
# Lecture 5: Tricky DVs

Dr. Rachel Blum

November 12, 2019

# Beyond OLS

# Motivation: inference

- We have theories. We want to know how "true" they are.
- We have data. We want to know how that data could help us evaluate our theories.

# Motivation: inference

- We have theories. We want to know how "true" they are.
- We have data. We want to know how that data could help us evaluate our theories.
- One key approach to these questions comes from **Bayes' Theorem**:

$$P(hypothesis|data) = \frac{P(data|hypothesis) \times P(hypothesis)}{P(data)}$$

# Bayes' Theorem

$$P(hypothesis|data) = \frac{P(data|hypothesis) \times P(hypothesis)}{P(data)}$$

- The problem is, we don't know $P(hypothesis)$.
- $P(data|hypothesis)$ is not so difficult. Recall the p-value.
- $P(data)$ is calculable. It's a function of $P(data|hypothesis)$ and $P(hypothesis)$ and it drops out.

# Bayes theorem example

Suppose we want to catch terrorists. We know that a large percentage of terrorists are Muslim. Should we not target Muslims in attempting to catch terrorists?

# Bayes theorem example

# Bayes theorem example

- Probability that a terrorist is a Muslim: 95%
  - A lot depends on how we define ‘Muslim’ and especially ‘terrorist.’

# Bayes theorem example

- Probability that a terrorist is a Muslim: 95%
  - A lot depends on how we define ‘Muslim’ and especially ‘terrorist.’
- Probability that a person is a Muslim: 23%
  - Smaller in United States or in western airports, etc.

# Bayes theorem example

- Probability that a terrorist is a Muslim: 95%
  - A lot depends on how we define ‘Muslim’ and especially ‘terrorist.’
- Probability that a person is a Muslim: 23%
  - Smaller in United States or in western airports, etc.
- Probability that a person is a terrorist: 0.01%
  - An absurdly large “guess”.

# Bayes theorem example

$$P(Muslim) = probabilityofbeingMuslim = 0.23$$

$$P(terrorist) = probabilityofbeingaterrorist = 0.0001$$

$$P(Muslim|terrorist) = probabilityaterroristisMuslim = 0.95$$

# Bayes theorem example

$P(Muslim) = probabilityofbeingMuslim = 0.23$

$P(terrorist) = probabilityofbeingaterrorist = 0.0001$

$P(Muslim|terrorist) = probabilityaterroristisMuslim = 0.95$

We want to know $P(terrorist|Muslim)$

$$P(terrorist|Muslim) = \frac{P(Muslim|terrorist) \times P(terrorist)}{P(Muslim)}$$
$$= \frac{(0.95)(0.0001)}{(0.23)}$$
$$= \frac{0.114}{0.23} = 0.0004$$

# Bayes theorem example

$P(Muslim) = probabilityofbeingMuslim = 0.23$

$P(terrorist) = probabilityofbeingaterrorist = 0.0001$

$P(Muslim|terrorist) = probabilityaterroristisMuslim = 0.95$

We want to know $P(terrorist|Muslim)$

$$P(terrorist|Muslim) = \frac{P(Muslim|terrorist) \times P(terrorist)}{P(Muslim)}$$
$$= \frac{(0.95)(0.0001)}{(0.23)}$$
$$= \frac{0.114}{0.23} = 0.0004$$

or a 0.004% chance that a Muslim is a terrorist.

# Bayesian and Frequentism compared

- **Bayesian statistics** attempts to use your beliefs about P(hypothesis) as a basis for calculating P(hypothesis|data).

# Bayesian and Frequentism compared

- **Bayesian statistics** attempts to use your beliefs about P(hypothesis) as a basis for calculating P(hypothesis|data).
  - Maximum Likelihood framework says P(data|hypothesis) is at least proportional to P(hypothesis|data), so P(data|hypothesis) is all we really care about.

# Bayesian and Frequentism compared

- **Bayesian statistics** attempts to use your beliefs about P(hypothesis) as a basis for calculating P(hypothesis|data).
  - Maximum Likelihood framework says P(data|hypothesis) is at least proportional to P(hypothesis|data), so P(data|hypothesis) is all we really care about.
- Frequentist statistics (OLS) will just learn to live with P(data|hypothesis)

# Bayesian and Frequentism compared

- **Bayesian statistics** attempts to use your beliefs about P(hypothesis) as a basis for calculating P(hypothesis|data).
  - Maximum Likelihood framework says P(data|hypothesis) is at least proportional to P(hypothesis|data), so P(data|hypothesis) is all we really care about.
- Frequentist statistics (OLS) will just learn to live with P(data|hypothesis)
  - Hypothesis testing: Reject null hypothesis when P(data|null hypothesis) is small.

# When OLS Doesn't Work

# When OLS works (Gauss Markoff)

When correctly specified, OLS is the **B**est **L**inear **U**nbiased **E**stimator of the regression line.

# OLS Conditions: 1

The model is correctly specified.

