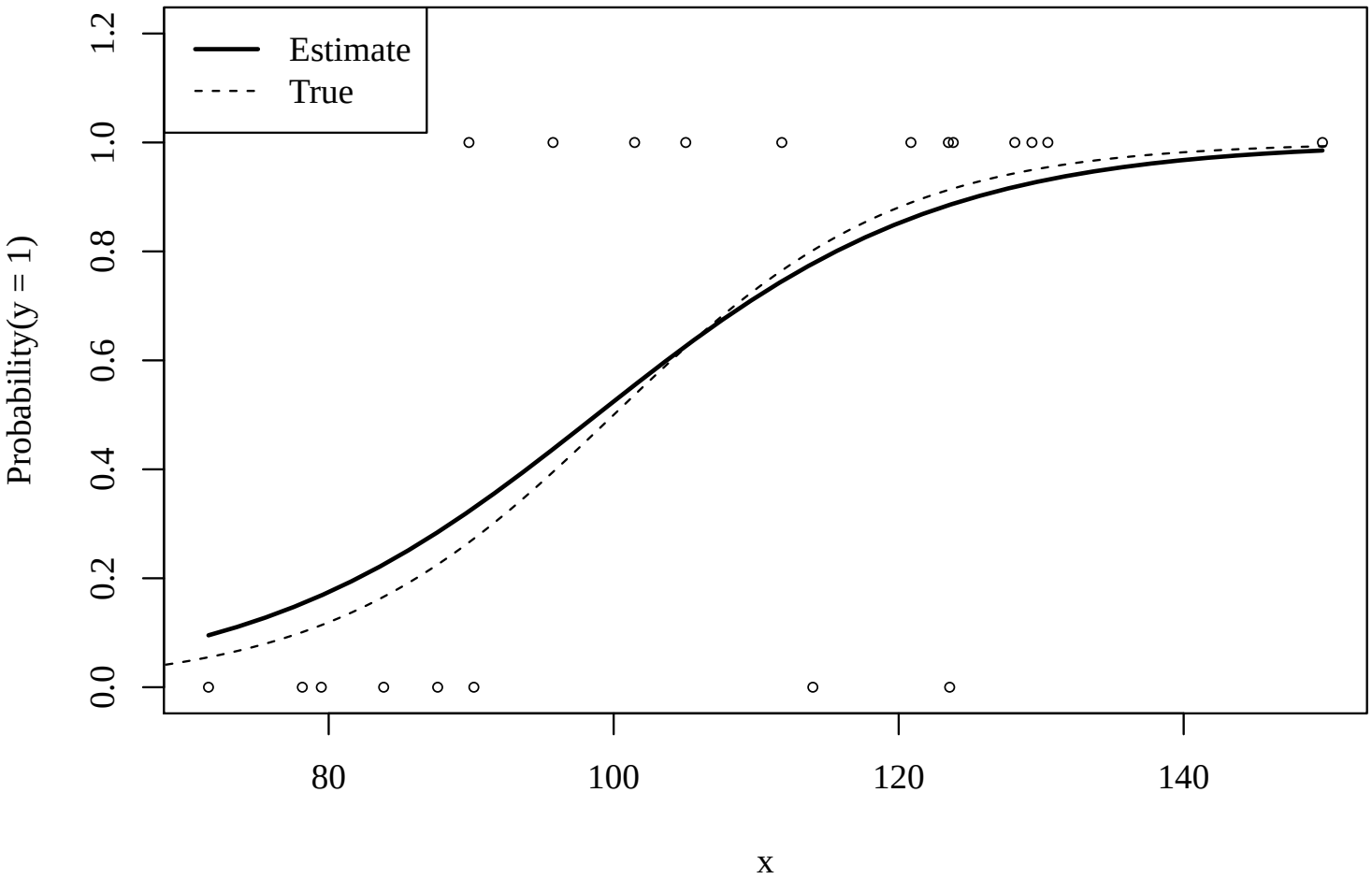
Sample 7


Estimate
True
Probability(y = 1)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
80
100
120
140
x

