GOV 2001/ 1002/ E-200 Section 5
Binary Dependent Variable Regression[1]

Anton Strezhnev

Harvard University

March 2, 2016

[1] These section notes are heavily indebted to past Gov 2001 TFs for slides and R code.

Logistics

Reading Assignment- 4 papers on binary dependent variable models - pay attention particularly to the applications and common pitfalls.

Problem Set 5- Due by 6pm, 3/9 on Canvas.

Assessment Question- Due by 6pm, 3/9 on on Canvas. You must work alone and only one attempt.

OVERVIEW

- In this section you will...

OVERVIEW

- In this section you will...
  - learn why the Fisher information is used to approximate standard errors of MLEs

Overview

- In this section you will...
  - learn why the Fisher information is used to approximate standard errors of MLEs
  - learn about generalized linear models and how logit models fit into that framework

Overview

- In this section you will...
  - learn why the Fisher information is used to approximate standard errors of MLEs
  - learn about generalized linear models and how logit models fit into that framework
  - learn how to evaluate binary dependent variable models.

OVERVIEW

- In this section you will...
  - learn why the Fisher information is used to approximate standard errors of MLEs
  - learn about generalized linear models and how logit models fit into that framework
  - learn how to evaluate binary dependent variable models.

Outline

Asymptotic properties of MLEs

Generalized Linear Models

Logit Model Applied

MLEs in Asymptopia

MLEs in Asymptopia

- For the entirety of this course, we'll be working with the "large sample" properties of Maximum Likelihood Estimators.

MLEs in Asymptopia

- For the entirety of this course, we'll be working with the "large sample" properties of Maximum Likelihood Estimators.
- This is because most MLEs – particularly for generalized linear models – are biased in finite samples.

MLEs in Asymptopia

- For the entirety of this course, we'll be working with the "large sample" properties of Maximum Likelihood Estimators.
- This is because most MLEs – particularly for generalized linear models – are biased in finite samples. But that bias goes away as the sample gets large.

MLEs in Asymptopia

- For the entirety of this course, we'll be working with the "large sample" properties of Maximum Likelihood Estimators.
- This is because most MLEs – particularly for generalized linear models – are biased in finite samples. But that bias goes away as the sample gets large.
- Also, small sample standard errors are hard to calculate analytically – no easy formula like OLS.

MLEs in Asymptopia

- For the entirety of this course, we'll be working with the "large sample" properties of Maximum Likelihood Estimators.
- This is because most MLEs – particularly for generalized linear models – are biased in finite samples. But that bias goes away as the sample gets large.
- Also, small sample standard errors are hard to calculate analytically – no easy formula like OLS.
- **Warning**: Be wary of published GLM results with small samples -

MLEs in Asymptopia

- For the entirety of this course, we'll be working with the "large sample" properties of Maximum Likelihood Estimators.
- This is because most MLEs – particularly for generalized linear models – are biased in finite samples. But that bias goes away as the sample gets large.
- Also, small sample standard errors are hard to calculate analytically – no easy formula like OLS.
- **Warning**: Be wary of published GLM results with small samples - they don't have the same small-sample unbiasedness properties as OLS

Large-sample properties of MLEs

Large-sample properties of MLEs

- Two big properties:

Large-sample properties of MLEs

- Two big properties:
- **Consistency:** $\hat{\theta}_{MLE} \xrightarrow{p} \theta_0$.

Large-sample properties of MLEs

- Two big properties:
- **Consistency:** $\hat{\theta}_{MLE} \xrightarrow{p} \theta_0$. The MLE estimator converges in probability to the true value θ_0 as n gets large.

Large-sample properties of MLEs

- Two big properties:
- **Consistency:** $\hat{\theta}_{MLE} \xrightarrow{p} \theta_0$. The MLE estimator converges in probability to the true value θ_0 as n gets large.
- **Asymptotic Normality:** $\hat{\theta}_{MLE} \sim \text{Normal}(\theta_0, \sigma^2_{MLE})$ in large samples.

Large-sample properties of MLEs

- Two big properties:
- **Consistency:** $\hat{\theta}_{MLE} \xrightarrow{p} \theta_0$. The MLE estimator converges in probability to the true value θ_0 as n gets large.
- **Asymptotic Normality:** $\hat{\theta}_{MLE} \sim \text{Normal}(\theta_0, \sigma^2_{MLE})$ in large samples.
- But how do we calculate σ^2_{MLE} !?

Review: Estimator Variance

Review: Estimator Variance

- Why does the MLE have a variance?

Review: Estimator Variance

- Why does the MLE have a variance?
- Because the data is random!

Review: Estimator Variance

- Why does the MLE have a variance?
- Because the data is random! We have a *stochastic* component in our model that describes how the observations are generated.
- If we drew another sample, or re-ran the experiment on another hypothetical group, we would get a different MLE.

Review: Estimator Variance

- Why does the MLE have a variance?
- Because the data is random! We have a *stochastic* component in our model that describes how the observations are generated.
- If we drew another sample, or re-ran the experiment on another hypothetical group, we would get a different MLE. Imagining repeating this process over and over gives us the theoretical *sampling distribution*.

Review: Estimator Variance

- Why does the MLE have a variance?
- Because the data is random! We have a *stochastic* component in our model that describes how the observations are generated.
- If we drew another sample, or re-ran the experiment on another hypothetical group, we would get a different MLE. Imagining repeating this process over and over gives us the theoretical *sampling distribution*.
- More simply, likelihoods are sums of random variables (e.g. Y_i).

Review: Estimator Variance

- Why does the MLE have a variance?
- Because the data is random! We have a *stochastic* component in our model that describes how the observations are generated.
- If we drew another sample, or re-ran the experiment on another hypothetical group, we would get a different MLE. Imagining repeating this process over and over gives us the theoretical *sampling distribution*.
- More simply, likelihoods are sums of random variables (e.g. Y_i). Therefore functions of them are also random variables!

Asymptotic variance of MLEs

ASYMPTOTIC VARIANCE OF MLES

- When we did OLS, we could get σ^2_{MLE} by just taking $Var(\hat{\theta})$ since there was a closed form solution

Asymptotic variance of MLEs

- When we did OLS, we could get σ^2_{MLE} by just taking $Var(\hat{\theta})$ since there was a closed form solution
- $Var(\hat{\theta}_{OLS}) = Var\left((X'X)^{-1}X'Y\right) = \sigma^2(X'X)^{-1}$

Asymptotic variance of MLEs

- When we did OLS, we could get σ^2_{MLE} by just taking $Var(\hat{\theta})$ since there was a closed form solution
- $Var(\hat{\theta}_{OLS}) = Var\left((X'X)^{-1}X'Y\right) = \sigma^2(X'X)^{-1}$
- But we don't have closed form solutions for almost all MLEs $\hat{\theta}_{MLE}$. What can we use?