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

# OLS Conditions: 2

Errors are distributed with mean zero.

$$E(\epsilon) = 0$$

# OLS Conditions: 3

Errors have constant variance (homoskedasticity).

$$Var(\epsilon) = \sigma^2 (vs. {\sigma_i}^2)$$

# OLS Conditions: 4

Errors are uncorrelated with each other (no autocorrelation).

$$Cov(\epsilon_i, \epsilon_j) = 0, i \neq j$$

# OLS Conditions: 5

Errors are uncorrelated with the independent variables.

$$Cov(\epsilon_i, X_i) = 0$$

# OLS Conditions: 6

*Helpful*: errors are normally distributed, especially for small $N$).

$$\epsilon \sim N(0, \sigma_2)$$

# OLS: the DV

The dependent variable is a linearly additive function of a set of linearly independent independent variables, plus a disturbance term.

# OLS: the IV(s), 1

There is no error associated with the measurement of the IV ($X$).

# OLS: the IV(s), 2

Regressors are linearly independent (i.e., *not* multicollinear).

# OLS: the IV(s), 3

You have more observations than regressors. If

$$X = n \times k \implies n > k$$

# OLS: THE PUNCHLINE

OLS only works if you are modeling a linear relationship.

# Consequences of Gauss Markoff violations

If linearity assumptions are violated $\Rightarrow$ we have estimated the wrong thing.

*Results are meaningless.*

# Consequences of Gauss Markoff violations

If autocorrelation of errors occurs $\Rightarrow$ standard errors are wrong.

*Sign of mis-specification.*

# Consequences of Gauss Markoff violations

If key variable is omitted $\Rightarrow$ estimates will be incorrect.

*Standard errors (and p-values, etc.) will also be incorrect.*

# Non-linearity: dichotomous outcomes

# Dichotomous outcomes

- A **dichotomous outcome** is divided in two parts: the outcome either occurred or it did not.

# Dichotomous outcomes

- A **dichotomous outcome** is divided in two parts: the outcome either occurred or it did not.
    - A policy intervention either works or does not work.

# Dichotomous outcomes

- A **dichotomous outcome** is divided in two parts: the outcome either occurred or it did not.
  - A policy intervention either works or does not work.
  - Votes can be either in favor of or against a policy or representative.

# Dichotomous outcomes

- A **dichotomous outcome** is divided in two parts: the outcome either occurred or it did not.
  - A policy intervention either works or does not work.
  - Votes can be either in favor of or against a policy or representative.
  - Survey respondents are either for or against a given prompt

# Enter: Maximum Likelihood Estimation (MLE)

- Practical motivation: A model for when OLS might not work so well.
- Theoretical motivation: Making inferences based on probabilities.
- The mechanics: What are we maximizing the likelihood of, and how?

# MLE, briefly

**MLE is a different estimator for the parameters of a model** (another way to estimate $\beta$.)

> Especially helpful when the relationship you're estimating is not linear.

# Logit and Probit

- Both `probit` and `logit` models estimate & predict probable relationship between *Y* and *X*.

# Logit and Probit

- Both `probit` and `logit` models estimate & predict probable relationship between *Y* and *X*.
- Predicted probabilities range between 0 and 1.

# Logit and Probit

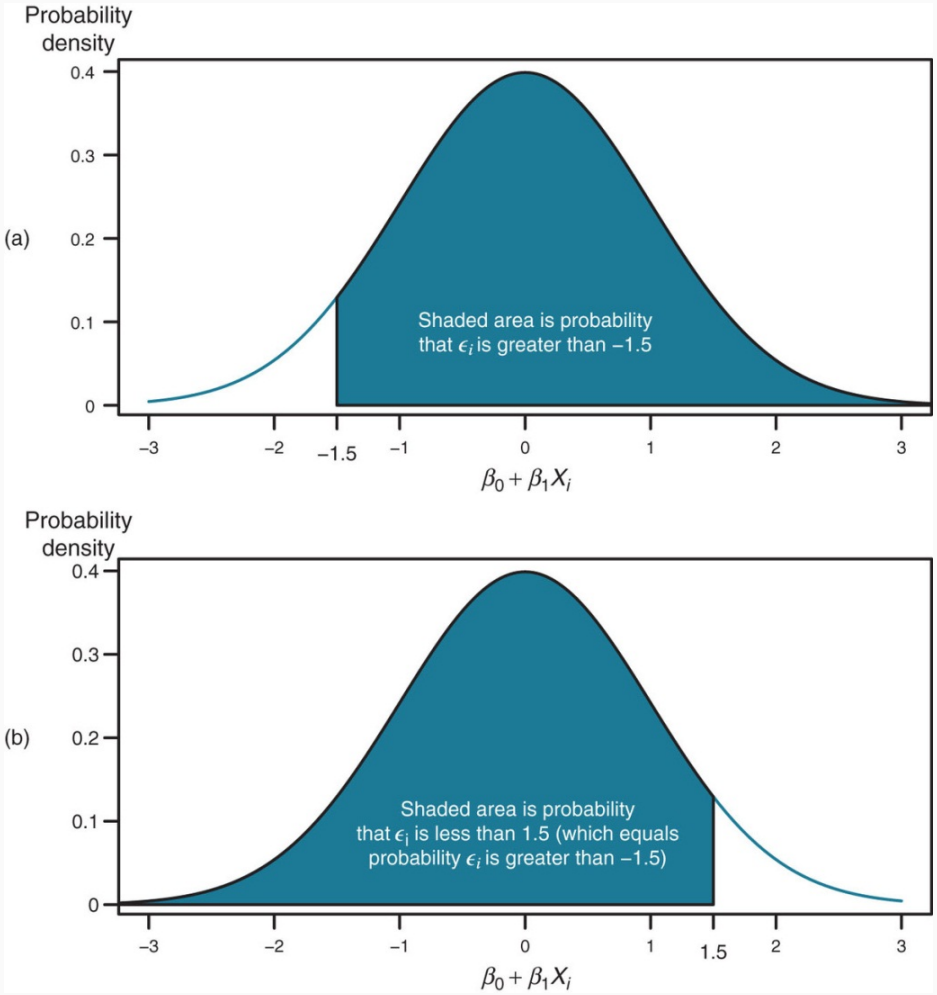
- Both `probit` and `logit` models estimate & predict probable relationship between *Y* and *X*.
- Predicted probabilities range between 0 and 1.
- Achieved by fitting `S` curve to the data

