

Political Methodology III: Model Based Inference

Justin Grimmer

Professor
Department of Political Science
Stanford University

May 7th, 2019

Model Based Inference

- 1) Likelihood inference
- 2) Logit/Probit
- 3) Ordered Probit
- 4) Choice Models:
- 5) Count Models
- 6) Survival Models
- 7) Hypothesis Tests + Model Checking in Likelihood
 - Likelihood Ratios, Wald, and Score tests
 - Model Checking: analysis of residuals, hat values, etc.

Simple Example: Antiobama Speech

We'll use the speech data from the problem set, as follows:

- $Y_i = 1$ if representative says obamacare or big government during the year, 0 otherwise
- $\mathbf{X}_i = (1, I(\text{Year} = 2010)_i, \text{Democrat}_i, \text{DW-Nom}_i)$

$$Y_i \sim \text{Bernoulli}(\pi_i)$$
$$\pi_i = \text{logit}^{-1}(\mathbf{X}_i' \boldsymbol{\beta}) = \frac{1}{1 + \exp(-\mathbf{X}_i' \boldsymbol{\beta})}$$

Which covariates do we include? \rightsquigarrow depends on goal.

- Predictive goal \rightsquigarrow replicate task
- Model fitting \rightsquigarrow do covariates increase likelihood? Can we drop them?

Hypothesis Testing — Likelihood Ratio Test

- Null (H_0): $h_1(\beta) = \dots = h_Q(\beta) = 0$ (Q equality constraints)
- Alternative (H_1): No such constraints
- Let $\hat{\beta}_R = \hat{\beta}_{MLE|H_0}$ (restricted MLE) and $\hat{\beta}_{UR} = \hat{\beta}_{MLE}$ (original MLE)
- **Likelihood ratio** (LR) test: If H_0 is true, $L(\hat{\beta}_R)$ should be equal to $L(\hat{\beta}_{UR})$ except for sampling variability
- LR statistic:

$$LR(Y) \equiv -2 \log \frac{L(\hat{\beta}_R)}{L(\hat{\beta}_{UR})} = 2 \left[\ell(\hat{\beta}_{UR}) - \ell(\hat{\beta}_R) \right]$$

- We can show that $LR(Y) \xrightarrow{d} \chi_Q^2$
- Works for testing any nested models
— model under H_0 has to be a special case of model under H_1

Hypothesis Testing — Likelihood Ratio Test

- Null (H_0): $h_1(\beta) = \dots = h_Q(\beta) = 0$ (Q equality constraints)
- Alternative (H_1): No such constraints
- Let $\hat{\beta}_R = \hat{\beta}_{MLE|H_0}$ (restricted MLE) and $\hat{\beta}_{UR} = \hat{\beta}_{MLE}$ (original MLE)
- **Likelihood ratio** (LR) test: If H_0 is true, $L(\hat{\beta}_R)$ should be equal to $L(\hat{\beta}_{UR})$ except for sampling variability
- LR statistic:

$$LR(Y) \equiv -2 \log \frac{L(\hat{\beta}_R)}{L(\hat{\beta}_{UR})} = 2 \left[\ell(\hat{\beta}_{UR}) - \ell(\hat{\beta}_R) \right]$$

- We can show that $LR(Y) \xrightarrow{d} \chi_Q^2$
- Works for testing any nested models
— model under H_0 has to be a special case of model under H_1

Hypothesis Testing — Likelihood Ratio Test

- Null (H_0): $h_1(\beta) = \dots = h_Q(\beta) = 0$ (Q equality constraints)
- Alternative (H_1): No such constraints
- Let $\hat{\beta}_R = \hat{\beta}_{MLE|H_0}$ (restricted MLE) and $\hat{\beta}_{UR} = \hat{\beta}_{MLE}$ (original MLE)
- **Likelihood ratio** (LR) test: If H_0 is true, $L(\hat{\beta}_R)$ should be equal to $L(\hat{\beta}_{UR})$ except for sampling variability
- LR statistic:

$$LR(Y) \equiv -2 \log \frac{L(\hat{\beta}_R)}{L(\hat{\beta}_{UR})} = 2 \left[\ell(\hat{\beta}_{UR}) - \ell(\hat{\beta}_R) \right]$$

- We can show that $LR(Y) \xrightarrow{d} \chi_Q^2$
- Works for testing any nested models
— model under H_0 has to be a special case of model under H_1

Hypothesis Testing — Likelihood Ratio Test

- Null (H_0): $h_1(\beta) = \dots = h_Q(\beta) = 0$ (Q equality constraints)
- Alternative (H_1): No such constraints
- Let $\hat{\beta}_R = \hat{\beta}_{MLE|H_0}$ (restricted MLE) and $\hat{\beta}_{UR} = \hat{\beta}_{MLE}$ (original MLE)
- **Likelihood ratio** (LR) test: If H_0 is true, $L(\hat{\beta}_R)$ should be equal to $L(\hat{\beta}_{UR})$ except for sampling variability
- LR statistic:

$$LR(Y) \equiv -2 \log \frac{L(\hat{\beta}_R)}{L(\hat{\beta}_{UR})} = 2 \left[\ell(\hat{\beta}_{UR}) - \ell(\hat{\beta}_R) \right]$$

- We can show that $LR(Y) \xrightarrow{d} \chi_Q^2$
- Works for testing any nested models
— model under H_0 has to be a special case of model under H_1

Hypothesis Testing — Likelihood Ratio Test

- Null (H_0): $h_1(\beta) = \dots = h_Q(\beta) = 0$ (Q equality constraints)
- Alternative (H_1): No such constraints
- Let $\hat{\beta}_R = \hat{\beta}_{MLE|H_0}$ (restricted MLE) and $\hat{\beta}_{UR} = \hat{\beta}_{MLE}$ (original MLE)
- **Likelihood ratio** (LR) test: If H_0 is true, $L(\hat{\beta}_R)$ should be equal to $L(\hat{\beta}_{UR})$ except for sampling variability
- LR statistic:

$$LR(Y) \equiv -2 \log \frac{L(\hat{\beta}_R)}{L(\hat{\beta}_{UR})} = 2 \left[\ell(\hat{\beta}_{UR}) - \ell(\hat{\beta}_R) \right]$$

- We can show that $LR(Y) \xrightarrow{d} \chi_Q^2$
- Works for testing any nested models
— model under H_0 has to be a special case of model under H_1

Hypothesis Testing — Likelihood Ratio Test

- Null (H_0): $h_1(\beta) = \dots = h_Q(\beta) = 0$ (Q equality constraints)
- Alternative (H_1): No such constraints
- Let $\hat{\beta}_R = \hat{\beta}_{MLE|H_0}$ (restricted MLE) and $\hat{\beta}_{UR} = \hat{\beta}_{MLE}$ (original MLE)
- **Likelihood ratio** (LR) test: If H_0 is true, $L(\hat{\beta}_R)$ should be equal to $L(\hat{\beta}_{UR})$ except for sampling variability
- LR statistic:

$$LR(Y) \equiv -2 \log \frac{L(\hat{\beta}_R)}{L(\hat{\beta}_{UR})} = 2 \left[\ell(\hat{\beta}_{UR}) - \ell(\hat{\beta}_R) \right]$$

- We can show that $LR(Y) \xrightarrow{d} \chi_Q^2$
- Works for testing any nested models
— model under H_0 has to be a special case of model under H_1

Hypothesis Testing — Likelihood Ratio Test

- Null (H_0): $h_1(\beta) = \dots = h_Q(\beta) = 0$ (Q equality constraints)
- Alternative (H_1): No such constraints
- Let $\hat{\beta}_R = \hat{\beta}_{MLE|H_0}$ (restricted MLE) and $\hat{\beta}_{UR} = \hat{\beta}_{MLE}$ (original MLE)
- **Likelihood ratio** (LR) test: If H_0 is true, $L(\hat{\beta}_R)$ should be equal to $L(\hat{\beta}_{UR})$ except for sampling variability
- LR statistic:

$$LR(Y) \equiv -2 \log \frac{L(\hat{\beta}_R)}{L(\hat{\beta}_{UR})} = 2 \left[\ell(\hat{\beta}_{UR}) - \ell(\hat{\beta}_R) \right]$$

- We can show that $LR(Y) \xrightarrow{d} \chi_Q^2$
- Works for testing any nested models
— model under H_0 has to be a special case of model under H_1

```

un_rest_reg<- glm(once~two_10 + dem + dw_nom,
  data = speech_dat, family = binomial(link = logit))

rest_reg<- glm(once~1, family= binomial(link = logit))

##calculating the likelihood ratio
log_lik<- function(pars, X, Y){
  y.tilde<- X%*%pars
  probs<- plogis(y.tilde)
  log_out<- Y%*%log(probs) + (1-Y)%*%log(1 - probs)
  return(log_out)
}
X<- cbind(1, two_10, dem, speech_dat$dw_nom)

un_rest<- log_lik(un_rest_reg$coef, X, once)

rest<- log_lik(rest_reg$coef, as.matrix(rep(1, nrow(X)))), once)

> 2 * un_rest - 2*rest
  [,1]
[1,] 433.996