THE SCORE

- Remember that the likelihood of θ: $L(\theta|X) = p(X|\theta)$.

The Score

- Remember that the likelihood of θ: $L(\theta|X) = p(X|\theta)$.
- The log-likelihood $\ell(\theta|X) = \ln p(X|\theta)$.

The Score

- Remember that the likelihood of θ: $L(\theta|X) = p(X|\theta)$.
- The log-likelihood $\ell(\theta|X) = \ln p(X|\theta)$.
- The derivative of the log-likelihood is known as the "score" function.

THE SCORE

- Remember that the likelihood of θ: $L(\theta|X) = p(X|\theta)$.
- The log-likelihood $\ell(\theta|X) = \ln p(X|\theta)$.
- The derivative of the log-likelihood is known as the "score" function.
  - $S(\theta) = \frac{\partial}{\partial\theta} \ln p(X|\theta) = \ell'(\theta|X)$

The Score

- Remember that the likelihood of θ: $L(\theta|X) = p(X|\theta)$.
- The log-likelihood $\ell(\theta|X) = \ln p(X|\theta)$.
- The derivative of the log-likelihood is known as the "score" function.
  - $S(\theta) = \frac{\partial}{\partial\theta} \ln p(X|\theta) = \ell'(\theta|X)$
- At the MLE, by definition, the score is 0: $S(\hat{\theta}) = 0$.

The Score

- Remember that the likelihood of θ: $L(\theta|X) = p(X|\theta)$.
- The log-likelihood $\ell(\theta|X) = \ln p(X|\theta)$.
- The derivative of the log-likelihood is known as the "score" function.
  - $S(\theta) = \frac{\partial}{\partial\theta} \ln p(X|\theta) = \ell'(\theta|X)$
- At the MLE, by definition, the score is 0: $S(\hat{\theta}) = 0$.
- We can also show that the expectation of the score function is also 0

$$E[S(\theta)] = 0$$

THE INFORMATION

- Where does the "information matrix" we talk about come from?

The Information

- Where does the "information matrix" we talk about come from? Well, it's the variance of the score.

$$
\begin{aligned}
I(\theta) &= Var[S(\theta)] \\
&= E[S(\theta)^2] - E[S(\theta)]^2
\end{aligned}
$$

The Information

- Where does the "information matrix" we talk about come from? Well, it's the variance of the score.

$$\begin{aligned} I(\theta) &= Var[S(\theta)] \\ &= E[S(\theta)^2] - E[S(\theta)]^2 \end{aligned}$$

From before, the second term is 0. So

$$I(\theta) = E[S(\theta)^2]$$

The Information

- Where does the "information matrix" we talk about come from? Well, it's the variance of the score.

$$\begin{aligned} I(\theta) &= Var[S(\theta)] \\ &= E[S(\theta)^2] - E[S(\theta)]^2 \end{aligned}$$

From before, the second term is 0. So

$$I(\theta) = E[S(\theta)^2]$$

It turns out that we can also show that it equals the expectation of the negative of the Hessian of the likelihood.

The Information

- Where does the "information matrix" we talk about come from? Well, it's the variance of the score.

$$\begin{aligned} I(\theta) &= Var[S(\theta)] \\ &= E[S(\theta)^2] - E[S(\theta)]^2 \end{aligned}$$

From before, the second term is 0. So

$$I(\theta) = E[S(\theta)^2]$$

It turns out that we can also show that it equals the expectation of the negative of the Hessian of the likelihood.

$$I(\theta) = E[S(\theta)^2] = E[-\ell''(\theta)]$$

POWER OF THE I.I.D. ASSUMPTION

Power of the i.i.d. assumption

- A lot of intuition for asymptotics comes from our i.i.d. assumption. Under this assumption, our log-likelihoods are sums of n separate log-likelihoods for each observation.

POWER OF THE I.I.D. ASSUMPTION

- A lot of intuition for asymptotics comes from our i.i.d. assumption. Under this assumption, our log-likelihoods are sums of n separate log-likelihoods for each observation.
- So $\ell(\theta|X) = \sum_{i=1}^{n} \ell(\theta|X_i)$ under i.i.d. observations.

Power of the i.i.d. assumption

- A lot of intuition for asymptotics comes from our i.i.d. assumption. Under this assumption, our log-likelihoods are sums of n separate log-likelihoods for each observation.
- So $\ell(\theta|X) = \sum_{i=1}^{n} \ell(\theta|X_i)$ under i.i.d. observations.
- Likewise, $S(\theta) = \ell'(\theta|X) = \sum_{i=1}^{n} \ell'(\theta|X_i)$

Power of the i.i.d. assumption

- A lot of intuition for asymptotics comes from our i.i.d. assumption. Under this assumption, our log-likelihoods are sums of n separate log-likelihoods for each observation.
- So $\ell(\theta|X) = \sum_{i=1}^{n} \ell(\theta|X_i)$ under i.i.d. observations.
- Likewise, $S(\theta) = \ell'(\theta|X) = \sum_{i=1}^{n} \ell'(\theta|X_i)$
- Under the i.i.d. assumption, the information also grows with n, so we often denote for i.i.d. observations

$$I_n(\theta) = E\left[-\sum_{i=1}^{n} \ell''(\theta|X_i)\right] = -nE\left[\ell''(\theta|X_1)\right]$$

Power of the i.i.d. assumption

- A lot of intuition for asymptotics comes from our i.i.d. assumption. Under this assumption, our log-likelihoods are sums of n separate log-likelihoods for each observation.
- So $\ell(\theta|X) = \sum_{i=1}^{n} \ell(\theta|X_i)$ under i.i.d. observations.
- Likewise, $S(\theta) = \ell'(\theta|X) = \sum_{i=1}^{n} \ell'(\theta|X_i)$
- Under the i.i.d. assumption, the information also grows with n, so we often denote for i.i.d. observations

$$I_n(\theta) = E\left[-\sum_{i=1}^{n} \ell''(\theta|X_i)\right] = -nE\left[\ell''(\theta|X_1)\right]$$

- **Key takeaway:** Under the i.i.d. assumption, likelihoods are sums of *i.i.d.* random variables.

Power of the i.i.d. assumption

- A lot of intuition for asymptotics comes from our i.i.d. assumption. Under this assumption, our log-likelihoods are sums of n separate log-likelihoods for each observation.
- So $\ell(\theta|X) = \sum_{i=1}^{n} \ell(\theta|X_i)$ under i.i.d. observations.
- Likewise, $S(\theta) = \ell'(\theta|X) = \sum_{i=1}^{n} \ell'(\theta|X_i)$
- Under the i.i.d. assumption, the information also grows with n, so we often denote for i.i.d. observations

$$I_n(\theta) = E\left[-\sum_{i=1}^{n} \ell''(\theta|X_i)\right] = -nE\left[\ell''(\theta|X_1)\right]$$

- **Key takeaway:** Under the i.i.d. assumption, likelihoods are sums of *i.i.d.* random variables. This lets us invoke the Law of Large Numbers and Central Limit Theorem.