# Logit and Probit

- Both `probit` and `logit` models estimate & predict probable relationship between *Y* and *X*.
- Predicted probabilities range between 0 and 1.
- Achieved by fitting **S** curve to the data
- Practically the same (but we will consider each in turn)

# Two models for dichotomous DVs: Logit and Probit

`Logit` and `Probit` both provide a curve bounded between 0 and 1.



# Probit model

- `Probit` model is one way to estimate the effects of IVs on a dichotomous outcome.
- **Key assumption**: error term $\epsilon$ has a normal distribution.
- Model estimates the **probability** of $Y$ taking a certain value, given whatever $X$'s value is:

$$Pr(Y = 1|X_1) = Pr(\epsilon \leq \beta_0 + \beta_1 X_1)$$

# Normal distribution and CDF

**CDF** = *Cumulative Density Function* of a normal distribution.

# Normal distribution and CDF

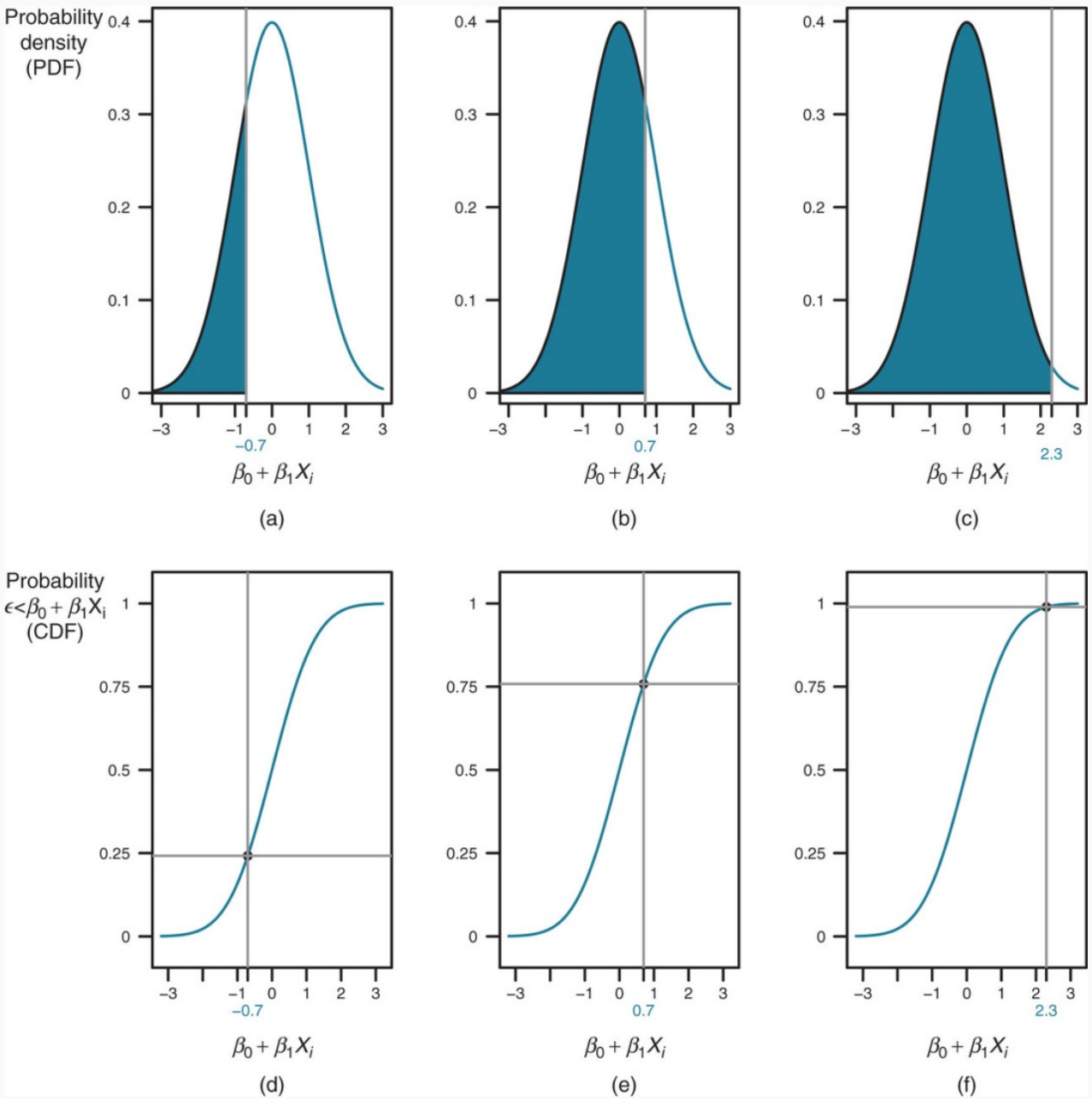
**CDF** = *Cumulative Density Function* of a normal distribution.