```

```
> 2 * un_rest - 2*rest
      [,1]
[1,] 433.996
##get the same statistic automatically from glm
diff<- un_rest_reg$null.deviance - un_rest_reg$deviance
> diff
[1] 433.996

1 - pchisq(diff, 3) ##very small!
[1] 0
```

Hypothesis Testing — Wald Test

- **Wald test:** If true, the null $h_1(\beta) = \dots = h_Q(\beta) = 0$ should approximately hold even if we substitute $\hat{\beta}_{UR}$ for β . Call $h(\beta) = (h_1(\beta), \dots, h_Q(\beta))$
- Wald statistic: Use asymptotic distribution of $\hat{\beta}$ and representation of restrictions, properties of normal distribution to obtain form

$$W \equiv h(\hat{\beta}_{UR})' \left[\left(\frac{\partial h(\beta)}{\partial \beta} \Big|_{\beta=\hat{\beta}_{UR}} \right)' \widehat{\text{Var}}(\hat{\beta}_{UR}) \left(\frac{\partial h(\beta)}{\partial \beta} \Big|_{\beta=\hat{\beta}_{UR}} \right) \right]^{-1} h(\hat{\beta}_{UR})$$

- The “meat” $\simeq \widehat{\text{Var}}(h(\hat{\beta}_{UR}))$ (**Delta method**)
- Choose any $\widehat{\text{Var}}(\hat{\beta}_{UR})$ as appropriate (e.g. Huber-White)
- We can show that $W \xrightarrow{d} \chi_Q^2$
- An important special case: $Q = 1$ and $H_0 : \beta = 0$
- In this case, we can use the **z statistic**:

$$z = W^{1/2} = \frac{\hat{\beta}_{UR}}{\text{s.e.}(\hat{\beta}_{UR})} \xrightarrow{d} N(0, 1)$$

Hypothesis Testing — Wald Test

- **Wald test:** If true, the null $h_1(\beta) = \dots = h_Q(\beta) = 0$ should approximately hold even if we substitute $\widehat{\beta}_{UR}$ for β . Call $h(\beta) = (h_1(\beta), \dots, h_Q(\beta))$
- Wald statistic: Use asymptotic distribution of $\widehat{\beta}$ and representation of restrictions, properties of normal distribution to obtain form

$$W \equiv h(\widehat{\beta}_{UR})' \left[\left(\frac{\partial h(\beta)}{\partial \beta} \Big|_{\beta=\widehat{\beta}_{UR}} \right)' \widehat{\text{Var}}(\widehat{\beta}_{UR}) \left(\frac{\partial h(\beta)}{\partial \beta} \Big|_{\beta=\widehat{\beta}_{UR}} \right) \right]^{-1} h(\widehat{\beta}_{UR})$$

- The “meat” $\simeq \widehat{\text{Var}}(h(\widehat{\beta}_{UR}))$ (**Delta method**)
- Choose any $\widehat{\text{Var}}(\widehat{\beta}_{UR})$ as appropriate (e.g. Huber-White)
- We can show that $W \xrightarrow{d} \chi_Q^2$
- An important special case: $Q = 1$ and $H_0 : \beta = 0$
- In this case, we can use the **z statistic**:

$$z = W^{1/2} = \frac{\widehat{\beta}_{UR}}{\text{s.e.}(\widehat{\beta}_{UR})} \xrightarrow{d} N(0, 1)$$

Hypothesis Testing — Wald Test

- **Wald test:** If true, the null $h_1(\beta) = \dots = h_Q(\beta) = 0$ should approximately hold even if we substitute $\widehat{\beta}_{UR}$ for β . Call $h(\beta) = (h_1(\beta), \dots, h_Q(\beta))$
- Wald statistic: Use asymptotic distribution of $\widehat{\beta}$ and representation of restrictions, properties of normal distribution to obtain form

$$W \equiv h(\widehat{\beta}_{UR})' \left[\left(\frac{\partial h(\beta)}{\partial \beta} \Big|_{\beta=\widehat{\beta}_{UR}} \right)' \widehat{\text{Var}}(\widehat{\beta}_{UR}) \left(\frac{\partial h(\beta)}{\partial \beta} \Big|_{\beta=\widehat{\beta}_{UR}} \right) \right]^{-1} h(\widehat{\beta}_{UR})$$

- The “meat” $\simeq \widehat{\text{Var}}(h(\widehat{\beta}_{UR}))$ (**Delta method**)
- Choose any $\widehat{\text{Var}}(\widehat{\beta}_{UR})$ as appropriate (e.g. Huber-White)
- We can show that $W \xrightarrow{d} \chi_Q^2$
- An important special case: $Q = 1$ and $H_0 : \beta = 0$
- In this case, we can use the **z statistic**:

$$z = W^{1/2} = \frac{\widehat{\beta}_{UR}}{\text{s.e.}(\widehat{\beta}_{UR})} \xrightarrow{d} N(0, 1)$$

Hypothesis Testing — Wald Test

- **Wald test:** If true, the null $h_1(\beta) = \dots = h_Q(\beta) = 0$ should approximately hold even if we substitute $\widehat{\beta}_{UR}$ for β . Call $h(\beta) = (h_1(\beta), \dots, h_Q(\beta))$
- Wald statistic: Use asymptotic distribution of $\widehat{\beta}$ and representation of restrictions, properties of normal distribution to obtain form

$$W \equiv h(\widehat{\beta}_{UR})' \left[\left(\frac{\partial h(\beta)}{\partial \beta} \Big|_{\beta=\widehat{\beta}_{UR}} \right)' \widehat{\text{Var}}(\widehat{\beta}_{UR}) \left(\frac{\partial h(\beta)}{\partial \beta} \Big|_{\beta=\widehat{\beta}_{UR}} \right) \right]^{-1} h(\widehat{\beta}_{UR})$$

- The “meat” $\simeq \widehat{\text{Var}}(h(\widehat{\beta}_{UR}))$ (**Delta method**)
- Choose any $\widehat{\text{Var}}(\widehat{\beta}_{UR})$ as appropriate (e.g. Huber-White)
- We can show that $W \xrightarrow{d} \chi_Q^2$
- An important special case: $Q = 1$ and $H_0 : \beta = 0$
- In this case, we can use the **z statistic**:

$$z = W^{1/2} = \frac{\widehat{\beta}_{UR}}{\text{s.e.}(\widehat{\beta}_{UR})} \xrightarrow{d} N(0, 1)$$

Hypothesis Testing — Wald Test

- **Wald test:** If true, the null $h_1(\beta) = \dots = h_Q(\beta) = 0$ should approximately hold even if we substitute $\widehat{\beta}_{UR}$ for β . Call $h(\beta) = (h_1(\beta), \dots, h_Q(\beta))$
- Wald statistic: Use asymptotic distribution of $\widehat{\beta}$ and representation of restrictions, properties of normal distribution to obtain form

$$W \equiv h(\widehat{\beta}_{UR})' \left[\left(\frac{\partial h(\beta)}{\partial \beta} \Big|_{\beta=\widehat{\beta}_{UR}} \right)' \widehat{\text{Var}}(\widehat{\beta}_{UR}) \left(\frac{\partial h(\beta)}{\partial \beta} \Big|_{\beta=\widehat{\beta}_{UR}} \right) \right]^{-1} h(\widehat{\beta}_{UR})$$

- The “meat” $\simeq \widehat{\text{Var}}(h(\widehat{\beta}_{UR}))$ (**Delta method**)
- Choose any $\widehat{\text{Var}}(\widehat{\beta}_{UR})$ as appropriate (e.g. Huber-White)
- We can show that $W \xrightarrow{d} \chi_Q^2$
- An important special case: $Q = 1$ and $H_0 : \beta = 0$
- In this case, we can use the **z statistic**:

$$z = W^{1/2} = \frac{\widehat{\beta}_{UR}}{\text{s.e.}(\widehat{\beta}_{UR})} \xrightarrow{d} N(0, 1)$$

Hypothesis Testing — Wald Test

- **Wald test:** If true, the null $h_1(\beta) = \dots = h_Q(\beta) = 0$ should approximately hold even if we substitute $\widehat{\beta}_{UR}$ for β . Call $h(\beta) = (h_1(\beta), \dots, h_Q(\beta))$
- Wald statistic: Use asymptotic distribution of $\widehat{\beta}$ and representation of restrictions, properties of normal distribution to obtain form

$$W \equiv h(\widehat{\beta}_{UR})' \left[\left(\frac{\partial h(\beta)}{\partial \beta} \Big|_{\beta=\widehat{\beta}_{UR}} \right)' \widehat{\text{Var}}(\widehat{\beta}_{UR}) \left(\frac{\partial h(\beta)}{\partial \beta} \Big|_{\beta=\widehat{\beta}_{UR}} \right) \right]^{-1} h(\widehat{\beta}_{UR})$$

- The “meat” $\simeq \widehat{\text{Var}}(h(\widehat{\beta}_{UR}))$ (**Delta method**)
- Choose any $\widehat{\text{Var}}(\widehat{\beta}_{UR})$ as appropriate (e.g. Huber-White)
- We can show that $W \xrightarrow{d} \chi_Q^2$
- An important special case: $Q = 1$ and $H_0 : \beta = 0$
- In this case, we can use the **z statistic**:

$$z = W^{1/2} = \frac{\widehat{\beta}_{UR}}{\text{s.e.}(\widehat{\beta}_{UR})} \xrightarrow{d} N(0, 1)$$