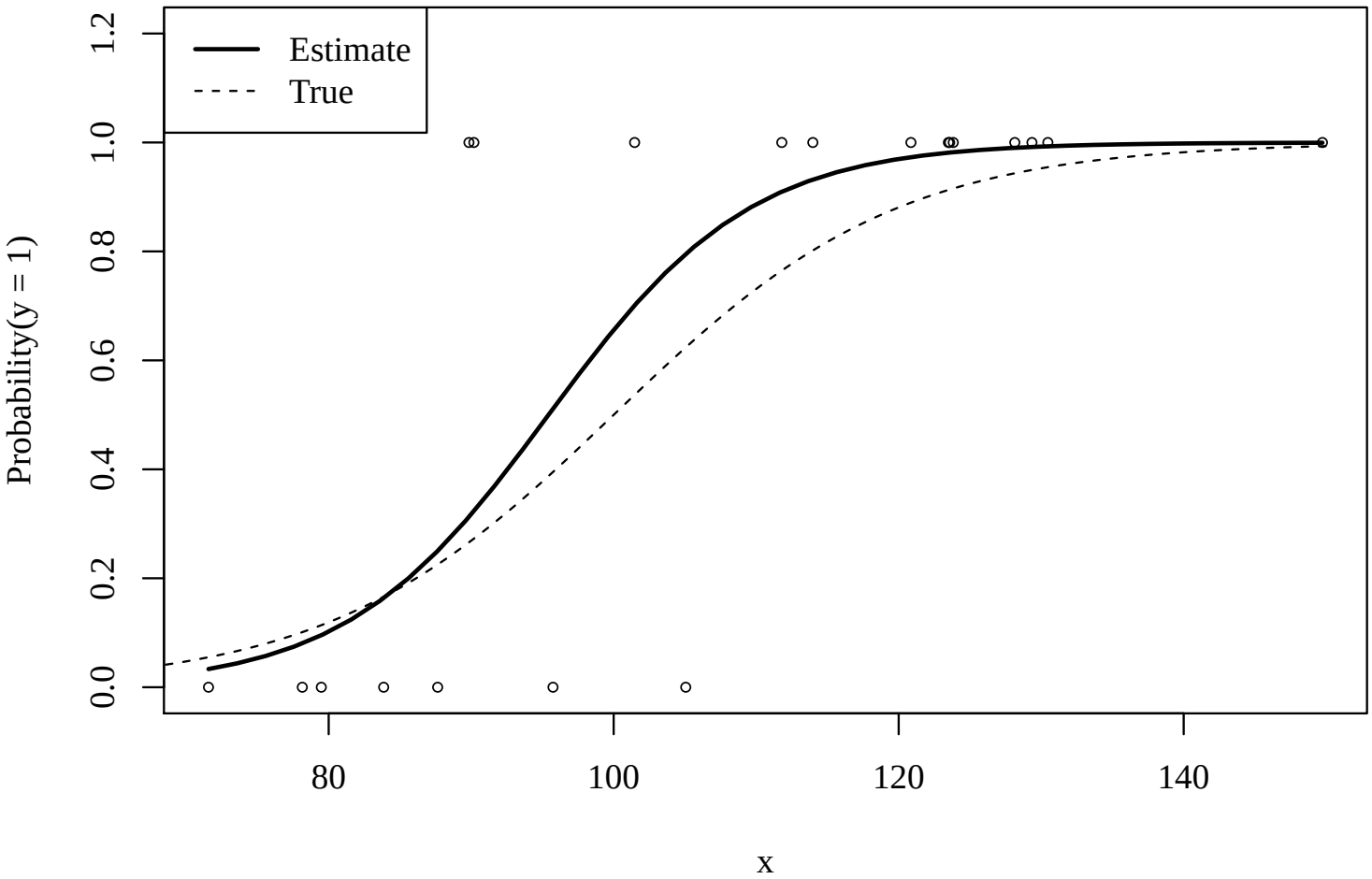
Sample 8


Estimate
True
Probability(y = 1)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
80
100
120
140
x