CDF written as probability of something being < a certain value in a *normal distribution*.

# Normal distribution and CDF

**CDF** = *Cumulative Density Function* of a normal distribution.

CDF written as probability of something being < a certain value in a *normal distribution*. If we supply the CDF with any number it will give us the probability that a normally distributed random variable is less than that number.


Probability density (PDF)
0.4
0.3
0.2
0.1
0
−3
−1
0
1
2
3
−0.7
0.7
2.3
$\beta_0 + \beta_1 X_i$
(a)
(b)
(c)
Probability $\epsilon < \beta_0 + \beta_1 X_i$ (CDF)
1
0.75
0.5
0.25
0
(d)
(e)
(f)

# Probit notation

CDF has its own notation:the Greek letter $\Phi$

# Probit notation

CDF has its own notation:the Greek letter $\Phi$ $\Phi$ indicates probability that a normally distributed random variable is less than whatever number we give it.

$$\begin{aligned} Pr(Y = 1) &= Pr(\epsilon \leq \beta_0 + \beta_1 X_1) \\ &= \boldsymbol{\Phi}(\beta_0 + \beta_1 X_1) \end{aligned}$$

# Boiled down

`Probit` estimates $\hat{\beta}$ that lead to high predicted probabilities (*close to 1*) for observations of *Y* that were actually **1**, and low predicted probabilities (*close to 0*) for observations of *Y* that were actually **0**.

# Logit model

`Logit` model is another way to estimate relationship between $X$ and $Y$ when the outcome is dichotomous.

# Logit model

`Logit` model is another way to estimate relationship between $X$ and $Y$ when the outcome is dichotomous.

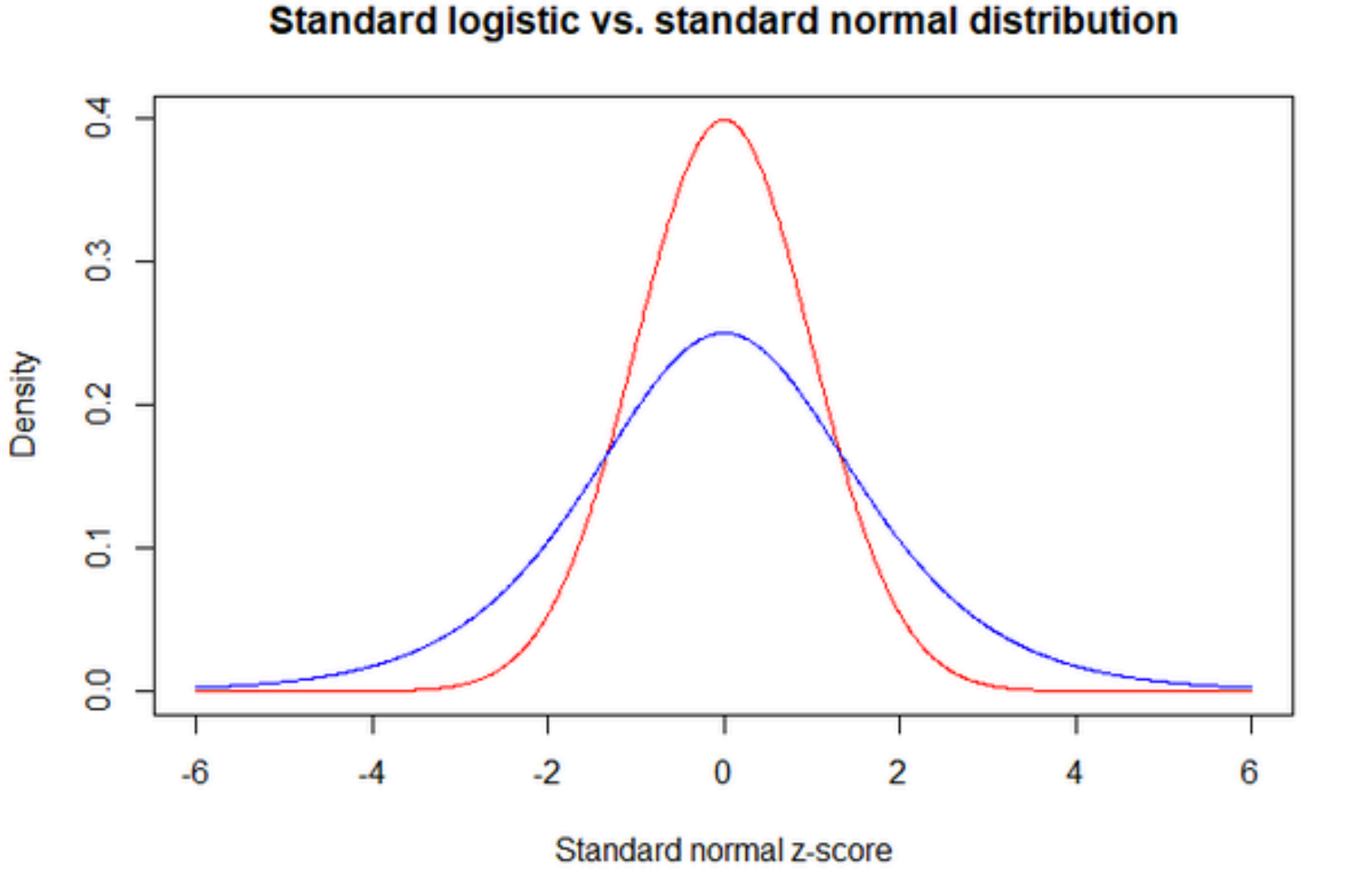
- The only mathematical difference between `logit` and `probit` is how we assume the error term is distributed...