Hypothesis Testing — Wald Test

- **Wald test:** If true, the null $h_1(\beta) = \dots = h_Q(\beta) = 0$ should approximately hold even if we substitute $\widehat{\beta}_{UR}$ for β . Call $h(\beta) = (h_1(\beta), \dots, h_Q(\beta))$
- Wald statistic: Use asymptotic distribution of $\widehat{\beta}$ and representation of restrictions, properties of normal distribution to obtain form

$$W \equiv h(\widehat{\beta}_{UR})' \left[\left(\frac{\partial h(\beta)}{\partial \beta} \Big|_{\beta=\widehat{\beta}_{UR}} \right)' \widehat{\text{Var}}(\widehat{\beta}_{UR}) \left(\frac{\partial h(\beta)}{\partial \beta} \Big|_{\beta=\widehat{\beta}_{UR}} \right) \right]^{-1} h(\widehat{\beta}_{UR})$$

- The “meat” $\simeq \widehat{\text{Var}}(h(\widehat{\beta}_{UR}))$ (**Delta method**)
- Choose any $\widehat{\text{Var}}(\widehat{\beta}_{UR})$ as appropriate (e.g. Huber-White)
- We can show that $W \xrightarrow{d} \chi_Q^2$
- An important special case: $Q = 1$ and $H_0 : \beta = 0$
- In this case, we can use the **z statistic**:

$$z = W^{1/2} = \frac{\widehat{\beta}_{UR}}{\text{s.e.}(\widehat{\beta}_{UR})} \xrightarrow{d} N(0, 1)$$

```
> un_rest_reg$coef%*%solve(vcov(un_rest_reg))%*%un_rest_reg$coef  
[,1]  
[1,] 225.2437  
  
> 1 - pchisq(225.2437, 3)  
[1] 0
```

Hypothesis Testing — Score Test

- At the unrestricted MLE $\hat{\theta}_{UR}$, $\sum_{i=1}^N s_i(\hat{\beta}_{UR}) = s(\hat{\beta}) = 0$ by construction
- **Score test**: If the null is true, $s(\hat{\beta}_R)$ should also equal zero except for sampling variability
- Score statistic: Use asymptotic distribution and properties of normal distribution to “standardize” $s(\hat{\beta}_R)$

$$LM = s(\hat{\beta}_R)' \widehat{\text{Var}}(\hat{\beta}_R) s(\hat{\beta}_R) \xrightarrow{d} \chi_Q^2$$

- For $\hat{\beta}_{QMLE}$, the expression is more complicated
- Also known as the **Lagrange multiplier** (LM) test due to an alternative derivation

Hypothesis Testing — Score Test

- At the unrestricted MLE $\hat{\theta}_{UR}$, $\sum_{i=1}^N s_i(\hat{\beta}_{UR}) = s(\hat{\beta}) = 0$ by construction
- **Score test**: If the null is true, $s(\hat{\beta}_R)$ should also equal zero except for sampling variability
- Score statistic: Use asymptotic distribution and properties of normal distribution to “standardize” $s(\hat{\beta}_R)$

$$LM = s(\hat{\beta}_R)' \widehat{\text{Var}}(\hat{\beta}_R) s(\hat{\beta}_R) \xrightarrow{d} \chi_Q^2$$

- For $\hat{\beta}_{QMLE}$, the expression is more complicated
- Also known as the **Lagrange multiplier** (LM) test due to an alternative derivation

Hypothesis Testing — Score Test

- At the unrestricted MLE $\hat{\theta}_{UR}$, $\sum_{i=1}^N s_i(\hat{\beta}_{UR}) = s(\hat{\beta}) = 0$ by construction
- **Score test**: If the null is true, $s(\hat{\beta}_R)$ should also equal zero except for sampling variability
- Score statistic: Use asymptotic distribution and properties of normal distribution to “standardize” $s(\hat{\beta}_R)$

$$LM = s(\hat{\beta}_R)' \widehat{\text{Var}}(\hat{\beta}_R) s(\hat{\beta}_R) \xrightarrow{d} \chi_Q^2$$

- For $\hat{\beta}_{QMLE}$, the expression is more complicated
- Also known as the **Lagrange multiplier** (LM) test due to an alternative derivation

Hypothesis Testing — Score Test

- At the unrestricted MLE $\hat{\theta}_{UR}$, $\sum_{i=1}^N s_i(\hat{\beta}_{UR}) = s(\hat{\beta}) = 0$ by construction
- **Score test**: If the null is true, $s(\hat{\beta}_R)$ should also equal zero except for sampling variability
- Score statistic: Use asymptotic distribution and properties of normal distribution to “standardize” $s(\hat{\beta}_R)$

$$LM = s(\hat{\beta}_R)' \widehat{\text{Var}}(\hat{\beta}_R) s(\hat{\beta}_R) \xrightarrow{d} \chi_Q^2$$

- For $\hat{\beta}_{QMLE}$, the expression is more complicated
- Also known as the **Lagrange multiplier** (LM) test due to an alternative derivation

Hypothesis Testing — Score Test

- At the unrestricted MLE $\hat{\theta}_{UR}$, $\sum_{i=1}^N s_i(\hat{\beta}_{UR}) = s(\hat{\beta}) = 0$ by construction
- **Score test**: If the null is true, $s(\hat{\beta}_R)$ should also equal zero except for sampling variability
- Score statistic: Use asymptotic distribution and properties of normal distribution to “standardize” $s(\hat{\beta}_R)$

$$LM = s(\hat{\beta}_R)' \widehat{\text{Var}}(\hat{\beta}_R) s(\hat{\beta}_R) \xrightarrow{d} \chi_Q^2$$

- For $\hat{\beta}_{QMLE}$, the expression is more complicated
- Also known as the **Lagrange multiplier** (LM) test due to an alternative derivation

```
score_func<- function(coef, X, Y){  
  y.tilde<- X%*%coef  
  probs<- plogis(y.tilde)  
  out<- t(Y - probs)%*%X  
  return(out) }
```

```
> round(score_func(un_rest_reg$coef, X, once), 2)
```

```
[1,] 0 0 0 0
```

```
rest_score<- score_func(c(rest_reg$coef, 0, 0, 0), X, once)
```

```
> round(rest_score,2)
```

```
[1,] 0 -6.30 -128.92 129.51
```

```

hess_func<- function(coef, X, Y){
  y.tilde<- X%*%coef
  probs<- plogis(y.tilde)
  base<- matrix(0, nrow = len(coef), ncol = len(coef))
  for(z in 1:nrow(X)){
    base<- base + probs[z]*(1 - probs[z])* X[z,]%*%t(X[z,])
  }
  return(base)
}

rest_hess<- solve(hess_func(c(rest_reg$coef, 0, 0, 0), X, once))
>rest_score%*%rest_hess%*%t(rest_score)
[1,] 395.0382
> 1- pchisq(395, 3)
[1] 0

```

Comparing The Three Tests

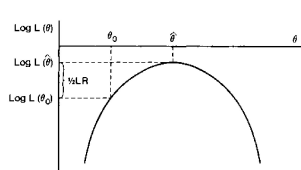


Figure 1. The Likelihood Ratio Test

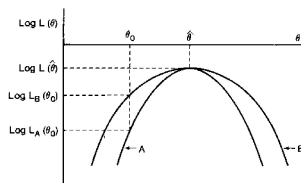


Figure 2. The Wald Test

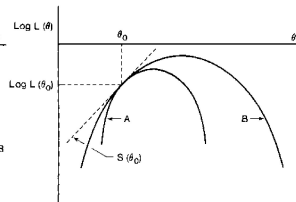


Figure 3. The Lagrange Multiplier Test

- All asymptotically equivalent
- But can be quite different in small samples

	<i>Pros</i>	<i>Cons</i>
LR	Most powerful (Neyman-Pearson)	Must compute both $\hat{\theta}_{UR}$ and $\hat{\theta}_R$ Cannot be easily robustified
W	Only need $\hat{\theta}_{UR}$ Easily robustified by sandwich	Not invariant to transformation (e.g. $\theta_1/\theta_2 = 1$ vs. $\theta_1 = \theta_2$)
LM	Only need $\hat{\theta}_R$	$\hat{\theta}_R$ often difficult to estimate

Comparing The Three Tests

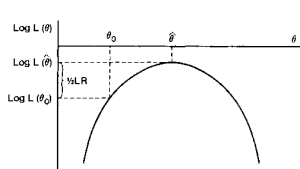


Figure 1. The Likelihood Ratio Test

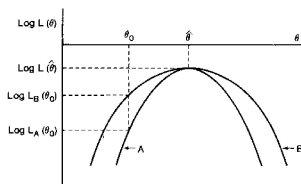


Figure 2. The Wald Test

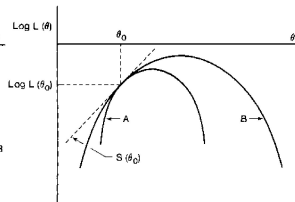


Figure 3. The Lagrange Multiplier Test

- All asymptotically equivalent
- But can be quite different in small samples

	Pros	Cons
LR	Most powerful (Neyman-Pearson)	Must compute both $\hat{\theta}_{UR}$ and $\hat{\theta}_R$ Cannot be easily robustified
W	Only need $\hat{\theta}_{UR}$ Easily robustified by sandwich	Not invariant to transformation (e.g. $\theta_1/\theta_2 = 1$ vs. $\theta_1 = \theta_2$)
LM	Only need $\hat{\theta}_R$	$\hat{\theta}_R$ often difficult to estimate