POWER OF THE I.I.D. ASSUMPTION

- A lot of intuition for asymptotics comes from our i.i.d. assumption. Under this assumption, our log-likelihoods are sums of n separate log-likelihoods for each observation.
- So $\ell(\theta|X) = \sum_{i=1}^{n} \ell(\theta|X_i)$ under i.i.d. observations.
- Likewise, $S(\theta) = \ell'(\theta|X) = \sum_{i=1}^{n} \ell'(\theta|X_i)$
- Under the i.i.d. assumption, the information also grows with n, so we often denote for i.i.d. observations

$$I_n(\theta) = E\left[-\sum_{i=1}^{n} \ell''(\theta|X_i)\right] = -nE\left[\ell''(\theta|X_1)\right]$$

- **Key takeaway:** Under the i.i.d. assumption, likelihoods are sums of *i.i.d.* random variables. This lets us invoke the Law of Large Numbers and Central Limit Theorem.

Connecting information to variance

Connecting information to variance

- Let's start with the quadratic Taylor approximation of the likelihood around the true value θ_0.

Connecting information to variance

- Let's start with the quadratic Taylor approximation of the likelihood around the true value θ_0.

$$\ell(\theta) \approx \ell(\theta_0) + \ell'(\theta_0)(\theta - \theta_0) + \frac{1}{2}\ell''(\theta_0)(\theta - \theta_0)^2$$

Connecting information to variance

- Let's start with the quadratic Taylor approximation of the likelihood around the true value θ_0.

$$\ell(\theta) \approx \ell(\theta_0) + \ell'(\theta_0)(\theta - \theta_0) + \frac{1}{2}\ell''(\theta_0)(\theta - \theta_0)^2$$

- Take the derivative to get the score

$$\ell'(\theta) = \ell'(\theta_0) + \ell''(\theta_0)(\theta - \theta_0)$$

CONNECTING INFORMATION TO VARIANCE

- Let's start with the quadratic Taylor approximation of the likelihood around the true value θ_0.

$$\ell(\theta) \approx \ell(\theta_0) + \ell'(\theta_0)(\theta - \theta_0) + \frac{1}{2}\ell''(\theta_0)(\theta - \theta_0)^2$$

- Take the derivative to get the score

$$\ell'(\theta) = \ell'(\theta_0) + \ell''(\theta_0)(\theta - \theta_0)$$

- At the MLE, $\hat{\theta}$, the score is 0, so we can write

Connecting information to variance

- Let's start with the quadratic Taylor approximation of the likelihood around the true value θ_0.

$$\ell(\theta) \approx \ell(\theta_0) + \ell'(\theta_0)(\theta - \theta_0) + \frac{1}{2}\ell''(\theta_0)(\theta - \theta_0)^2$$

- Take the derivative to get the score

$$\ell'(\theta) = \ell'(\theta_0) + \ell''(\theta_0)(\theta - \theta_0)$$

- At the MLE, $\hat{\theta}$, the score is 0, so we can write

$$0 = \ell'(\theta_0) + \ell''(\theta_0)(\hat{\theta} - \theta_0)$$

Connecting information to variance

- Re-arranging terms

$$(\hat{\theta} - \theta_0) = \frac{\ell'(\theta_0)}{-\ell''(\theta_0)}$$

Connecting information to variance

- Re-arranging terms

$$(\hat{\theta} - \theta_0) = \frac{\ell'(\theta_0)}{-\ell''(\theta_0)}$$

- Multiply both sides by $\sqrt{n}$

$$\sqrt{n}(\hat{\theta} - \theta_0) = \frac{\frac{1}{\sqrt{n}}\ell'(\theta_0)}{-\frac{1}{n}\ell''(\theta_0)}$$

CONNECTING INFORMATION TO VARIANCE

- Re-arranging terms

$$(\hat{\theta} - \theta_0) = \frac{\ell'(\theta_0)}{-\ell''(\theta_0)}$$

- Multiply both sides by $\sqrt{n}$

$$\sqrt{n}(\hat{\theta} - \theta_0) = \frac{\frac{1}{\sqrt{n}}\ell'(\theta_0)}{-\frac{1}{n}\ell''(\theta_0)}$$

- Now we make use of two convergence rules. Recall that both the score and the information are sums of i.i.d. random variables

Connecting information to variance

- First, by the Law of Large Numbers, the denominator converges to the information for a single observation

$$-\frac{1}{n}\ell''(\theta_0) = -\frac{1}{n}\sum_{i=1}^{n}\ell''(\theta_0|X_i) \xrightarrow{p} -E[\ell''(\theta_0|X_i)] = I(\theta)$$

Connecting information to variance

- First, by the Law of Large Numbers, the denominator converges to the information for a single observation

$$-\frac{1}{n}\ell''(\theta_0) = -\frac{1}{n}\sum_{i=1}^{n}\ell''(\theta_0|X_i) \xrightarrow{p} -E[\ell''(\theta_0|X_i)] = I(\theta)$$

- Second, by the Central Limit Theorem, the numerator converges in distribution to a normal distribution with mean $E[S(\theta|X_i)] = 0$ and variance $Var(S(\theta|X_i)] = I(\theta)$

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\ell'(\theta|X_i) \xrightarrow{d} \text{Normal}(0, I(\theta))$$

Connecting information to variance

- First, by the Law of Large Numbers, the denominator converges to the information for a single observation

$$-\frac{1}{n}\ell''(\theta_0) = -\frac{1}{n}\sum_{i=1}^{n}\ell''(\theta_0|X_i) \xrightarrow{p} -E[\ell''(\theta_0|X_i)] = I(\theta)$$

- Second, by the Central Limit Theorem, the numerator converges in distribution to a normal distribution with mean $E[S(\theta|X_i)] = 0$ and variance $Var(S(\theta|X_i)] = I(\theta)$

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\ell'(\theta|X_i) \xrightarrow{d} \text{Normal}(0, I(\theta))$$

Connecting information to variance

- By Slutsky's theorem, therefore, we have

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \text{Normal}\left(0, \frac{1}{I(\theta)}\right)$$

Connecting information to variance

- By Slutsky's theorem, therefore, we have

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \text{Normal}\left(0, \frac{1}{I(\theta)}\right)$$

- And by extension

$$\hat{\theta} \xrightarrow{d} \text{Normal}\left(\theta_0, \frac{1}{I_n(\theta)}\right)$$

Connecting information to variance

- By Slutsky's theorem, therefore, we have

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \text{Normal}\left(0, \frac{1}{I(\theta)}\right)$$

- And by extension

$$\hat{\theta} \xrightarrow{d} \text{Normal}\left(\theta_0, \frac{1}{I_n(\theta)}\right)$$

- Finally, $I_n(\theta)$ is an expectation that's hard to calculate.