# Logit model

`Logit` model is another way to estimate relationship between $X$ and $Y$ when the outcome is dichotomous.

- The only mathematical difference between `logit` and `probit` is how we assume the error term is distributed...
- In a logit model, we assume the error term follows a **logistic distribution**.

# Logistic vs. normal distribution



# Logit notation

We again use **CDF** notation to specify how we estimate our model.

# Logit notation

We again use **CDF** notation to specify how we estimate our model.

- When $\beta_0 + \beta_1 X_1$ is really large, logit function approaches **1**.

# Logit notation

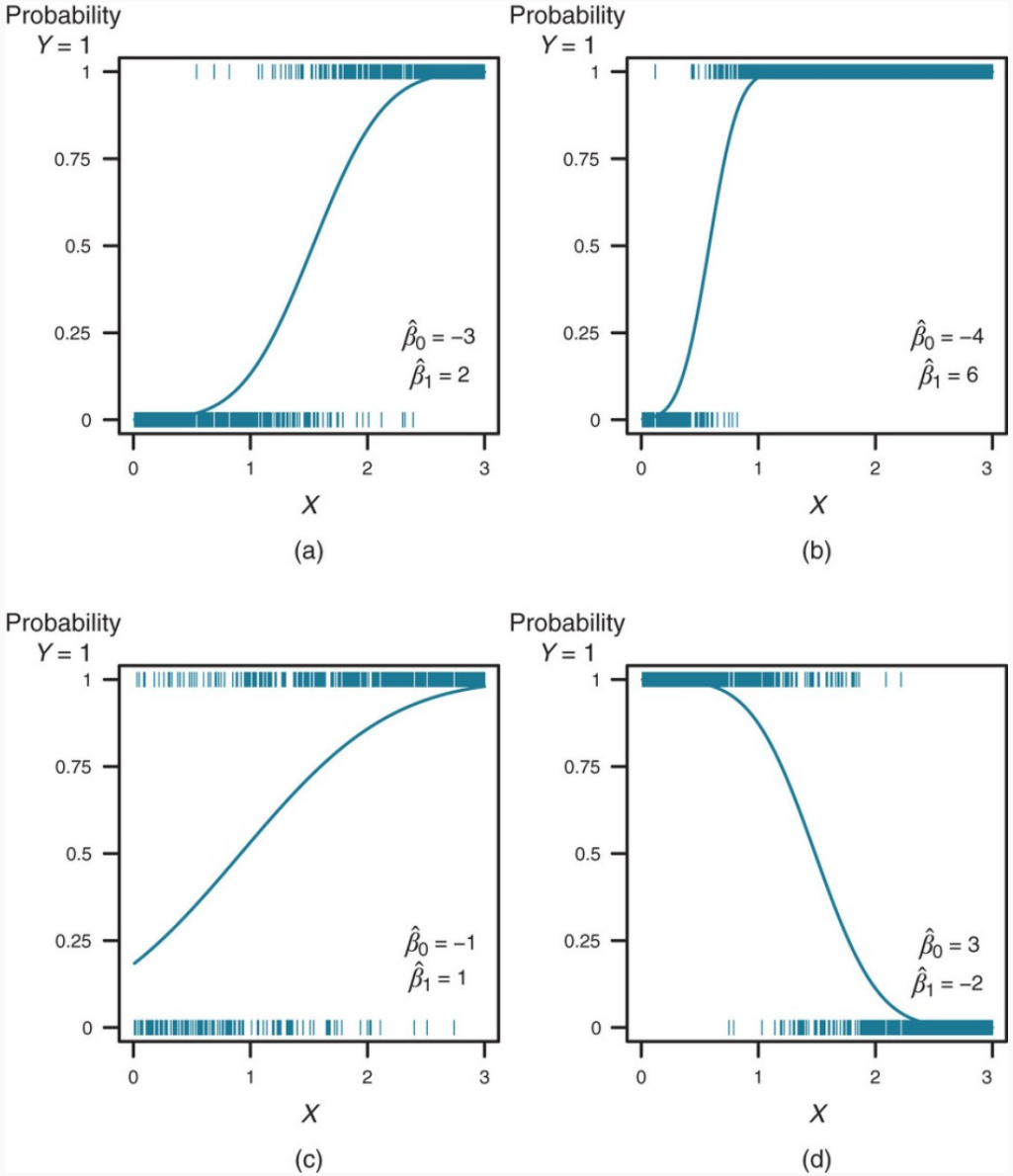
We again use **CDF** notation to specify how we estimate our model.