Sample 9


Estimate
True
Probability(y = 1)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
80
100
120
140
x

Sample 10


Estimate
True
Probability(y = 1)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
80
100
120
140
x

Estimates table

	M1 Estimate (S.E.)	M2 Estimate (S.E.)	M3 Estimate (S.E.)	M4 Estimate (S.E.)	M5 Estimate (S.E.)
(Intercept)	-9.596*	-3.335	-3.760	-27.407	-33.668
	(4.044)	(2.942)	(2.835)	(15.776)	(23.613)
x	0.097*	0.043	0.045	0.290	0.370
	(0.040)	(0.029)	(0.028)	(0.169)	(0.262)
N	20	20	20	20	20
Deviance	16.294	19.905	21.366	5.998	5.656
$-2LLR(Model\chi^2)$	10.627**	2.589	3.068	19.900***	18.779***

$*p \leq 0.05$** $p \leq 0.01$***$p \leq 0.001$

	M6 Estimate (S.E.)	M7 Estimate (S.E.)	M8 Estimate (S.E.)	M9 Estimate (S.E.)	M10 Estima (S.E.
(Intercept)	-10.935*	-16.649*	-727.991	-8.162*	-13.462
	(5.128)	(6.955)	(574012.197)	(3.602)	(5.941)
x	0.120*	0.155*	7.383	0.083*	0.141*
	(0.055)	(0.064)	(5819.051)	(0.035)	(0.062)
N	20	20	20	20	20
Deviance	13.337	10.839	0.000	18.123	11.817
$-2LLR(Model\chi^2)$	11.098**	16.887***	26.920***	8.798*	14.081*

$*p \leq 0.05** p \leq 0.01***p \leq 0.001$

Funny Business in model 8

- Scroll back to slide 115, the graph of model 8
- The glm function does not throw an error. But, if we are running interactively, there is a warning, which looks like this:

```
> glm(yobs ~ x, dat8, family=binomial)

Call:  glm(formula = yobs ~ x, family = binomial,
    data = dat)

Coefficients:
(Intercept)            x
   -727.991        7.383

Degrees of Freedom: 19 Total (i.e. Null);  18
    Residual
Null Deviance:      26.92
Residual Deviance: 2.652e-09    AIC: 4
```

Funny Business in model 8 ...

```
Warning messages:
1: glm.fit: algorithm did not converge
2: glm.fit: fitted probabilities numerically 0 or
   1 occurred
```

How to comprehend that

- Theoretical model says $Pr(y = 1)$ must always be between 0 and 1, never exactly equal to it.
- But the data wants to say y is certainly 0 up to a point, and then 1 after.
- To "fit" that, the S shaped curve would need kinks
- Why doesn't optimizer notice and stop
  - A well behaved ML has a single peak in the interior of the parameter space
  - An ill-behaved ML surface has a maximum at infinity, so the estimator might keep guessing bigger and bigger values
  - The stopping algorithm quits on the way to an infinite estimate and glm reports that non-converged value.

KU

Funny Business in model 8

One defense method is the safeBinaryRegression package, which provokes an error, rather than a warning. Run like this:

```
> library(safeBinaryRegression)
> glm(yobs ~ x, dat = dat8, family = "binomial")
Error in glm(yobs ~ x, dat = dat8, family =
    "binomial") :
  The following terms are causing separation among
     the sample points: (Intercept), x
```

Outline

KU

Outline ...