Connecting information to variance

- By Slutsky's theorem, therefore, we have

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \text{Normal}\left(0, \frac{1}{I(\theta)}\right)$$

- And by extension

$$\hat{\theta} \xrightarrow{d} \text{Normal}\left(\theta_0, \frac{1}{I_n(\theta)}\right)$$

- Finally, $I_n(\theta)$ is an expectation that's hard to calculate. Instead, we plug in a consistent estimate – the observed information $I_n(\hat{\theta})$. This is the negative Hessian evaluated at the MLE.

Connecting information to variance

- By Slutsky's theorem, therefore, we have

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \text{Normal}\left(0, \frac{1}{I(\theta)}\right)$$

- And by extension

$$\hat{\theta} \xrightarrow{d} \text{Normal}\left(\theta_0, \frac{1}{I_n(\theta)}\right)$$

- Finally, $I_n(\theta)$ is an expectation that's hard to calculate. Instead, we plug in a consistent estimate – the observed information $I_n\hat{(\theta)}$. This is the negative Hessian evaluated at the MLE.

Regularity conditions in practice

- We've been handwaving a lot about what makes a likelihood "nice" for the asymptotics to apply.

Regularity conditions in practice

- We've been handwaving a lot about what makes a likelihood "nice" for the asymptotics to apply. But what does this mean in practice?

REGULARITY CONDITIONS IN PRACTICE

- We've been handwaving a lot about what makes a likelihood "nice" for the asymptotics to apply. But what does this mean in practice?
- Three big "regularity" conditions for convergence:

REGULARITY CONDITIONS IN PRACTICE

- We've been handwaving a lot about what makes a likelihood "nice" for the asymptotics to apply. But what does this mean in practice?
- Three big "regularity" conditions for convergence:
- **Identifiability** - If $L(\hat{\theta}) = L(\theta)$, then $\hat{\theta} = \theta$. That is, there is a single value of θ that maximizes the likelihood. One place where this doesn't hold is when there are more parameters than data points.

REGULARITY CONDITIONS IN PRACTICE

- We've been handwaving a lot about what makes a likelihood "nice" for the asymptotics to apply. But what does this mean in practice?
- Three big "regularity" conditions for convergence:
- **Identifiability** - If $L(\hat{\theta}) = L(\theta)$, then $\hat{\theta} = \theta$. That is, there is a single value of θ that maximizes the likelihood. One place where this doesn't hold is when there are more parameters than data points.
- **I.i.d. observations** - The likelihood can be factored into n identical and independent densities – that is, $L(\theta|X) = \prod_{i=1}^{n} L(\theta|X_i)$

Regularity conditions in practice

- We've been handwaving a lot about what makes a likelihood "nice" for the asymptotics to apply. But what does this mean in practice?
- Three big "regularity" conditions for convergence:
- **Identifiability** - If $L(\hat{\theta}) = L(\theta)$, then $\hat{\theta} = \theta$. That is, there is a single value of θ that maximizes the likelihood. One place where this doesn't hold is when there are more parameters than data points.
- **I.i.d. observations** - The likelihood can be factored into n identical and independent densities – that is, $L(\theta|X) = \prod_{i=1}^{n} L(\theta|X_i)$
- **Parameter space fixed relative to n** - As sample size increases, the number of parameters being estimated *doesn't* grow with it.

Regularity conditions in practice

- We've been handwaving a lot about what makes a likelihood "nice" for the asymptotics to apply. But what does this mean in practice?
- Three big "regularity" conditions for convergence:
- **Identifiability** - If $L(\hat{\theta}) = L(\theta)$, then $\hat{\theta} = \theta$. That is, there is a single value of θ that maximizes the likelihood. One place where this doesn't hold is when there are more parameters than data points.
- **I.i.d. observations** - The likelihood can be factored into n identical and independent densities – that is, $L(\theta|X) = \prod_{i=1}^{n} L(\theta|X_i)$
- **Parameter space fixed relative to** n - As sample size increases, the number of parameters being estimated *doesn't* grow with it.

OUTLINE

Asymptotic properties of MLEs

Generalized Linear Models

Logit Model Applied

Generalized Linear Models

- Most models that we work with in this course are part of a class of models called *generalized linear models (GLM)*.

Generalized Linear Models

- Most models that we work with in this course are part of a class of models called *generalized linear models (GLM)*.
- Takes the "linear" component $X\beta$ from OLS and allows it to model outcomes with different types of distributions.

Generalized Linear Models

- Most models that we work with in this course are part of a class of models called *generalized linear models (GLM)*.
- Takes the "linear" component $X\beta$ from OLS and allows it to model outcomes with different types of distributions.
- Three components:
  - A distribution for Y (stochastic component)
  - A linear predictor for $X\beta$ (systematic component)
  - A link function that connects the linear predictor to parameters of the distribution on Y

1. Pick a distribution for Y

- We start by assuming our data comes from some distribution.
- Examples:
  - Continuous and Unbounded: **Normal**(μ, σ^2)

1. Pick a distribution for Y

- We start by assuming our data comes from some distribution.
- Examples:
  - Continuous and Unbounded: **Normal**(μ, σ^2)
  - Binary: **Bernoulli**(π)

1. Pick a distribution for Y

- We start by assuming our data comes from some distribution.
- Examples:
  - Continuous and Unbounded: **Normal**(μ, σ^2)
  - Binary: **Bernoulli**(π)
  - Event Count: **Poisson**(λ),

1. Pick a distribution for Y

- We start by assuming our data comes from some distribution.
- Examples:
  - Continuous and Unbounded: **Normal**(μ, σ^2)
  - Binary: **Bernoulli**(π)
  - Event Count: **Poisson**(λ), **Negative Binomial**(r, p)

1. Pick a distribution for Y

- We start by assuming our data comes from some distribution.
- Examples:
  - Continuous and Unbounded: **Normal**(μ, σ^2)
  - Binary: **Bernoulli**(π)
  - Event Count: **Poisson**(λ), **Negative Binomial**(r, p)
  - Duration: **Exponential**(λ),

1. Pick a distribution for Y

- We start by assuming our data comes from some distribution.
- Examples:
  - Continuous and Unbounded: **Normal**(μ, σ^2)
  - Binary: **Bernoulli**(π)
  - Event Count: **Poisson**(λ), **Negative Binomial**(r, p)
  - Duration: **Exponential**(λ), **Weibull**(λ, k)

1. Pick a distribution for Y

- We start by assuming our data comes from some distribution.
- Examples:
  - Continuous and Unbounded: **Normal**(μ, σ^2)
  - Binary: **Bernoulli**(π)
  - Event Count: **Poisson**(λ), **Negative Binomial**(r, p)
  - Duration: **Exponential**(λ), **Weibull**(λ, k)
  - Unordered Categories: **Multinomial**(π)