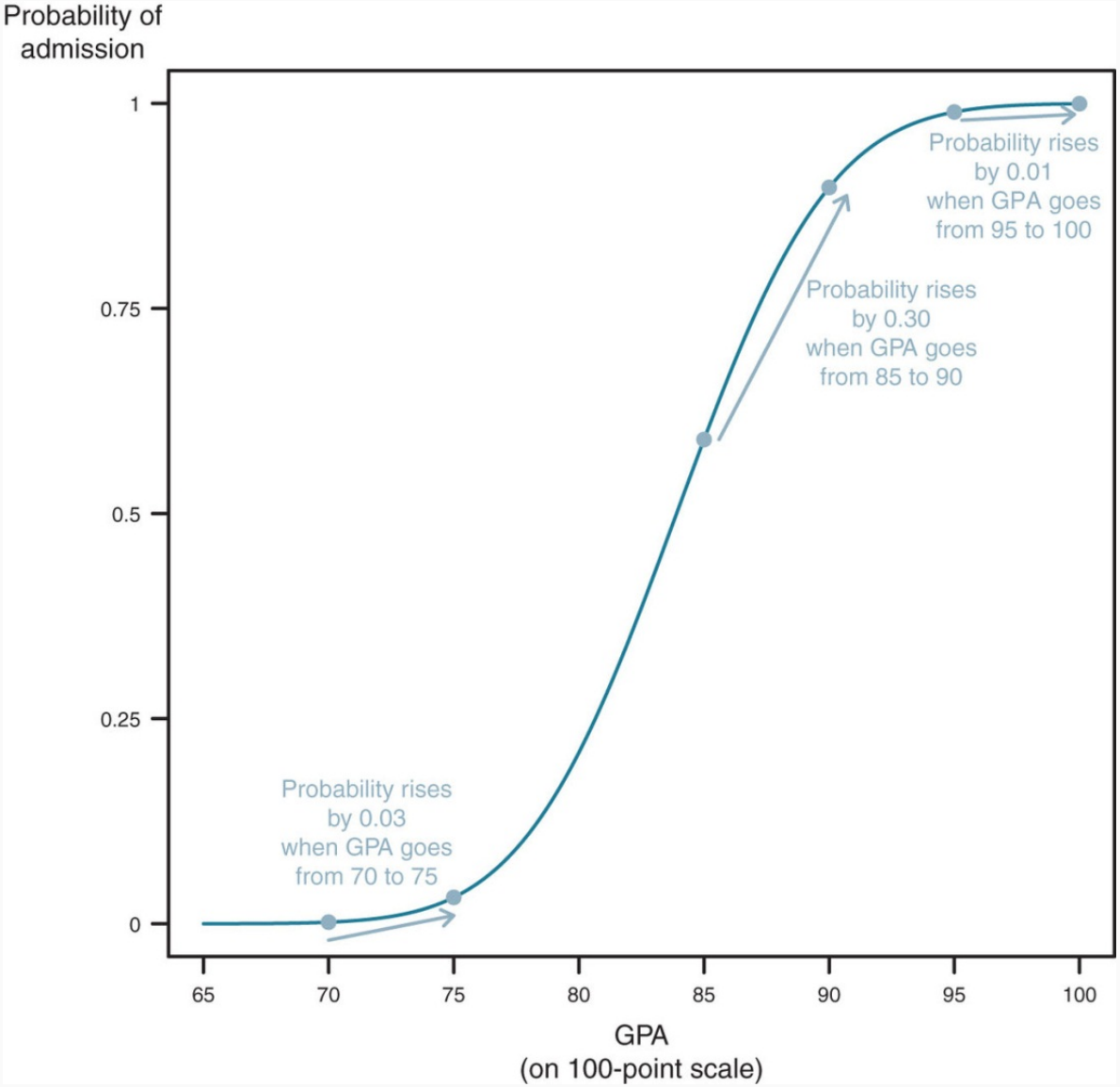
- When $\beta_0 + \beta_1 X_1$ is really large, logit function approaches **1**.
- When $\beta_0 + \beta_1 X_1$ is really small, logit function approaches *0*.

# Logit notation

We again use **CDF** notation to specify how we estimate our model.

- When $\beta_0 + \beta_1 X_1$ is really large, logit function approaches **1**.
- When $\beta_0 + \beta_1 X_1$ is really small, logit function approaches *0*.
- When $\beta_0 + \beta_1 X_1$ is equal to 0, logit function is *0.5*.