Comparing The Three Tests

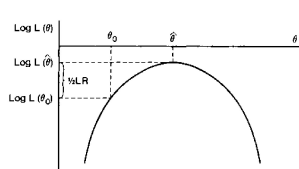


Figure 1. The Likelihood Ratio Test

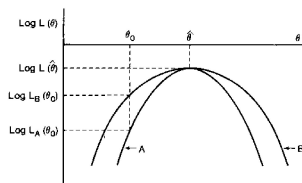


Figure 2. The Wald Test

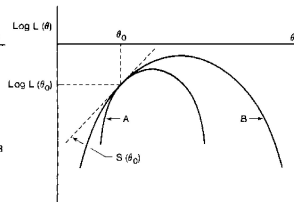


Figure 3. The Lagrange Multiplier Test

- All asymptotically equivalent
- But can be quite different in small samples

	Pros	Cons
LR	Most powerful (Neyman-Pearson)	Must compute both $\hat{\theta}_{UR}$ and $\hat{\theta}_R$ Cannot be easily robustified
W	Only need $\hat{\theta}_{UR}$ Easily robustified by sandwich	Not invariant to transformation (e.g. $\theta_1/\theta_2 = 1$ vs. $\theta_1 = \theta_2$)
LM	Only need $\hat{\theta}_R$	$\hat{\theta}_R$ often difficult to estimate

Comparing The Three Tests

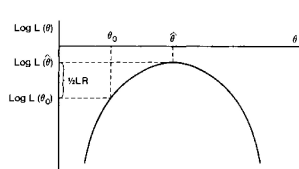


Figure 1. The Likelihood Ratio Test

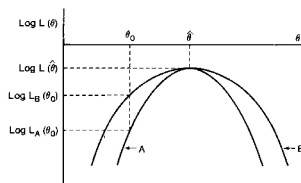


Figure 2. The Wald Test

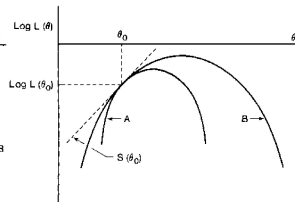


Figure 3. The Lagrange Multiplier Test

- All asymptotically equivalent
- But can be quite different in small samples

	<i>Pros</i>	<i>Cons</i>
LR	Most powerful (Neyman-Pearson)	Must compute both $\hat{\theta}_{UR}$ and $\hat{\theta}_R$ Cannot be easily robustified
W	Only need $\hat{\theta}_{UR}$ Easily robustified by sandwich	Not invariant to transformation (e.g. $\theta_1/\theta_2 = 1$ vs. $\theta_1 = \theta_2$)
LM	Only need $\hat{\theta}_R$	$\hat{\theta}_R$ often difficult to estimate

Generalized Linear Models

- Many of the models we have learned so far all assume that Y_i is a (stochastic) function of the **linear predictor**, $X_i^\top \beta$
- They also share many characteristics, e.g. the form of the score and Hessian functions
- In fact, many are special cases of the **generalized linear model (GLM)**
- Here, we provide a general treatment of GLMs to study those models more systematically
- 3 components of a GLM
 - 1 Systematic component: $X_i^\top \beta$
 - Must be a linear function of X_i
 - 2 Random component: $f(Y; \theta, \phi)$
 - Must be in the **exponential family**
 - θ is called the canonical parameter
 - ϕ : is called the dispersion parameter
 - 3 Link function: $g(\mu_i) = X_i^\top \beta$ where $\mu_i = E(Y_i | X_i)$
 - Must be monotonic and differentiable wrt μ_i

Generalized Linear Models

- Many of the models we have learned so far all assume that Y_i is a (stochastic) function of the **linear predictor**, $X_i^\top \beta$
- They also share many characteristics, e.g. the form of the score and Hessian functions
- In fact, many are special cases of the **generalized linear model (GLM)**
- Here, we provide a general treatment of GLMs to study those models more systematically
- 3 components of a GLM
 - 1 Systematic component: $X_i^\top \beta$
 - Must be a linear function of X_i
 - 2 Random component: $f(Y; \theta, \phi)$
 - Must be in the **exponential family**
 - θ is called the canonical parameter
 - ϕ : is called the dispersion parameter
 - 3 Link function: $g(\mu_i) = X_i^\top \beta$ where $\mu_i = E(Y_i | X_i)$
 - Must be monotonic and differentiable wrt μ_i

Generalized Linear Models

- Many of the models we have learned so far all assume that Y_i is a (stochastic) function of the **linear predictor**, $X_i^\top \beta$
- They also share many characteristics, e.g. the form of the score and Hessian functions
- In fact, many are special cases of the **generalized linear model (GLM)**
- Here, we provide a general treatment of GLMs to study those models more systematically
- 3 components of a GLM
 - 1 Systematic component: $X_i^\top \beta$
 - Must be a linear function of X_i
 - 2 Random component: $f(Y; \theta, \phi)$
 - Must be in the **exponential family**
 - θ is called the canonical parameter
 - ϕ : is called the dispersion parameter
 - 3 Link function: $g(\mu_i) = X_i^\top \beta$ where $\mu_i = E(Y_i | X_i)$
 - Must be monotonic and differentiable wrt μ_i

Generalized Linear Models

- Many of the models we have learned so far all assume that Y_i is a (stochastic) function of the **linear predictor**, $X_i^\top \beta$
- They also share many characteristics, e.g. the form of the score and Hessian functions
- In fact, many are special cases of the **generalized linear model (GLM)**
- Here, we provide a general treatment of GLMs to study those models more systematically
- 3 components of a GLM
 - 1 Systematic component: $X_i^\top \beta$
 - Must be a linear function of X_i
 - 2 Random component: $f(Y; \theta, \phi)$
 - Must be in the **exponential family**
 - θ is called the canonical parameter
 - ϕ is called the dispersion parameter
 - 3 Link function: $g(\mu_i) = X_i^\top \beta$ where $\mu_i = E(Y_i | X_i)$
 - Must be monotonic and differentiable wrt μ_i

Generalized Linear Models

- Many of the models we have learned so far all assume that Y_i is a (stochastic) function of the **linear predictor**, $X_i^\top \beta$
- They also share many characteristics, e.g. the form of the score and Hessian functions
- In fact, many are special cases of the **generalized linear model (GLM)**
- Here, we provide a general treatment of GLMs to study those models more systematically
- 3 components of a GLM
 - 1 Systematic component: $X_i^\top \beta$
 - Must be a linear function of X_i
 - 2 Random component: $f(Y; \theta, \phi)$
 - Must be in the **exponential family**
 - θ is called the canonical parameter
 - ϕ is called the dispersion parameter
 - 3 Link function: $g(\mu_i) = X_i^\top \beta$ where $\mu_i = E(Y_i | X_i)$
 - Must be monotonic and differentiable wrt μ_i

Generalized Linear Models

- Many of the models we have learned so far all assume that Y_i is a (stochastic) function of the **linear predictor**, $X_i^\top \beta$
- They also share many characteristics, e.g. the form of the score and Hessian functions
- In fact, many are special cases of the **generalized linear model (GLM)**
- Here, we provide a general treatment of GLMs to study those models more systematically
- 3 components of a GLM
 - 1 Systematic component: $X_i^\top \beta$
 - Must be a linear function of X_i
 - 2 Random component: $f(Y; \theta, \phi)$
 - Must be in the **exponential family**
 - θ is called the canonical parameter
 - ϕ : is called the dispersion parameter
 - 3 Link function: $g(\mu_i) = X_i^\top \beta$ where $\mu_i = E(Y_i | X_i)$
 - Must be monotonic and differentiable wrt μ_i

Exponential Family of Distributions

Any distribution in the exponential family has the density of the following form:

$$f(y|\theta, \phi) = \exp \left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}$$

Example: Poisson(λ):

$$\Pr(Y_i = y \mid \lambda) = \frac{\exp(-\lambda)\lambda^y}{y!} = \exp \{y \log \lambda - \exp(\log \lambda) - \log y!\}$$

$\implies \theta = \log \lambda$, $\phi = 1$, $a(\phi) = \phi$, $b(\theta) = \exp(\theta)$, and $c = -\log y!$

Exponential Family of Distributions

Any distribution in the exponential family has the density of the following form:

$$f(y|\theta, \phi) = \exp \left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}$$

Example: Poisson(λ):

$$\Pr(Y_i = y \mid \lambda) = \frac{\exp(-\lambda)\lambda^y}{y!} = \exp \{y \log \lambda - \exp(\log \lambda) - \log y!\}$$