1. Pick a distribution for Y

- We start by assuming our data comes from some distribution.
- Examples:
  - Continuous and Unbounded: **Normal**(μ, σ^2)
  - Binary: **Bernoulli**(π)
  - Event Count: **Poisson**(λ), **Negative Binomial**(r, p)
  - Duration: **Exponential**(λ), **Weibull**(λ, k)
  - Unordered Categories: **Multinomial**(π)
- Sometimes, instead of directly putting a distribution on Y, we can put a distribution on an unobserved "latent" variable Y^* and treat Y as a function of Y^*

1. Pick a distribution for Y

- We start by assuming our data comes from some distribution.
- Examples:
  - Continuous and Unbounded: **Normal**(μ, σ^2)
  - Binary: **Bernoulli**(π)
  - Event Count: **Poisson**(λ), **Negative Binomial**(r, p)
  - Duration: **Exponential**(λ), **Weibull**(λ, k)
  - Unordered Categories: **Multinomial**(π)
- Sometimes, instead of directly putting a distribution on Y, we can put a distribution on an unobserved "latent" variable Y^* and treat Y as a function of Y^* – e.g. for ordered categorical data, Y^* is unbounded, but Y is a piece-wise function of Y^*.

2. Specify a linear predictor $\eta = X\beta$

- Our covariates X enter into a GLM in a very specific way.

2. Specify a linear predictor $\eta = X\beta$

- Our covariates X enter into a GLM in a very specific way. As in OLS, the linear predictor is a linear combination of the parameters β.

2. Specify a linear predictor $\eta = X\beta$

- Our covariates X enter into a GLM in a very specific way. As in OLS, the linear predictor is a linear combination of the parameters β.
- If we have k covariates, our linear predictor $\eta = X\beta$ is:

$$\eta = X\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + x_k \beta_k$$

2. Specify a linear predictor $\eta = X\beta$

- Our covariates X enter into a GLM in a very specific way. As in OLS, the linear predictor is a linear combination of the parameters β.
- If we have k covariates, our linear predictor $\eta = X\beta$ is:

$$\eta = X\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + x_k \beta_k$$

- Just like OLS, we can have (non-linear) functions of X as covariates (e.g. X^2), but our *parameters* are a linear combination.

3. Specify a link function $g(\mu) = X\beta$

- Finally, we need to connect the linear predictor to the mean μ of the distribution on Y.

3. Specify a link function $g(\mu) = X\beta$

- Finally, we need to connect the linear predictor to the mean μ of the distribution on Y. Often this will be a parameter of that distribution.

3. Specify a link function $g(\mu) = X\beta$

- Finally, we need to connect the linear predictor to the mean μ of the distribution on Y. Often this will be a parameter of that distribution.
- Lots of choices – we need the domain of the link to match the range of the mean.
- We pick a $g(\dot{})$ and set $g(\mu) = X\beta$
- Then solve back to get the inverse link $\mu = g^{-1}(X\beta)$.

Link function example: Logit

- If $Y_i \sim \text{Bernoulli}(\pi_i)$, then $E[Y_i] = \pi_i$.

Link function example: Logit

- If $Y_i \sim \text{Bernoulli}(\pi_i)$, then $E[Y_i] = \pi_i$.
- $\pi_i \in (0, 1)$, so we need a function that maps from $(0, 1)$ to $(-\infty, \infty)$.

Link function example: Logit

- If $Y_i \sim \text{Bernoulli}(\pi_i)$, then $E[Y_i] = \pi_i$.
- $\pi_i \in (0, 1)$, so we need a function that maps from $(0, 1)$ to $(-\infty, \infty)$.
- One function is the "logit" $\text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$.

Link function example: Logit

- If $Y_i \sim \text{Bernoulli}(\pi_i)$, then $E[Y_i] = \pi_i$.
- $\pi_i \in (0, 1)$, so we need a function that maps from $(0, 1)$ to $(-\infty, \infty)$.
- One function is the "logit" $\text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$.
- So we set $\ln\left(\frac{\pi_i}{1-\pi_i}\right) = X_i\beta$.

Link function example: Logit

- If $Y_i \sim \text{Bernoulli}(\pi_i)$, then $E[Y_i] = \pi_i$.
- $\pi_i \in (0, 1)$, so we need a function that maps from $(0, 1)$ to $(-\infty, \infty)$.
- One function is the "logit" $\text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$.
- So we set $\ln\left(\frac{\pi_i}{1-\pi_i}\right) = X_i\beta$. And take the inverse to solve for π_i.

LINK FUNCTION EXAMPLE: LOGIT

- If $Y_i \sim \text{Bernoulli}(\pi_i)$, then $E[Y_i] = \pi_i$.
- $\pi_i \in (0, 1)$, so we need a function that maps from $(0, 1)$ to $(-\infty, \infty)$.
- One function is the "logit" $\text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$.
- So we set $\ln\left(\frac{\pi_i}{1-\pi_i}\right) = X_i\beta$. And take the inverse to solve for π_i.

$$\ln\left(\frac{1-\pi_i}{\pi_i}\right) = -X_i\beta$$

Link function example: Logit

- If $Y_i \sim \text{Bernoulli}(\pi_i)$, then $E[Y_i] = \pi_i$.
- $\pi_i \in (0,1)$, so we need a function that maps from $(0,1)$ to $(-\infty, \infty)$.
- One function is the "logit" $\text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$.
- So we set $\ln\left(\frac{\pi_i}{1-\pi_i}\right) = X_i\beta$. And take the inverse to solve for π_i.

$$\ln\left(\frac{1-\pi_i}{\pi_i}\right) = -X_i\beta$$

$$\frac{1-\pi_i}{\pi_i} = \exp(-X_i\beta)$$

LINK FUNCTION EXAMPLE: LOGIT

- If $Y_i \sim \text{Bernoulli}(\pi_i)$, then $E[Y_i] = \pi_i$.
- $\pi_i \in (0, 1)$, so we need a function that maps from $(0, 1)$ to $(-\infty, \infty)$.
- One function is the "logit" $\text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$.
- So we set $\ln\left(\frac{\pi_i}{1-\pi_i}\right) = X_i\beta$. And take the inverse to solve for π_i.

$$\ln\left(\frac{1-\pi_i}{\pi_i}\right) = -X_i\beta$$
$$\frac{1-\pi_i}{\pi_i} = \exp(-X_i\beta)$$
$$\frac{1}{\pi_i} - 1 = \exp(-X_i\beta)$$

Link function example: Logit

- If $Y_i \sim \text{Bernoulli}(\pi_i)$, then $E[Y_i] = \pi_i$.
- $\pi_i \in (0, 1)$, so we need a function that maps from $(0, 1)$ to $(-\infty, \infty)$.
- One function is the "logit" $\text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$.
- So we set $\ln\left(\frac{\pi_i}{1-\pi_i}\right) = X_i\beta$. And take the inverse to solve for π_i.