- Small Sample with Separation

8. Testing Statistical Significance

9. Model Goodness
   - Percent Correctly Predicted and ROC
   - LLR equivalent of an F test
   - Deviance
   - Why no R square
   - Hosmer and Lemeshow test

Hypo Tests in practice

- Deciding which variable to keep in the model:
  - Is the true effect really 0?
  - Is it safe to omit that variable from the equation
- The best practice is to fit a larger model, then fit a model that omits a variable, and do the analysis of deviance (liklihood ratio test) to test (Hastie, Tibshirani, Friedman, ESL 2ed, p. 124).
- However, because that requires calculation, there is emphasis
- Testing one parameter at a time is somewhat dangerous.
  - z test, or Wald χ^2 test, is approximate unless sample is very large
  - remember ML coefficient estimates are known to be biased
- Venables and Ripley (MASS) caution about a particular kind of numerical instability in Logit models, known as the Hauck Donner effect.
  - They suggest that hypothesis tests be done with a likelihood ratio test comparing models that do and don't have a single variable.

z-test and t-test: two approximations

- Asymptotically, $\hat{\beta}$ is Normal (recall fundamental ML Theory). What about the quantity:

$$\frac{\hat{\beta} - \beta_{null}}{s.e.(\hat{\beta})}$$

- it looks like a $t - test$, doesn't it?
- More correct to say it is a z statistic, approximately Normally distributed

$$z = \frac{\hat{\beta} - \beta_{null}}{s.e.(\hat{\beta})} \tag{21}$$

Wald χ^2 test: another approximation

- The Wald Chi-square is the ratio of the squared estimate to the variance,$\hat{\beta}^2/Var(\hat{\beta})$. But alert users will note that it is simply z^2.

$$z^2 = \left(\frac{\hat{\beta}}{s.e.(\hat{\beta})} \right)^2 \tag{22}$$

- Wald contended that value is distributed as a χ^2 variable. The square root $\hat{\beta}/se(\hat{\beta})$ is approximately Normal (and also approximately t-distributed). That explains why some programs call the statistic $\hat{\beta}/se(\hat{\beta})$ a t variable, while others call it a Z statistic.
- From elementary statistics, we know that the square root of a χ^2 variable with one degree of freedom is Normal(0,1), so the Wald Chi-Square test for a single parameter is actually substantively IDENTICAL to the t or z approaches.

Wald χ^2 test: another approximation ...

- The Wald Chi-Square can be used to simultaneously test several coefficients.

$$\hat{\beta} Var(\hat{b})^{-1} \hat{\beta}$$

- Note that if we were testing only one parameter, this degenerates to the preceding equation.

likelihood ratio test (Compare nested models)

The Unrestricted "Full" Model. Let L_{max} be the value of the likelihood function at its maximum, when all coefficients, the slope and the intercept, are estimated to maximize the likelihood.

The Restricted Model. Set some parameter at a fixed value, possibly 0. Let L_0 be the value of the likelihood function in which the "slope" coefficient β_1 (or other coefficients if they are in the model) is 0.

- The likelihood ratio test can be used to compare **nested** models that are estimated ON THE **SAME DATA**

Compare L_{max} and L_0 with a χ^2 distribution.

Let λ, (Greek "lambda"), be the ratio of L_0 to L_{max}:

$$\lambda = \frac{L_0}{L_{max.}} \tag{23}$$

$-2 \cdot \ln(\lambda)$ has a χ^2 distribution with k degrees of freedom, where k is the difference in the number of coefficients in L_0 versus L_{max}).

Where do they get standard errors?

- Fisher showed that the second derivative matrix of the likelihood function, dubbed the "Information matrix", can be used to derive a variance-covariance matrix for the coefficient estimates.
  - The var-covar matrix is -1 times the inverse of the information matrix
- The estimator for $\hat{\beta}$ is wandering about in a p dimensional space, trying to find the h

Where do they get standard errors?...

- Here's an intuition:

- If the likelihood surface looks like a sharply peaked mountain, then we are very confident we are at the maximum, and the standard errors will be small.

- If the likelihood surface looks like a rounded weathered foothill, then we are uncertain about the estimates, and the standard errors will be big.

- And if the likelihood surface looks like a mole hill, we stop doing this kind of work for a while

Anything to say about variance-covariance?