$\implies \theta = \log \lambda, \phi = 1, a(\phi) = \phi, b(\theta) = \exp(\theta), \text{ and } c = -\log y!$

Exponential Family of Distributions

Any distribution in the exponential family has the density of the following form:

$$f(y|\theta, \phi) = \exp \left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}$$

Example: Poisson(λ):

$$\Pr(Y_i = y \mid \lambda) = \frac{\exp(-\lambda)\lambda^y}{y!} = \exp \{y \log \lambda - \exp(\log \lambda) - \log y!\}$$

$\implies \theta = \log \lambda, \phi = 1, a(\phi) = \phi, b(\theta) = \exp(\theta), \text{ and } c = -\log y!$

Exponential Family of Distributions

Any distribution in the exponential family has the density of the following form:

$$f(y|\theta, \phi) = \exp \left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}$$

Example: Poisson(λ):

$$\Pr(Y_i = y \mid \lambda) = \frac{\exp(-\lambda)\lambda^y}{y!} = \exp \{y \log \lambda - \exp(\log \lambda) - \log y!\}$$

$\implies \theta = \log \lambda, \phi = 1, a(\phi) = \phi, b(\theta) = \exp(\theta), \text{ and } c = -\log y!$

Exponential Family of Distributions

Any distribution in the exponential family has the density of the following form:

$$f(y|\theta, \phi) = \exp \left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}$$

Example: Poisson(λ):

$$\Pr(Y_i = y \mid \lambda) = \frac{\exp(-\lambda)\lambda^y}{y!} = \exp \{y \log \lambda - \exp(\log \lambda) - \log y!\}$$

$\implies \theta = \log \lambda$, $\phi = 1$, $a(\phi) = \phi$, $b(\theta) = \exp(\theta)$, and $c = -\log y!$

$$f(y_i|\theta_i, \phi) = \exp \left\{ \frac{y\theta_i - b(\theta)}{a(\phi)} + c(y, \phi) \right\}$$

Example: Bernoulli(π_i):

$$\begin{aligned} \Pr(Y_i = y_i|\pi_i) &= \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \\ &= \exp \left[\frac{y_i\theta_i - \log[1 + \exp(\theta_i)]}{1} - 0 \right] \end{aligned}$$

where:

$$\begin{aligned} \theta_i &= \text{logit}(\pi_i) = \log \left(\frac{\pi_i}{1 - \pi_i} \right) \\ b(\theta_i) &= \log[1 + \exp(\theta_i)] \\ a(\phi) &= 1; c(y_i, \phi) = 0 \end{aligned}$$

$$f(y_i|\theta_i, \phi) = \exp \left\{ \frac{y\theta_i - b(\theta)}{a(\phi)} + c(y, \phi) \right\}$$

Example: Bernoulli(π_i):

$$\begin{aligned} \Pr(Y_i = y_i|\pi_i) &= \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \\ &= \exp \left[\frac{y_i\theta_i - \log[1 + \exp(\theta_i)]}{1} - 0 \right] \end{aligned}$$

where:

$$\begin{aligned} \theta_i &= \text{logit}(\pi_i) = \log \left(\frac{\pi_i}{1 - \pi_i} \right) \\ b(\theta_i) &= \log[1 + \exp(\theta_i)] \\ a(\phi) &= 1; c(y_i, \phi) = 0 \end{aligned}$$

$$f(y_i|\theta_i, \phi) = \exp \left\{ \frac{y\theta_i - b(\theta)}{a(\phi)} + c(y, \phi) \right\}$$

Example: Bernoulli(π_i):

$$\begin{aligned} \Pr(Y_i = y_i|\pi_i) &= \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \\ &= \exp \left[\frac{y_i\theta_i - \log[1 + \exp(\theta_i)]}{1} - 0 \right] \end{aligned}$$

where:

$$\begin{aligned} \theta_i &= \text{logit}(\pi_i) = \log \left(\frac{\pi_i}{1 - \pi_i} \right) \\ b(\theta_i) &= \log[1 + \exp(\theta_i)] \\ a(\phi) &= 1; c(y_i, \phi) = 0 \end{aligned}$$

$$f(y_i|\theta_i, \phi) = \exp \left\{ \frac{y\theta_i - b(\theta)}{a(\phi)} + c(y, \phi) \right\}$$

Example: Bernoulli(π_i):

$$\begin{aligned} \Pr(Y_i = y_i|\pi_i) &= \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \\ &= \exp \left[\frac{y_i\theta_i - \log[1 + \exp(\theta_i)]}{1} - 0 \right] \end{aligned}$$

where:

$$\begin{aligned} \theta_i &= \text{logit}(\pi_i) = \log \left(\frac{\pi_i}{1 - \pi_i} \right) \\ b(\theta_i) &= \log[1 + \exp(\theta_i)] \\ a(\phi) &= 1; c(y_i, \phi) = 0 \end{aligned}$$

Properties of the Exponential Family

- Mean is a function of θ and given by

$$E(Y) \equiv \mu = b'(\theta)$$

- Variance is a function of θ and ϕ and given by

$$\text{Var}(Y) \equiv V = b''(\theta)a(\phi)$$

- Common forms of $a(\phi)$: 1 (Poisson, Bernoulli), ϕ (normal, Gamma), and ϕ/ω_i (binomial)
- $b''(\theta)$ is called the **variance function**

Properties of the Exponential Family

- Mean is a function of θ and given by

$$E(Y) \equiv \mu = b'(\theta)$$

- Variance is a function of θ and ϕ and given by

$$\text{Var}(Y) \equiv V = b''(\theta)a(\phi)$$

- Common forms of $a(\phi)$: 1 (Poisson, Bernoulli), ϕ (normal, Gamma), and ϕ/ω_i (binomial)
- $b''(\theta)$ is called the **variance function**

Properties of the Exponential Family

- Mean is a function of θ and given by

$$E(Y) \equiv \mu = b'(\theta)$$

- Variance is a function of θ and ϕ and given by

$$\text{Var}(Y) \equiv V = b''(\theta)a(\phi)$$

- Common forms of $a(\phi)$: 1 (Poisson, Bernoulli), ϕ (normal, Gamma), and ϕ/ω_i (binomial)
- $b''(\theta)$ is called the variance function

Properties of the Exponential Family

- Mean is a function of θ and given by

$$E(Y) \equiv \mu = b'(\theta)$$

- Variance is a function of θ and ϕ and given by

$$\text{Var}(Y) \equiv V = b''(\theta)a(\phi)$$

- Common forms of $a(\phi)$: 1 (Poisson, Bernoulli), ϕ (normal, Gamma), and ϕ/ω_i (binomial)
- $b''(\theta)$ is called the **variance function**

Properties of the Exponential Family

- Mean is a function of θ and given by

$$E(Y) \equiv \mu = b'(\theta)$$

- Variance is a function of θ and ϕ and given by

$$\text{Var}(Y) \equiv V = b''(\theta)a(\phi)$$

- Common forms of $a(\phi)$: 1 (Poisson, Bernoulli), ϕ (normal, Gamma), and ϕ/ω_i (binomial)
- $b''(\theta)$ is called the **variance function**

- In the Poisson model, $\theta_i = \log \lambda_i$, $a(\phi) = 1$ and $b(\theta_i) = \exp(\theta_i)$

$$\Rightarrow E(Y_i) = \frac{\partial b(\theta_i)}{\partial \theta_i} = \exp(\theta_i) = \lambda_i \text{ and}$$

$$\text{Var}(Y_i) = \frac{\partial^2 b(\theta_i)}{\partial \theta_i^2} = \exp(\theta_i) = \lambda_i$$

- In the Bernoulli model, $\theta_i = \text{logit}(\pi_i)$, $a(\phi) = 1$ and $b(\theta_i) = \log[1 + \exp(\theta_i)]$.

\Rightarrow

$$E(Y_i) = \frac{\partial b(\theta_i)}{\partial \theta_i} = \frac{\exp(\theta_i)}{1 + \exp(\theta_i)} = \text{logit}^{-1}(\theta_i) = \text{logit}^{-1}(\text{logit}(\pi_i)) = \pi_i$$

$$\text{Var}(Y_i) = \frac{\partial^2 b(\theta_i)}{\partial \theta_i^2} = \frac{\exp(\theta_i)(1 + \exp(\theta_i)) - \exp(\theta_i)\exp(\theta_i)}{(1 + \exp(\theta_i))^2} =$$

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

- In the Poisson model, $\theta_i = \log \lambda_i$, $a(\phi) = 1$ and $b(\theta_i) = \exp(\theta_i)$

$$\Rightarrow E(Y_i) = \frac{\partial b(\theta_i)}{\partial \theta_i} = \exp(\theta_i) = \lambda_i \text{ and}$$

$$\text{Var}(Y_i) = \frac{\partial^2 b(\theta_i)}{\partial \theta_i^2} = \exp(\theta_i) = \lambda_i$$

- In the Bernoulli model, $\theta_i = \text{logit}(\pi_i)$, $a(\phi) = 1$ and $b(\theta_i) = \log[1 + \exp(\theta_i)]$.