$$\ln\left(\frac{1-\pi_i}{\pi_i}\right) = -X_i\beta$$

$$\frac{1-\pi_i}{\pi_i} = \exp(-X_i\beta)$$

$$\frac{1}{\pi_i} - 1 = \exp(-X_i\beta)$$

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

Link function example: Logit

- If $Y_i \sim \text{Bernoulli}(\pi_i)$, then $E[Y_i] = \pi_i$.
- $\pi_i \in (0, 1)$, so we need a function that maps from $(0, 1)$ to $(-\infty, \infty)$.
- One function is the "logit" $\text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$.
- So we set $\ln\left(\frac{\pi_i}{1-\pi_i}\right) = X_i\beta$. And take the inverse to solve for π_i.

$$\begin{aligned}
\ln\left(\frac{1-\pi_i}{\pi_i}\right) &= -X_i\beta \\
\frac{1-\pi_i}{\pi_i} &= \exp(-X_i\beta) \\
\frac{1}{\pi_i} - 1 &= \exp(-X_i\beta) \\
\frac{1}{\pi_i} &= 1 + \exp(-X_i\beta) \\
\pi_i &= \frac{1}{1+\exp(-X_i\beta)}
\end{aligned}$$

Link function example: Logit

- If $Y_i \sim \text{Bernoulli}(\pi_i)$, then $E[Y_i] = \pi_i$.
- $\pi_i \in (0,1)$, so we need a function that maps from $(0,1)$ to $(-\infty, \infty)$.
- One function is the "logit" $\text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$.
- So we set $\ln\left(\frac{\pi_i}{1-\pi_i}\right) = X_i\beta$. And take the inverse to solve for π_i.

$$\begin{aligned}
\ln\left(\frac{1-\pi_i}{\pi_i}\right) &= -X_i\beta \\
\frac{1-\pi_i}{\pi_i} &= \exp(-X_i\beta) \\
\frac{1}{\pi_i} - 1 &= \exp(-X_i\beta) \\
\frac{1}{\pi_i} &= 1 + \exp(-X_i\beta) \\
\pi_i &= \frac{1}{1+\exp(-X_i\beta)}
\end{aligned}$$

Visualizing the inverse logit

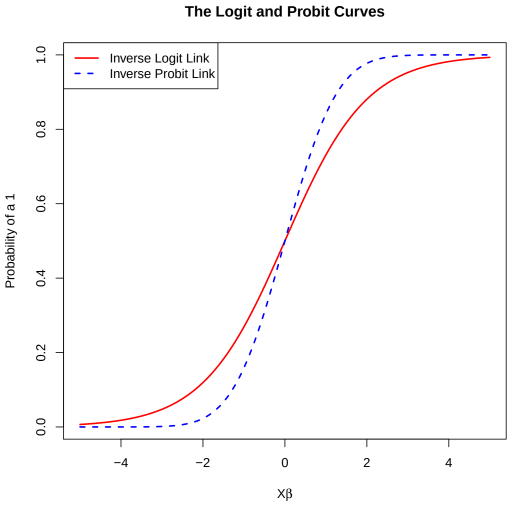

Figure : Comparison of inverse-logit and inverse-probit links

Outline

Asymptotic properties of MLEs

Generalized Linear Models

Logit Model Applied

Forecasting Congressional Elections

Suppose we want to forecast whether or not the incumbent party will win the U.S. House general election.

Forecasting - Define the Model

- First, let's define a model.

Forecasting - Define the Model

- First, let's define a model. We observe n observations each with outcome Y_i and covariates X_i.

Forecasting - Define the Model

- First, let's define a model. We observe n observations each with outcome Y_i and covariates X_i.
- Our distribution on the data is simple:

$$Y_i \sim \text{Bernoulli}(\pi_i)$$

Forecasting - Define the Model

- First, let's define a model. We observe n observations each with outcome Y_i and covariates X_i.
- Our distribution on the data is simple:

$$Y_i \sim \text{Bernoulli}(\pi_i)$$

- Our linear predictor is a function of covariates.

Forecasting - Define the Model

- First, let's define a model. We observe n observations each with outcome Y_i and covariates X_i.
- Our distribution on the data is simple:

$$Y_i \sim \text{Bernoulli}(\pi_i)$$

- Our linear predictor is a function of covariates. In this case, we observe three: whether the seat is open: *open*;

Forecasting - Define the Model

- First, let's define a model. We observe n observations each with outcome Y_i and covariates X_i.
- Our distribution on the data is simple:

$$Y_i \sim \text{Bernoulli}(\pi_i)$$

- Our linear predictor is a function of covariates. In this case, we observe three: whether the seat is open: *open*; whether the incumbent is a freshman: *fresh*;

FORECASTING - DEFINE THE MODEL

- First, let's define a model. We observe n observations each with outcome Y_i and covariates X_i.
- Our distribution on the data is simple:

$$Y_i \sim \text{Bernoulli}(\pi_i)$$

- Our linear predictor is a function of covariates. In this case, we observe three: whether the seat is open: *open*; whether the incumbent is a freshman: *fresh*; and the vote share of the incumbent party in the district in the last presidential election: *incshare*.

Forecasting - Define the Model

- First, let's define a model. We observe n observations each with outcome Y_i and covariates X_i.
- Our distribution on the data is simple:

$$Y_i \sim \text{Bernoulli}(\pi_i)$$

- Our linear predictor is a function of covariates. In this case, we observe three: whether the seat is open: *open*; whether the incumbent is a freshman: *fresh*; and the vote share of the incumbent party in the district in the last presidential election: *incshare*.

Forecasting - Define the Model

- So our linear predictor η_i is

$$\eta_i = X_i\beta = \beta_0 + \beta_1 X_{i,open} + \beta_2 X_{i,fresh} + \beta_3 X_{i,incshare}$$

- Finally, we pick a link function.

Forecasting - Define the Model

- So our linear predictor η_i is

$$\eta_i = X_i\beta = \beta_0 + \beta_1 X_{i,open} + \beta_2 X_{i,fresh} + \beta_3 X_{i,incshare}$$

- Finally, we pick a link function. For simplicity, we'll pick the logit link, which yields

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

Forecasting - Estimate the Model

- In class, we derived the log-likelihood of this model.

Forecasting - Estimate the Model

- In class, we derived the log-likelihood of this model. Which we can maximize numerically in R.