- The Variance matrix in logistic regression is interestingly similar to OLS
  - OLS:

$$Var(\hat{\beta}^{OLS}) = \sigma^2(X'X)^{-1} \tag{24}$$

  - Logistic Regression:

$$Var(\hat{\beta}^{logistic}) = \left[X'WX\right]^{-1} \tag{25}$$

  - W has the estimated probabilities for the individual cases:

$$W = \begin{bmatrix} P_1(1-P_1) & 0 & 0 & 0 & 0 \\ 0 & P_2(1-P_2) & 0 & 0 & 0 \\ 0 & \cdots & & & \\ 0 & \cdots & & & \\ 0 & 0 & 0 & 0 & P_N(1-P_N) \end{bmatrix} \tag{26}$$

- This variance-covariance matrix is rather similar to what we would get if we used WLS in the linear probability model.

Outline

Outline ...

- Small Sample with Separation

8 Testing Statistical Significance

9 Model Goodness

- Percent Correctly Predicted and ROC
- LLR equivalent of an F test
- Deviance
- Why no R square
- Hosmer and Lemeshow test

Predicted value table

- Use the predicted probabilities P_i to predict 0 or 1 for each case
- The *percent correctly predicted* has sometimes been emphasized as a summary of model accuracy
  - However, usually it is uninformative
  - E.g., if 80% of the observations are 1, then a model that predicts 1 for all cases will be correct 80% of the time.
- Need a way to penalize by accounting for incorrect predictions.

KU

Confusion matrix

- Make a table like so. Commonly called a "confusion matrix"

		Predicted		
		0	1	Count in Data
Observed	0	true-negative (TN)	false-positive(FP)	N (#0's)
	1	false-negative (FN)	true-positive(TP)	P (#1's)

- true positive rate (TPR), also called "sensitivity"

$$sensitivity = \frac{TP}{TP + FN}$$

- true negative rate (TNR), also called "specificity"

$$specificity = \frac{TN}{TN + FP}$$

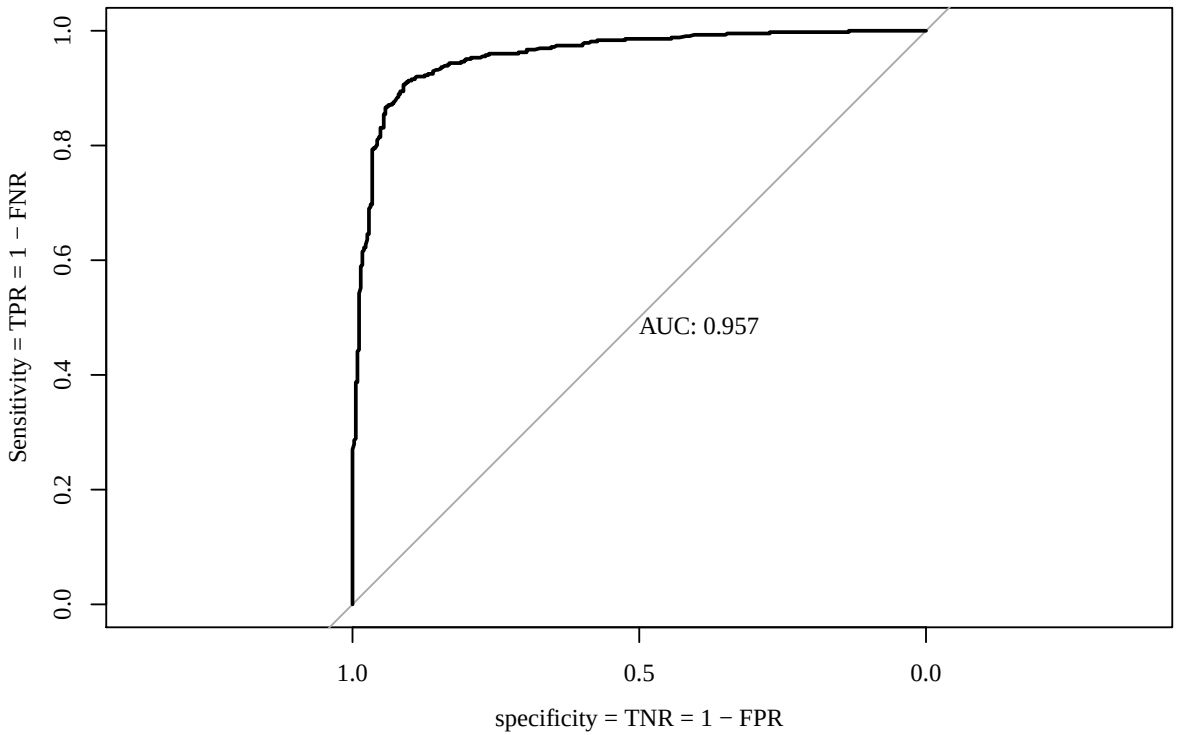
- accuracy is the plain old "percent correctly predicted"