Probability
Y = 1
1
0.75
0.5
0.25
0
0
1
2
3
X
$\hat{\beta}_0 = -3$
$\hat{\beta}_1 = 2$
(a)
Probability
Y = 1
$\hat{\beta}_0 = -4$
$\hat{\beta}_1 = 6$
(b)
$\hat{\beta}_0 = -1$
$\hat{\beta}_1 = 1$
(c)
$\hat{\beta}_0 = 3$
$\hat{\beta}_1 = -2$
(d)


Probability of admission
1
0.75
0.5
0.25
0
Probability rises by 0.01 when GPA goes from 95 to 100
Probability rises by 0.30 when GPA goes from 85 to 90
Probability rises by 0.03 when GPA goes from 70 to 75
65
70
75
80
85
90
95
100
GPA
(on 100-point scale)

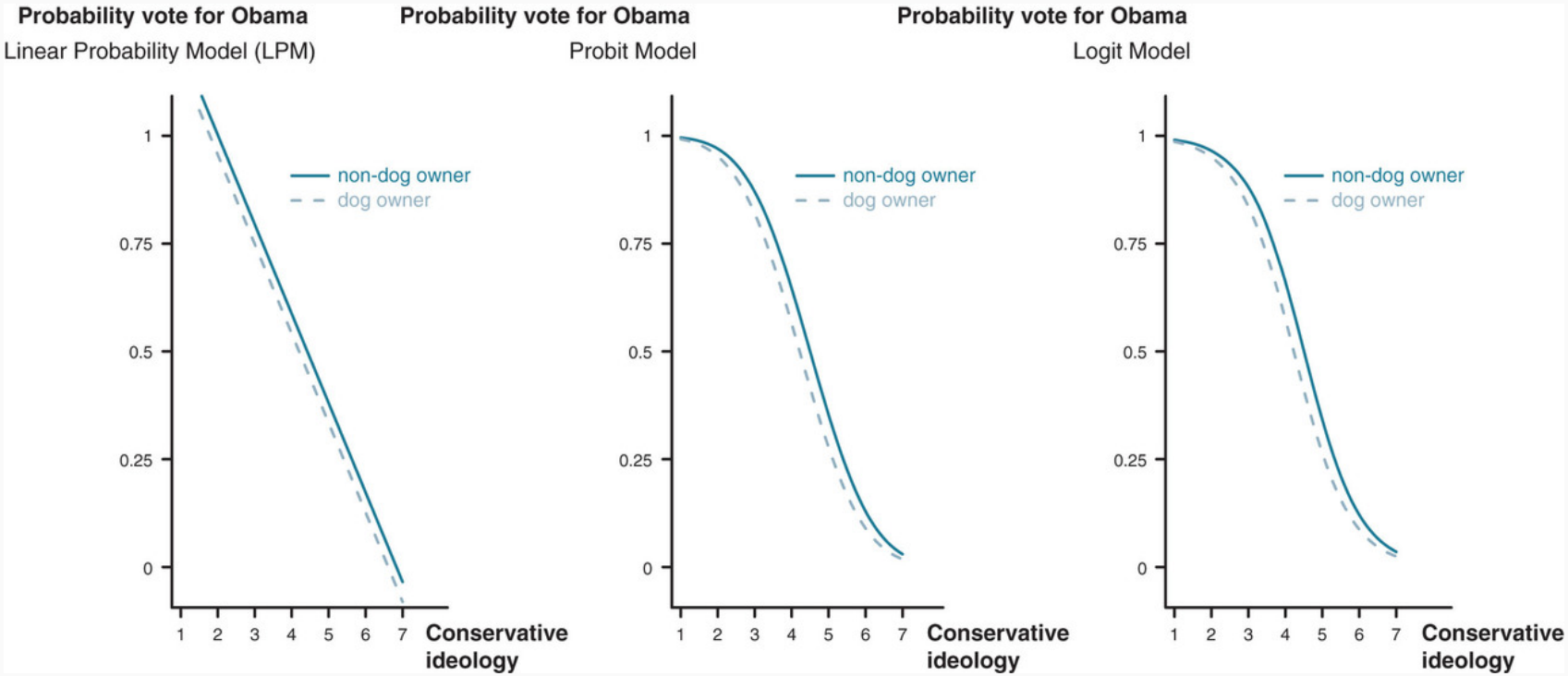

Probability vote for Obama
Linear Probability Model (LPM)
Probability vote for Obama
Probit Model
Probability vote for Obama
Logit Model
non-dog owner
dog owner
1
0.75
0.5
0.25
0
1 2 3 4 5 6 7
Conservative ideology

**TABLE 12.3 Dog Ownership and Probability of Supporting Obama in 2008 Election**

| | LPM | Probit | Logit |
|---|---|---|---|
| Dog owner | −0.050* | −0.213* | −0.368* |
| | (0.006) | (0.025) | (0.044) |
| | [$t = 8.14$] | [$z = 8.53$] | [$z = 8.44$] |
| Ideology | −0.207* | −0.753* | −1.326* |
| | (0.002) | (0.011) | (0.021) |
| | [$t = 106.0$] | [$z = 71.68$] | [$z = 63.98$] |
| Constant | 1.421* | 3.386* | 5.977* |
| | (0.009) | (0.049) | (0.095) |
| | [$t = 159.9$] | [$z = 69.74$] | [$z = 62.84$] |
| $N$ | 15,596 | 15,596 | 15,596 |
| $R^2$ | 0.42 | | |
| $\log L$ | | −6,720.87 | −6,693.98 |
| Minimum $\hat{Y}_i$ | −0.080 | 0.018 | 0.024 |
| Maximum $\hat{Y}_i$ | 1.213 | 0.995 | 0.990 |

*Standard errors in parentheses.*

* *indicates significance at* $p < 0.05$, *two-tailed.*

**TABLE 12.4 Estimated Effect of Dog Ownership and Ideology on Probability of Supporting Obama in 2008 Election**

| Variable | Simulated change | Probit | Logit |
|---|---|---|---|
| Dog owner | From 0 to 1, ideology at actual value | −0.052 | −0.051 |
| Ideology | Increase by 1, dog owner at actual value | −0.180 | −0.181 |

# Dealing with categorical DV's

# Data in ordered categories

Suppose we have data in *categories*, and we are sure of their order.