\Rightarrow

$$E(Y_i) = \frac{\partial b(\theta_i)}{\partial \theta_i} = \frac{\exp(\theta_i)}{1 + \exp(\theta_i)} = \text{logit}^{-1}(\theta_i) = \text{logit}^{-1}(\text{logit}(\pi_i)) = \pi_i$$

$$\text{Var}(Y_i) = \frac{\partial^2 b(\theta_i)}{\partial \theta_i^2} = \frac{\exp(\theta_i)(1 + \exp(\theta_i)) - \exp(\theta_i)\exp(\theta_i)}{(1 + \exp(\theta_i))^2} =$$

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

- In the Poisson model, $\theta_i = \log \lambda_i$, $a(\phi) = 1$ and $b(\theta_i) = \exp(\theta_i)$

$$\Rightarrow E(Y_i) = \frac{\partial b(\theta_i)}{\partial \theta_i} = \exp(\theta_i) = \lambda_i \text{ and}$$

$$\text{Var}(Y_i) = \frac{\partial^2 b(\theta_i)}{\partial \theta_i^2} = \exp(\theta_i) = \lambda_i$$

- In the Bernoulli model, $\theta_i = \text{logit}(\pi_i)$, $a(\phi) = 1$ and $b(\theta_i) = \log[1 + \exp(\theta_i)]$.

\Rightarrow

$$E(Y_i) = \frac{\partial b(\theta_i)}{\partial \theta_i} = \frac{\exp(\theta_i)}{1 + \exp(\theta_i)} = \text{logit}^{-1}(\theta_i) = \text{logit}^{-1}(\text{logit}(\pi_i)) = \pi_i$$

$$\text{Var}(Y_i) = \frac{\partial^2 b(\theta_i)}{\partial \theta_i^2} = \frac{\exp(\theta_i)(1 + \exp(\theta_i)) - \exp(\theta_i) \exp(\theta_i)}{(1 + \exp(\theta_i))^2} =$$

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

Link Functions

- **Link function:** $g(\mu_i) = X_i^\top \beta$
- Defines the relationship between $X_i^\top \beta$ and the mean μ_i
- Must map the real line onto the possible range of μ_i
- Recall that $\mu_i = b'(\theta_i)$
- Therefore, θ_i is always a (often simple) function of $X_i^\top \beta$
- In particular, when $\theta_i = X_i^\top \beta$, the link is called the **canonical link**

- In Poisson, $\theta_i = \log(\lambda_i) = \log(\exp(X_i' \beta)) = X_i' \beta$

→ $\exp^{-1} = \log$ is the canonical link function

- In Bernoulli $\theta_i = \text{logit}(\pi_i); \pi_i = \frac{\exp(X_i' \beta)}{1 + \exp(X_i' \beta)}$

$$\theta_i = \text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \log\left(\frac{\frac{\exp(X_i \beta)}{1 + \exp(X_i \beta)}}{\frac{1}{1 + \exp(X_i \beta)}}\right) = \log(\exp(X_i \beta)) = X_i \beta$$

- Must be monotonic and differentiable
- This allows us to express the **mean function** as: $\mu_i = g^{-1}(X_i^\top \beta)$

Link Functions

- **Link function:** $g(\mu_i) = X_i^\top \beta$
- Defines the relationship between $X_i^\top \beta$ and the mean μ_i
- Must map the real line onto the possible range of μ_i
- Recall that $\mu_i = b'(\theta_i)$
- Therefore, θ_i is always a (often simple) function of $X_i^\top \beta$
- In particular, when $\theta_i = X_i^\top \beta$, the link is called the **canonical link**

- In Poisson, $\theta_i = \log(\lambda_i) = \log(\exp(X_i^\top \beta)) = X_i^\top \beta$

→ $\exp^{-1} = \log$ is the canonical link function

- In Bernoulli $\theta_i = \text{logit}(\pi_i); \pi_i = \frac{\exp(X_i^\top \beta)}{1 + \exp(X_i^\top \beta)}$

$$\theta_i = \text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \log\left(\frac{\frac{\exp(X_i^\top \beta)}{1 + \exp(X_i^\top \beta)}}{\frac{1}{1 + \exp(X_i^\top \beta)}}\right) = \log(\exp(X_i^\top \beta)) = X_i^\top \beta$$

- Must be monotonic and differentiable
- This allows us to express the **mean function** as: $\mu_i = g^{-1}(X_i^\top \beta)$

Link Functions

- **Link function:** $g(\mu_i) = X_i^\top \beta$
- Defines the relationship between $X_i^\top \beta$ and the mean μ_i
- Must map the real line onto the possible range of μ_i
- Recall that $\mu_i = b'(\theta_i)$
- Therefore, θ_i is always a (often simple) function of $X_i^\top \beta$
- In particular, when $\theta_i = X_i^\top \beta$, the link is called the **canonical link**

- In Poisson, $\theta_i = \log(\lambda_i) = \log(\exp(X_i' \beta)) = X_i' \beta$

→ $\exp^{-1} = \log$ is the canonical link function

- In Bernoulli $\theta_i = \text{logit}(\pi_i); \pi_i = \frac{\exp(X_i' \beta)}{1 + \exp(X_i' \beta)}$

$$\theta_i = \text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \log\left(\frac{\frac{\exp(X_i \beta)}{1 + \exp(X_i \beta)}}{\frac{1}{1 + \exp(X_i \beta)}}\right) = \log(\exp(X_i \beta)) = X_i \beta$$

- Must be monotonic and differentiable
- This allows us to express the **mean function** as: $\mu_i = g^{-1}(X_i^\top \beta)$

Link Functions

- **Link function:** $g(\mu_i) = X_i^\top \beta$
- Defines the relationship between $X_i^\top \beta$ and the mean μ_i
- Must map the real line onto the possible range of μ_i
- Recall that $\mu_i = b'(\theta_i)$
- Therefore, θ_i is always a (often simple) function of $X_i^\top \beta$
- In particular, when $\theta_i = X_i^\top \beta$, the link is called the **canonical link**

- In Poisson, $\theta_i = \log(\lambda_i) = \log(\exp(X_i' \beta)) = X_i' \beta$

→ $\exp^{-1} = \log$ is the canonical link function

- In Bernoulli $\theta_i = \text{logit}(\pi_i); \pi_i = \frac{\exp(X_i' \beta)}{1 + \exp(X_i' \beta)}$

$$\theta_i = \text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \log\left(\frac{\frac{\exp(X_i' \beta)}{1 + \exp(X_i' \beta)}}{\frac{1}{1 + \exp(X_i' \beta)}}\right) = \log(\exp(X_i' \beta)) = X_i' \beta$$

- Must be monotonic and differentiable
- This allows us to express the **mean function** as: $\mu_i = g^{-1}(X_i^\top \beta)$

Link Functions

- **Link function:** $g(\mu_i) = X_i^\top \beta$
- Defines the relationship between $X_i^\top \beta$ and the mean μ_i
- Must map the real line onto the possible range of μ_i
- Recall that $\mu_i = b'(\theta_i)$
- Therefore, θ_i is always a (often simple) function of $X_i^\top \beta$
- In particular, when $\theta_i = X_i^\top \beta$, the link is called the **canonical link**

- In Poisson, $\theta_i = \log(\lambda_i) = \log(\exp(X_i' \beta)) = X_i' \beta$
 $\longrightarrow \exp^{-1} = \log$ is the canonical link function

- In Bernoulli $\theta_i = \text{logit}(\pi_i); \pi_i = \frac{\exp(X_i' \beta)}{1 + \exp(X_i' \beta)}$

$$\theta_i = \text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \log\left(\frac{\frac{\exp(X_i' \beta)}{1 + \exp(X_i' \beta)}}{\frac{1}{1 + \exp(X_i' \beta)}}\right) = \log(\exp(X_i' \beta)) = X_i' \beta$$

- Must be monotonic and differentiable
- This allows us to express the **mean function** as: $\mu_i = g^{-1}(X_i^\top \beta)$

Link Functions

- **Link function:** $g(\mu_i) = X_i^\top \beta$
- Defines the relationship between $X_i^\top \beta$ and the mean μ_i
- Must map the real line onto the possible range of μ_i
- Recall that $\mu_i = b'(\theta_i)$
- Therefore, θ_i is always a (often simple) function of $X_i^\top \beta$
- In particular, when $\theta_i = X_i^\top \beta$, the link is called the **canonical link**
 - In Poisson, $\theta_i = \log(\lambda_i) = \log(\exp(X_i' \beta)) = X_i' \beta$
 $\longrightarrow \exp^{-1} = \log$ is the canonical link function
 - In Bernoulli $\theta_i = \text{logit}(\pi_i); \pi_i = \frac{\exp(X_i' \beta)}{1 + \exp(X_i' \beta)}$

$$\theta_i = \text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \log\left(\frac{\frac{\exp(X_i' \beta)}{1 + \exp(X_i' \beta)}}{\frac{1}{1 + \exp(X_i' \beta)}}\right) = \log(\exp(X_i' \beta)) = X_i' \beta$$

- Must be monotonic and differentiable
- This allows us to express the **mean function** as: $\mu_i = g^{-1}(X_i^\top \beta)$