$$\frac{TN + TP}{TP + TN + FP + FN}$$

ROC

- ROC Curve: Receiver Operating Characteristic curve is a visual summary of how the model's fit with the data will depend on alternative methods of calculating predictions.
- Before I used simple idea that dividing between 0 and 1 is $P_i = 0.5$
- Now we change the threshold, called τ, or lower it, to somehow make the predictions fit the data more accurately.
- Perhaps if we raise the threshold to predict 1 to .6, we will make fewer "false positive predictions" and have more "true negatives".

ROC diagram (from the LogitProbit-Worked Example)



Explaining that ROC diagram

Take the predicted probabilities from the model, $\hat{p}_i$, and calculate predictions for a cutoff point, τ.

Predicted y (0 or 1) is based on $\hat{p}_i$

$$\hat{y}_i(\tau) = \begin{cases} 1 & \hat{p}_i \geq \tau \\ 0 & else \end{cases} \tag{27}$$

We vary τ from 1 to 0 (predictions begin as all 0's, and end up at all 1's).

As we vary τ, we "trace" the curve in previous slide.

anova test in linear models

- when our linear predictor is $\beta_0 + \beta_1 x1_i + \beta_2 x2_i + \beta_3 x3_i + \ldots + \beta_p xp_i$
- Linear regression output includes an F test, which is a test of that larger fitted model against a linear predictor that includes only the intercept

$$y_i = \beta_0 + \varepsilon_i$$

- That F asks "are all of my predictor coefficients equal to 0".
  - not a very informative test, but we often do report it.

The ML equivalent of and F test

- A statistic, often referred to as a "model χ^2" test, can be calculated as a likelihood ratio test.
- The value referred to in my regression tables, -2LLR, is approximately distributed as a χ^2 statistic:

$$-2ln\left[\frac{likelihood\,of\,intercept\,only\,model}{likelihood\,of\,model\,including\,predictors}\right]$$

- If this value is large, it means at least one of the predictor coefficients is not 0.
  - again, not very informative

Deviance to diagnose model specification

- Here we think of model comparison in a different direction
- A "**saturated model**" is one in which we are include unique a unique predictors for each combination of the input variables.
  - The saturated model typically has likelihood 1, its predicted values exactly match the observed data.
  - Saturated model log likelihood is thus usually 0
- If a fitted model's likelihood is very far from the saturated model, then perhaps "something is wrong."
  - Maybe more predictors are needed?
  - Maybe the functional form should be changed?
  - New probability model?
- The standard output for a logistic regression includes 2 values
  - Null Deviance
  - Residual Deviance

KU

Deviance to diagnose model specification ...

- Residual deviance is difference between a "saturated" model and your "fitted" model

$$-2ln\left[\frac{Likelihood\,of\ fitted\,model}{Likelihood\,of\ saturated\,model}\right] \tag{28}$$

  it usually boils down to

$$-2ln(Likelihood\,of\ fitted\,model)$$

  .

- Residual deviance, which is commonly referred to as Deviance, indicates of "how bad" your model is when compared against the saturated model.

Deviance to diagnose model specification ...