For example, we have this question from the ANES:

> *There has been some discussion of abortion in recent years. Which of the following best agrees with your view?*

# ANES abortion question

1. By law, abortion should never be permitted.
2. The law should permit abortion only in case of rape, incest, or when the woman's life is in danger.
3. The law should permit abortion for reasons other than rape, incest, or danger to the woman's life, but only after the need for the abortion has been clearly established.
4. By law, a woman should always be able to obtain an abortion as a matter of personal choice.

# How do we model this?

- If we used OLS, we'd be assuming that each category is equally far apart.
- Instead, let's just assume that the range for each option is the same for everyone.
- This is the domain of `ordered logit/probit`.

# Modeling a latent trait

- **Ordered** logit/probit take a **latent variable** approach.
- Latent means an unobserved tendency to have the observed outcome.
- We can't actually observe this underlying trait, but we can observe the outcome.

# Unobserved outcome

In ordered probit/logit, we assume that this unobserved $y^*$ is:

- A function of $X$ and whatever $\beta$ we're estimating.
- Monotonically increasing (preserving order) in the value of the latent variable.

# Unobserved outcome

In ordered probit/logit, we assume that this unobserved $y^*$ is:

- A function of $X$ and whatever $\beta$ we're estimating.
- Monotonically increasing (preserving order) in the value of the latent variable.

  If the highest level of our outcome is **5** ("strongly agree") on a 1-5 scale, then as $y^*$ increases, our observed outcome should *never decrease*.

# Categories as boundaries

Values where the outcome "jumps" are **thresholds** (or boundaries or cutpoints): $\tau$

# Categories as boundaries

Values where the outcome "jumps" are **thresholds** (or boundaries or cutpoints): $\tau$

- Number of thresholds corresponds with the number of categories in our outcome variable ($k$).

$$\tau = k - 1$$
$$\tau = 5 - 1$$
$$\tau = 4$$

# Thresholds in the abortion example

For our abortion question, suppose we want to predict the answer with gender, ideology and income. We are thus saying:

$$y^* = \beta_1 Female + \beta_2 Ideology + \beta_3 Income + \epsilon$$

We don't observe $y^*$, so we model $y$.

We introduce three $\tau$'s to separate our four categories.

# Interpreting non-linear output

# Interpretation

1. What is the quantity of interest?
2. What is the effect on $Y$ of a small change in $X$?
3. What is the prediction of $Y$ given a profile of $X$?
4. What is the difference in $Y$ between two profiles of $X$?

# Parameters in non-linear models

- In logit, the coefficient is a change in the "log-odds" of an outcome.
- Log-odds is not intuitive. You could convert to an odds-ratio. Still not intuitive.
- Can simply look at the sign of the coefficient, as a scan for substantive significance.
- Hard to compare coefficients because the effects of each depends on one others.

# Two main ways

1. **Predicted probabilities**: probability of outcome being $K$ given our $X$'s.
2. **Marginal effects**:discrete change associated with one unit increase in one covariate.

# Predicted probabilities

- Predicted probabilities in R
- Predicted probs in logit and ordered logit
- Discrete marginal effects in R
- Marginal effects in R with cool plots
- Data camp practice course

# Reviewing probability

# An information problem

1. You have a question you want to ask about the world, or even a hunch about how things work, but you don't have complete information.

# An information problem

1. You have a question you want to ask about the world, or even a hunch about how things work, but you don't have complete information.
2. We usually start somewhere, even with incomplete info (with a hunch, or a *prior*.

# An information problem

1. You have a question you want to ask about the world, or even a hunch about how things work, but you don't have complete information.
2. We usually start somewhere, even with incomplete info (with a hunch, or a *prior*.
3. Our goal is to produce a **posterior probability estimate** (probability of observing evidence of your hypothesis given the data).

# Shift from OLS

In the regression context, we're tempted to talk about $\beta$ (or here, $\Theta$) by asking:

> What is the probability that the true $\beta$ is or is not this big?

# Shift from OLS

In the regression context, we're tempted to talk about $\beta$ (or here, $\Theta$) by asking:

> What is the probability that the true $\beta$ is or is not this big?

In reality, we have to say:

> What is the probability, assuming that $\beta = 0$, that we'd observe an estimated $\hat{\beta}$ this far from 0?

# Likely probabilities

Can't say anything about the true probability, but we can say something about relative probabilities.

1. Find the most likely $\Theta$, given the data.
2. Compare different $\Theta$'s, given the data.

# For Thursday

- Class won't meet.
- Online tutorial and assignment instead.
- Homework will be up later today.