Link Functions

- **Link function:** $g(\mu_i) = X_i^\top \beta$
- Defines the relationship between $X_i^\top \beta$ and the mean μ_i
- Must map the real line onto the possible range of μ_i
- Recall that $\mu_i = b'(\theta_i)$
- Therefore, θ_i is always a (often simple) function of $X_i^\top \beta$
- In particular, when $\theta_i = X_i^\top \beta$, the link is called the **canonical link**

- In Poisson, $\theta_i = \log(\lambda_i) = \log(\exp(X_i' \beta)) = X_i' \beta$
 $\longrightarrow \exp^{-1} = \log$ is the canonical link function

- In Bernoulli $\theta_i = \text{logit}(\pi_i); \pi_i = \frac{\exp(X_i' \beta)}{1 + \exp(X_i' \beta)}$

$$\theta_i = \text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \log\left(\frac{\frac{\exp(X_i \beta)}{1 + \exp(X_i \beta)}}{\frac{1}{1 + \exp(X_i \beta)}}\right) = \log(\exp(X_i \beta)) = X_i \beta$$

- Must be monotonic and differentiable
- This allows us to express the **mean function** as: $\mu_i = g^{-1}(X_i^\top \beta)$

Link Functions

- **Link function:** $g(\mu_i) = X_i^\top \beta$
- Defines the relationship between $X_i^\top \beta$ and the mean μ_i
- Must map the real line onto the possible range of μ_i
- Recall that $\mu_i = b'(\theta_i)$
- Therefore, θ_i is always a (often simple) function of $X_i^\top \beta$
- In particular, when $\theta_i = X_i^\top \beta$, the link is called the **canonical link**

- In Poisson, $\theta_i = \log(\lambda_i) = \log(\exp(X_i' \beta)) = X_i' \beta$
 $\longrightarrow \exp^{-1} = \log$ is the canonical link function

- In Bernoulli $\theta_i = \text{logit}(\pi_i); \pi_i = \frac{\exp(X_i' \beta)}{1 + \exp(X_i' \beta)}$

$$\theta_i = \text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \log\left(\frac{\frac{\exp(X_i \beta)}{1 + \exp(X_i \beta)}}{\frac{1}{1 + \exp(X_i \beta)}}\right) = \log(\exp(X_i \beta)) = X_i \beta$$

- Must be monotonic and differentiable
- This allows us to express the **mean function** as: $\mu_i = g^{-1}(X_i^\top \beta)$

Likelihood Function, Score, and Information Matrix

■ Log-likelihood function:

$$l_n(\theta, \phi; Y) = \sum_{i=1}^n \left\{ \frac{Y_i \theta_i - b(\theta_i)}{a(\phi)} + c(Y_i, \phi) \right\}$$

■ Recall our notation: $\mu_i = E[Y_i | X_i]$ and $V_i = \text{Var}[Y_i | X_i]$

■ Score:

$$s(\beta) = \frac{\partial l_n(\theta, \phi; Y)}{\partial \beta} = \sum_{i=1}^n \frac{Y_i - b'(\theta_i)}{a(\phi)b''(\theta_i)} \frac{\partial \mu_i}{\partial \beta} = \sum_{i=1}^n \frac{Y_i - \mu_i}{V_i} \left(\frac{\partial \mu_i}{\partial X_i^\top \beta} \right) X_i$$

■ Information:

$$I(\beta) = -E[H(\beta)] = -E \left(\frac{\partial^2 l_n(\theta, \phi; Y)}{\partial \beta \partial \beta^\top} \right) = \sum_{i=1}^n \frac{1}{V_i} \left(\frac{\partial \mu_i}{\partial X_i^\top \beta} \right)^2 X_i X_i^\top$$

■ Exercise: Check these hold for logit, probit, Poisson, etc.!

Likelihood Function, Score, and Information Matrix

■ Log-likelihood function:

$$l_n(\theta, \phi; Y) = \sum_{i=1}^n \left\{ \frac{Y_i \theta_i - b(\theta_i)}{a(\phi)} + c(Y_i, \phi) \right\}$$

■ Recall our notation: $\mu_i = E[Y_i | X_i]$ and $V_i = \text{Var}[Y_i | X_i]$

■ Score:

$$s(\beta) = \frac{\partial l_n(\theta, \phi; Y)}{\partial \beta} = \sum_{i=1}^n \frac{Y_i - b'(\theta_i)}{a(\phi)b''(\theta_i)} \frac{\partial \mu_i}{\partial \beta} = \sum_{i=1}^n \frac{Y_i - \mu_i}{V_i} \left(\frac{\partial \mu_i}{\partial X_i^\top \beta} \right) X_i$$

■ Information:

$$I(\beta) = -E[H(\beta)] = -E \left(\frac{\partial^2 l_n(\theta, \phi; Y)}{\partial \beta \partial \beta^\top} \right) = \sum_{i=1}^n \frac{1}{V_i} \left(\frac{\partial \mu_i}{\partial X_i^\top \beta} \right)^2 X_i X_i^\top$$

■ Exercise: Check these hold for logit, probit, Poisson, etc.!

Likelihood Function, Score, and Information Matrix

■ Log-likelihood function:

$$l_n(\theta, \phi; Y) = \sum_{i=1}^n \left\{ \frac{Y_i \theta_i - b(\theta_i)}{a(\phi)} + c(Y_i, \phi) \right\}$$

■ Recall our notation: $\mu_i = E[Y_i | X_i]$ and $V_i = \text{Var}[Y_i | X_i]$

■ Score:

$$s(\beta) = \frac{\partial l_n(\theta, \phi; Y)}{\partial \beta} = \sum_{i=1}^n \frac{Y_i - b'(\theta_i)}{a(\phi)b''(\theta_i)} \frac{\partial \mu_i}{\partial \beta} = \sum_{i=1}^n \frac{Y_i - \mu_i}{V_i} \left(\frac{\partial \mu_i}{\partial X_i^\top \beta} \right) X_i$$

■ Information:

$$I(\beta) = -E[H(\beta)] = -E \left(\frac{\partial^2 l_n(\theta, \phi; Y)}{\partial \beta \partial \beta^\top} \right) = \sum_{i=1}^n \frac{1}{V_i} \left(\frac{\partial \mu_i}{\partial X_i^\top \beta} \right)^2 X_i X_i^\top$$

■ Exercise: Check these hold for logit, probit, Poisson, etc.!

Likelihood Function, Score, and Information Matrix

■ Log-likelihood function:

$$l_n(\theta, \phi; Y) = \sum_{i=1}^n \left\{ \frac{Y_i \theta_i - b(\theta_i)}{a(\phi)} + c(Y_i, \phi) \right\}$$

■ Recall our notation: $\mu_i = E[Y_i | X_i]$ and $V_i = \text{Var}[Y_i | X_i]$

■ Score:

$$s(\beta) = \frac{\partial l_n(\theta, \phi; Y)}{\partial \beta} = \sum_{i=1}^n \frac{Y_i - b'(\theta_i)}{a(\phi)b''(\theta_i)} \frac{\partial \mu_i}{\partial \beta} = \sum_{i=1}^n \frac{Y_i - \mu_i}{V_i} \left(\frac{\partial \mu_i}{\partial X_i^\top \beta} \right) X_i$$

■ Information:

$$I(\beta) = -E[H(\beta)] = -E \left(\frac{\partial^2 l_n(\theta, \phi; Y)}{\partial \beta \partial \beta^\top} \right) = \sum_{i=1}^n \frac{1}{V_i} \left(\frac{\partial \mu_i}{\partial X_i^\top \beta} \right)^2 X_i X_i^\top$$

■ Exercise: Check these hold for logit, probit, Poisson, etc.!

Likelihood Function, Score, and Information Matrix

- Log-likelihood function:

$$l_n(\theta, \phi; Y) = \sum_{i=1}^n \left\{ \frac{Y_i \theta_i - b(\theta_i)}{a(\phi)} + c(Y_i, \phi) \right\}$$

- Recall our notation: $\mu_i = E[Y_i | X_i]$ and $V_i = \text{Var}[Y_i | X_i]$

- Score:

$$s(\beta) = \frac{\partial l_n(\theta, \phi; Y)}{\partial \beta} = \sum_{i=1}^n \frac{Y_i - b'(\theta_i)}{a(\phi)b''(\theta_i)} \frac{\partial \mu_i}{\partial \beta} = \sum_{i=1}^n \frac{Y_i - \mu_i}{V_i} \left(\frac{\partial \mu_i}{\partial X_i^\top \beta} \right) X_i$$

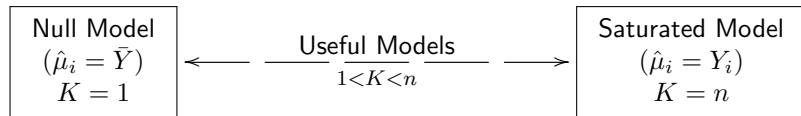
- Information:

$$I(\beta) = -E[H(\beta)] = -E \left(\frac{\partial^2 l_n(\theta, \phi; Y)}{\partial \beta \partial \beta^\top} \right) = \sum_{i=1}^n \frac{1}{V_i} \left(\frac{\partial \mu_i}{\partial X_i^\top \beta} \right)^2 X_i X_i^\top$$

- Exercise: Check these hold for logit, probit, Poisson, etc.!