FORECASTING - ESTIMATE THE MODEL

- In class, we derived the log-likelihood of this model. Which we can maximize numerically in R.

```
## Load election results from 04-08
votes <- read.dta("votes0408.dta")

## Our log-likelihood function, logit.ll takes three arguments:
## par: the parameters
## outcome: the Y variable
## covariates: the X matrix (including an intercept column

## Create the X matrix
design.matrix <- as.matrix(cbind(1,votes[,c("open","freshman","incpres")]))
## Estimate the MLE:
opt <- optim(par = rep(0, ncol(votes[,2:4]) + 1),
             fn = logit.ll,
             covariates = design.matrix,
             outcome = votes$incwin,
             control = list(fnscale = -1),
             hessian = T,
             method = "BFGS")
```

Forecasting - Estimates for 2004-2008

- Our coefficient estimates are

```
coefs <- opt$par # Beta values for the intercept and 3 coefficients
names(coefs) <- c("Intercept", "open","freshman","incpres")
coefs
 Intercept       open   freshman    incpres
-2.9064379 -2.1266744 -0.3568115  0.1112137
```

Forecasting - Estimates for 2004-2008

- Our coefficient estimates are

```
coefs <- opt$par # Beta values for the intercept and 3 coefficients
names(coefs) <- c("Intercept", "open","freshman","incpres")
coefs
 Intercept       open   freshman    incpres
-2.9064379 -2.1266744 -0.3568115  0.1112137
```

- And our estimates of the standard errors are

```
fisher_info <- -opt$hessian
vcov <- solve(fisher_info)
se <- sqrt(diag(vcov))
names(se) <-  c("Intercept", "open","freshman","incpres")
se
 Intercept       open   freshman    incpres
0.85244799 0.32525364 0.39999344 0.01641946
```

Forecasting - Quantities of Interest

- By themselves, the coefficients are hard to interpret (log-odds ratios).

Forecasting - Quantities of Interest

- By themselves, the coefficients are hard to interpret (log-odds ratios). We want to obtain more informative quantities.

Forecasting - Quantities of Interest

- By themselves, the coefficients are hard to interpret (log-odds ratios). We want to obtain more informative quantities. One intuitive quantity is a predicted probability $\hat{\pi}_i$ for some set of covariates X_i.
- By MLE invariance,

$$\hat{\pi}_i = \frac{1}{1 + \exp(-X_i \hat{\beta})}$$

- So applying the inverse-logit to our linear predictor us an MLE estimate of $\hat{\pi}_i$!

Forecasting - Quantities of Interest

- By themselves, the coefficients are hard to interpret (log-odds ratios). We want to obtain more informative quantities. One intuitive quantity is a predicted probability $\hat{\pi}_i$ for some set of covariates X_i.
- By MLE invariance,

$$\hat{\pi}_i = \frac{1}{1 + \exp(-X_i\hat{\beta})}$$

- So applying the inverse-logit to our linear predictor us an MLE estimate of $\hat{\pi}_i$! This is how we get “fitted values” for a logit model.

Forecasting - Quantities of Interest

- By themselves, the coefficients are hard to interpret (log-odds ratios). We want to obtain more informative quantities. One intuitive quantity is a predicted probability $\hat{\pi}_i$ for some set of covariates X_i.
- By MLE invariance,

$$\hat{\pi}_i = \frac{1}{1 + \exp(-X_i \hat{\beta})}$$

- So applying the inverse-logit to our linear predictor us an MLE estimate of $\hat{\pi}_i$! This is how we get "fitted values" for a logit model. It's also how we make predictions for *new* or hypothetical observations of X_i.

Forecasting - Validation

- How well does our model explain our data?

Forecasting - Validation

- How well does our model explain our data? Couple of ways of evaluating this in the binary D.V. context?

Forecasting - Validation

- How well does our model explain our data? Couple of ways of evaluating this in the binary D.V. context? The Greenhill et. al. reading for this week gives a general overview.

Forecasting - Validation

- How well does our model explain our data? Couple of ways of evaluating this in the binary D.V. context? The Greenhill et. al. reading for this week gives a general overview.
- People often look at accuracy – given some "cut-off" probability, how many cases are correctly predicted by the model.

FORECASTING - VALIDATION

- How well does our model explain our data? Couple of ways of evaluating this in the binary D.V. context? The Greenhill et. al. reading for this week gives a general overview.
- People often look at accuracy – given some "cut-off" probability, how many cases are correctly predicted by the model. Accuracy can be misleading!

FORECASTING - VALIDATION

- How well does our model explain our data? Couple of ways of evaluating this in the binary D.V. context? The Greenhill et. al. reading for this week gives a general overview.
- People often look at accuracy – given some "cut-off" probability, how many cases are correctly predicted by the model. Accuracy can be misleading!
- Suppose we have a very rare event – e.g. only 1% of cases are 1s.

Forecasting - Validation

- How well does our model explain our data? Couple of ways of evaluating this in the binary D.V. context? The Greenhill et. al. reading for this week gives a general overview.
- People often look at accuracy – given some “cut-off” probability, how many cases are correctly predicted by the model. Accuracy can be misleading!
- Suppose we have a very rare event – e.g. only 1% of cases are 1s. Then a model that just always predicted 0 would have 99% accuracy!
- Instead, in binary classification, we often care about *sensitivity* vs. *specificity*

Forecasting - Sensitivity and Specificity

	Actual Outcome	
Predicted Outcome	Negative	Positive
Negative	True Negative	**False Negative**
Positive	**False Positive**	True Positive

Table : Confusion matrix for binary predictions

- Sensitivity = True Positive Rate = $\frac{\sum \text{True Positive}}{\sum \text{Actual Positives}}$
- Specificity = True Negative Rate = $\frac{\sum \text{True Negative}}{\sum \text{Actual Negatives}}$

Forecasting - Sensitivity and Specificity

- Given a model, there's a trade-off between Sensitivity and Specificity.

Forecasting - Sensitivity and Specificity

- Given a model, there's a trade-off between Sensitivity and Specificity. In a naive model, we can always get 100% sensitivity by labeling everything as a positive.

Forecasting - Sensitivity and Specificity

- Given a model, there's a trade-off between Sensitivity and Specificity. In a naive model, we can always get 100% sensitivity by labeling everything as a positive. But this would yield 0% specificity.

Forecasting - Sensitivity and Specificity

- Given a model, there's a trade-off between Sensitivity and Specificity. In a naive model, we can always get 100% sensitivity by labeling everything as a positive. But this would yield 0% specificity.
- Sensitivity and specificity are going to depend on the "cutoff" $\hat{\pi}_0$ that we use to classify observations as either 0s or 1s.

FORECASTING - SENSITIVITY AND SPECIFICITY

- Given a model, there's a trade-off between Sensitivity and Specificity. In a naive model, we can always get 100% sensitivity by labeling everything as a positive. But this would yield 0% specificity.
- Sensitivity and specificity are going to depend on the "cutoff" $\hat{\pi}_0$ that we use to classify observations as either 0s or 1s. So we can control one, and see how well we do on the other.

FORECASTING - SENSITIVITY AND SPECIFICITY

- Given a model, there's a trade-off between Sensitivity and Specificity. In a naive model, we can always get 100% sensitivity by labeling everything as a positive. But this would yield 0% specificity.
- Sensitivity and specificity are going to depend on the "cutoff" $\hat{\pi}_0$ that we use to classify observations as either 0s or 1s. So we can control one, and see how well we do on the other.
- How do we quantify the trade-off for our particular model.