- Myers, Montgomery, and Vining (2002) observe, (I'm paraphrasing notation here to match the above notation) "Formally, an insignificant value of (deviance) in a one-tailed test implies that the fit of the model is not significantly worse than that of the saturated model. ... Often the rule of thumb is applied that the quality of fit is reasonable if $\frac{deviance}{N-p}$ is not appreciably larger than 1. The rule of thumb is prompted by the fact that N-p is the mean of the χ^2_{N-p} distribution"(p. 113).

R^2?

- several approximate R^2 statistics have been proposed
- Since I think R^2 is silly in ordinary linear models, you can guess how excited I am to work on approximate or "pseudo R^2" for logit models.
- What is silly about R^2. Read King(1986)

H-L test: Do your Predicted Probabilities Match Observed Percentages?

- Calculate predicted probabilities, $\hat{P}_i$ for all cases.
- Sort the data by $\hat{P}_i$. Subdivide the sample into subgroups.
- Find out if the observed frequency of 1's and 0's matches the estimated probabilities from the model.

H-L Test boils down to a χ^2 statistic

- Pick some pleasant number of subgroups, say 10. For each subgroup, one can calculate the observed "success rate" O_i and an expected (from the model) success rate, and the χ^2 test is used to find out if the model is grossly out-of-whack.

$$homer.and.lemeshow^2_\chi = \sum_{i=1}^{10} \left[\frac{(O_i - E_i)^2}{E_i} \right] \tag{29}$$

If the χ^2 value is extreme by that standard, it means that your predicted probabilities do not match the observations very well.

That is informative, but not too informative. It does not tell you if the model is "off" for any particular reason, and there could be many suspects in your search for the criminal.

References

Cummings, P. (2009). The Relative Merits of Risk Ratios and Odds Ratios. *Archives of Pediatrics & Adolescent Medicine*, 163(5), 438–445.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1), 27–38.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning data mining, inference, and prediction*. New York: Springer.

King, G. & Zeng, L. (2001). Logistic Regression in Rare Events Data. *Political Analysis*, 9(2), 137–163.

Lee, J., Tan, K. S., & Chia, K. S. (2009). A Practical Guide for Multivariate Analysis of Dichotomous Outcomes. *Annals Academy of Medicine, Singapore*, 38(8), 714–719.

KU

References ...

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*. Boca Raton: Chapman and Hall/CRC, 2nd edition.

Sainani, K. L. (2011). Understanding Odds Ratios. *PM&R*, 3(3), 263–267.

Session

```
sessionInfo()
```

```
R version 3.5.1 (2018-07-02)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 18.04.1 LTS

Matrix products: default
BLAS: /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.7.1
LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.7.1

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
     LC_TIME=en_US.UTF-8
 [4] LC_COLLATE=en_US.UTF-8     LC_MONETARY=en_US.UTF-8
     LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C                  LC_ADDRESS=C
[10] LC_TELEPHONE=C             LC_MEASUREMENT=en_US.UTF-8
    LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] rockchalk_1.8.124   stationery_0.98.5.5
```

Session ...

```
loaded via a namespace (and not attached):
 [1] Rcpp_0.12.17     knitr_1.20       magrittr_1.5     splines_3.5.1
     kutils_1.49      MASS_7.3-50
 [7] mnormt_1.5-5     lattice_0.20-35 pbivnorm_0.6.0   xtable_1.8-2
     minqa_1.2.4      carData_3.0-1
[13] stringr_1.3.1    plyr_1.8.4       tools_3.5.1      grid_3.5.1
     nlme_3.1-137     htmltools_0.3.6
[19] lme4_1.1-17      digest_0.6.15    rprojroot_1.3-2 lavaan_0.6-1
     Matrix_1.2-14    zip_1.0.0
[25] nloptr_1.0.4     evaluate_0.10.1 rmarkdown_1.10   openxlsx_4.1.0
     stringi_1.2.3    compiler_3.5.1
[31] backports_1.1.2 stats4_3.5.1     foreign_0.8-70
```