Forecasting - Sensitivity and Specificity

- Given a model, there's a trade-off between Sensitivity and Specificity. In a naive model, we can always get 100% sensitivity by labeling everything as a positive. But this would yield 0% specificity.
- Sensitivity and specificity are going to depend on the "cutoff" $\hat{\pi}_0$ that we use to classify observations as either 0s or 1s. So we can control one, and see how well we do on the other.
- How do we quantify the trade-off for our particular model. Receiver Operating Characteristic (ROC) plots!

FORECASTING - SENSITIVITY AND SPECIFICITY

- Given a model, there's a trade-off between Sensitivity and Specificity. In a naive model, we can always get 100% sensitivity by labeling everything as a positive. But this would yield 0% specificity.
- Sensitivity and specificity are going to depend on the "cutoff" $\hat{\pi}_0$ that we use to classify observations as either 0s or 1s. So we can control one, and see how well we do on the other.
- How do we quantify the trade-off for our particular model. Receiver Operating Characteristic (ROC) plots!
  - Basically, a plot of True Positive Rate on Y axis against False Positive Rate (1 - True Negative Rate) on the X axis.

FORECASTING - MAKING AN ROC PLOT

Forecasting - Making an ROC plot

- How to create an ROC:
  - Pick a threshold $\pi_0 \in [0, 1]$.
  - For your test data, generate predictions for each observation $\hat{\pi}_i$.
  - Predict $\hat{Y}_i = 0$ if $\pi_i < \pi_0$ and $\hat{Y}_i = 1$ otherwise.
  - Calculate sensitivity and specificity.
  - Repeat for values of π_0 from 0 to 1 and plot.

Forecasting - Making an ROC plot

FORECASTING - MAKING AN ROC PLOT

Here's what it looks like in R.

Forecasting - Making an ROC plot

Here's what it looks like in R. First we get our predicted probabilities.

Forecasting - Making an ROC plot

Here's what it looks like in R. First we get our predicted probabilities.

```
#### ROC curve
thresholds <- seq(0, 1, by=.001) ## Vector of thresholds to test
sensitivity <- rep(NA, length(thresholds))
specificity <- rep(NA, length(thresholds))

### Get predicted probabilities
pred.probs <- 1/(1 + exp(-design.matrix%*%coefs))
```

Forecasting - Making an ROC plot

Forecasting - Making an ROC plot

Then we calculate true positive rate and true negative rate for each threshold

```
### For each threshold
for(i in 1:length(thresholds)){
  ### Select the threshold
  thresh <- thresholds[i]
  ### Make a prediction
  y_hat <- ifelse(pred.probs < thresh, 0, 1)
  ### Compare to true Y
  cross_tab <- table(y_hat, votes$incwin)
  ### R-hack - Make sure cross_tab is a 2x2.
  if (nrow(cross_tab) == 2 & ncol(cross_tab) == 2){
    ## True positive rate (1s correctly predicted/total 1s)
    tpr <- cross_tab[2,2]/(cross_tab[2,2] + cross_tab[1,2])
    ## True negative rate
    tnr <- cross_tab[1,1]/(cross_tab[1,1] + cross_tab[2,1])
  }else{
  ### If we only predicted one class
    if (max(y_hat) == 0){
      ### If we only predict zeroes, no false positives, but no true positives
      tpr <- 0
      tnr <- 1
    }else if (min(y_hat) == 1){
      ### If we only predict 1s, no true negatives, but all true positives
      tpr <- 1
      tnr <- 0
    }
  }
  sensitivity[i] <- tpr
  specificity[i] <- tnr
}
```

FORECASTING - MAKING AN ROC PLOT

FORECASTING - MAKING AN ROC PLOT

Finally, we plot it!

```
pdf("ROC_house.pdf")
plot(x=1-specificity, y=sensitivity, type="s", xlab="False Positive Rate", col="
      dodgerblue", lwd=4,
     ylab="True Positive Rate", main="ROC Curve for in-sample forecasts\nof House
          elections 04-08", xlim=c(0,1), ylim=c(0,1))
abline(0,1, lty=2, lwd=2) ## 45 degree line
abline(v=1)
abline(v=0)
abline(h=1)
abline(h=0)
dev.off()
```

Forecasting - Making an ROC plot

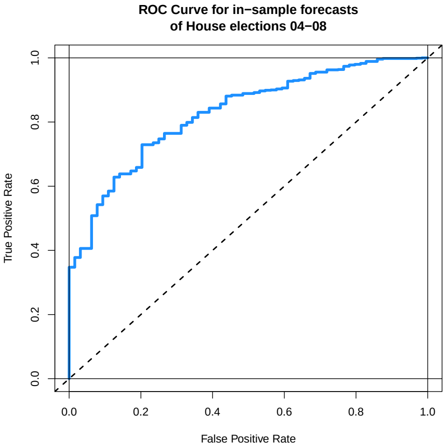

Figure : ROC Plot for House Election Logit Model

Forecasting - In-Sample vs. Out-of-sample forecasting

- Careful when validating models to be wary of over-fitting.

Forecasting - In-Sample vs. Out-of-sample forecasting

- Careful when validating models to be wary of over-fitting. We can have a model fit perfectly to our sample that does terribly on other samples.

Forecasting - In-Sample vs. Out-of-sample forecasting

- Careful when validating models to be wary of over-fitting. We can have a model fit perfectly to our sample that does terribly on other samples. This is because our model is *over-fit* – the estimates are highly sensitive to arbitrary noise in the sample.
- Analogy – Can always get a better R^2 by adding more junk to a linear model – does that make a model with millions of covariates better?

Forecasting - In-Sample vs. Out-of-sample forecasting

- Careful when validating models to be wary of over-fitting. We can have a model fit perfectly to our sample that does terribly on other samples. This is because our model is *over-fit* – the estimates are highly sensitive to arbitrary noise in the sample.
- Analogy – Can always get a better R^2 by adding more junk to a linear model – does that make a model with millions of covariates better?
- Solution is to fit a model to one part of the data and use it to forecast another part – "cross-validation."

Forecasting - In-Sample vs. Out-of-sample forecasting

- Careful when validating models to be wary of over-fitting. We can have a model fit perfectly to our sample that does terribly on other samples. This is because our model is *over-fit* – the estimates are highly sensitive to arbitrary noise in the sample.
- Analogy – Can always get a better R^2 by adding more junk to a linear model – does that make a model with millions of covariates better?
- Solution is to fit a model to one part of the data and use it to forecast another part – “cross-validation.”
- Often see all of these prediction diagnostics used on a “held-out” set of data that the model does not “see” during estimation.

QUESTIONS

Questions?