Assessing Goodness of Fit for GLMs

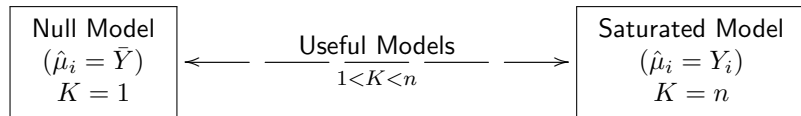
- Null model: Predict every observation by sample mean \bar{Y}
 \Rightarrow one parameter, maximum data reduction
- **Saturated model**: Predict every observation by its own value Y_i
 $\Rightarrow n$ parameters, no data reduction
- A useful model sits somewhere in between



- Goodness of fit can be measured by comparing the model likelihood to the saturated model likelihood

Assessing Goodness of Fit for GLMs

- Null model: Predict every observation by sample mean \bar{Y}
 \Rightarrow one parameter, maximum data reduction
- **Saturated model**: Predict every observation by its own value Y_i
 $\Rightarrow n$ parameters, no data reduction
- A useful model sits somewhere in between



- Goodness of fit can be measured by comparing the model likelihood to the saturated model likelihood

Analysis of Deviance

- **Scaled Deviance** (Unscaled: $\phi \times$ Deviance $D(Y, \hat{\theta})$):

$$\begin{aligned} D^*(Y; \hat{\theta}) &\equiv 2\{l_n(\tilde{\theta}; Y, \phi) - l_n(\hat{\theta}; Y, \phi)\} \\ &= 2 \sum_{i=1}^n \left\{ Y_i(\tilde{\theta}_i - \hat{\theta}_i) - (b(\tilde{\theta}_i) - b(\hat{\theta}_i)) \right\} / a(\phi) \end{aligned}$$

where $\begin{cases} \hat{\theta}_i = \theta(\hat{\mu}_i) & \text{(estimate from the model of interest)} \\ \tilde{\theta}_i = \theta(Y_i) & \text{("estimate" from the saturated model)} \end{cases}$

- Note that D^* is a likelihood-ratio statistic
- This implies $D^*(Y; \hat{\theta}) \stackrel{approx.}{\sim} \chi_{n-k}^2$ if the model fits the data well
- We can also compare models: $D_1^* - D_2^* \sim \chi_{k_1 - k_2}^2$ (LR test)
- McFadden's **pseudo- R^2** :

$$\tilde{R}^2 = \frac{l_n(\hat{\theta}; Y, \phi) - l_n(\theta(\bar{Y}); Y, \phi)}{l_n(\tilde{\theta}; Y, \phi) - l_n(\theta(\bar{Y}); Y, \phi)} = 1 - \frac{D^*(Y; \hat{\theta})}{D^*(Y; \theta(\bar{Y}))}$$

where $\bar{\theta}_i = \theta(\bar{Y})$ for all i

Analysis of Deviance

- **Scaled Deviance** (Unscaled: $\phi \times$ Deviance $D(Y, \hat{\theta})$):

$$\begin{aligned} D^*(Y; \hat{\theta}) &\equiv 2\{l_n(\tilde{\theta}; Y, \phi) - l_n(\hat{\theta}; Y, \phi)\} \\ &= 2 \sum_{i=1}^n \left\{ Y_i(\tilde{\theta}_i - \hat{\theta}_i) - (b(\tilde{\theta}_i) - b(\hat{\theta}_i)) \right\} / a(\phi) \end{aligned}$$

where $\begin{cases} \hat{\theta}_i = \theta(\hat{\mu}_i) & \text{(estimate from the model of interest)} \\ \tilde{\theta}_i = \theta(Y_i) & \text{("estimate" from the saturated model)} \end{cases}$

- Note that D^* is a likelihood-ratio statistic
- This implies $D^*(Y; \hat{\theta}) \stackrel{approx.}{\sim} \chi_{n-k}^2$ if the model fits the data well
- We can also compare models: $D_1^* - D_2^* \sim \chi_{k_1 - k_2}^2$ (LR test)
- McFadden's **pseudo- R^2** :

$$\tilde{R}^2 = \frac{l_n(\hat{\theta}; Y, \phi) - l_n(\theta(\bar{Y}); Y, \phi)}{l_n(\tilde{\theta}; Y, \phi) - l_n(\theta(\bar{Y}); Y, \phi)} = 1 - \frac{D^*(Y; \hat{\theta})}{D^*(Y; \theta(\bar{Y}))}$$

where $\bar{\theta}_i = \theta(\bar{Y})$ for all i

Analysis of Deviance

- **Scaled Deviance** (Unscaled: $\phi \times$ Deviance $D(Y, \hat{\theta})$):

$$\begin{aligned} D^*(Y; \hat{\theta}) &\equiv 2\{l_n(\tilde{\theta}; Y, \phi) - l_n(\hat{\theta}; Y, \phi)\} \\ &= 2 \sum_{i=1}^n \left\{ Y_i(\tilde{\theta}_i - \hat{\theta}_i) - (b(\tilde{\theta}_i) - b(\hat{\theta}_i)) \right\} / a(\phi) \end{aligned}$$

where $\begin{cases} \hat{\theta}_i = \theta(\hat{\mu}_i) & \text{(estimate from the model of interest)} \\ \tilde{\theta}_i = \theta(Y_i) & \text{("estimate" from the saturated model)} \end{cases}$

- Note that D^* is a likelihood-ratio statistic
- This implies $D^*(Y; \hat{\theta}) \stackrel{approx.}{\sim} \chi_{n-k}^2$ if the model fits the data well
- We can also compare models: $D_1^* - D_2^* \sim \chi_{k_1 - k_2}^2$ (LR test)
- McFadden's **pseudo- R^2** :

$$\tilde{R}^2 = \frac{l_n(\hat{\theta}; Y, \phi) - l_n(\theta(\bar{Y}); Y, \phi)}{l_n(\tilde{\theta}; Y, \phi) - l_n(\theta(\bar{Y}); Y, \phi)} = 1 - \frac{D^*(Y; \hat{\theta})}{D^*(Y; \theta(\bar{Y}))}$$

where $\bar{\theta}_i = \theta(\bar{Y})$ for all i

Analysis of Deviance

- **Scaled Deviance** (Unscaled: $\phi \times$ Deviance $D(Y, \hat{\theta})$):

$$\begin{aligned} D^*(Y; \hat{\theta}) &\equiv 2\{l_n(\tilde{\theta}; Y, \phi) - l_n(\hat{\theta}; Y, \phi)\} \\ &= 2 \sum_{i=1}^n \left\{ Y_i(\tilde{\theta}_i - \hat{\theta}_i) - (b(\tilde{\theta}_i) - b(\hat{\theta}_i)) \right\} / a(\phi) \end{aligned}$$

where $\begin{cases} \hat{\theta}_i = \theta(\hat{\mu}_i) & \text{(estimate from the model of interest)} \\ \tilde{\theta}_i = \theta(Y_i) & \text{("estimate" from the saturated model)} \end{cases}$

- Note that D^* is a likelihood-ratio statistic
- This implies $D^*(Y; \hat{\theta}) \stackrel{approx.}{\sim} \chi_{n-k}^2$ if the model fits the data well
- We can also compare models: $D_1^* - D_2^* \sim \chi_{k_1 - k_2}^2$ (LR test)
- McFadden's **pseudo- R^2** :

$$\tilde{R}^2 = \frac{l_n(\hat{\theta}; Y, \phi) - l_n(\theta(\bar{Y}); Y, \phi)}{l_n(\tilde{\theta}; Y, \phi) - l_n(\theta(\bar{Y}); Y, \phi)} = 1 - \frac{D^*(Y; \hat{\theta})}{D^*(Y; \theta(\bar{Y}))}$$

where $\bar{\theta}_i = \theta(\bar{Y})$ for all i

Unscaled Deviance for Normal GLMs

$$\hat{\theta}_i = \mathbf{X}_i' \boldsymbol{\beta}$$

$$b(\hat{\theta}_i) = \hat{\theta}_i^2 / 2$$

$$\text{Saturated model} \Rightarrow \tilde{\theta}_i = y_i; b(\tilde{\theta}_i) = y_i^2 / 2$$

$$\text{Deviance} = 2 \sum_{i=1}^N [y_i(y_i - \hat{\mu}_i) - y_i^2 / 2 + \hat{\mu}_i^2 / 2] = \sum_{i=1}^N (y_i - \hat{\mu}_i)^2$$

Deviance for Poisson

Saturated model $\tilde{\lambda}_i = y_i$. This implies a log-likelihood of:

$$\begin{aligned}\tilde{\theta}_i &= \log y_i \\ b(\tilde{\theta})_i &= \exp(\log y_i) = y_i \\ \text{Deviance} &= 2 \sum_{i=1}^N \left[y_i \log \frac{y_i}{\hat{\lambda}_i} - y_i + \hat{\lambda}_i \right]\end{aligned}$$

Residuals and Model Checking

- In normal linear regression, $D = \mathcal{X}^2 = \sum_{i=1}^n \hat{\epsilon}_i^2 = RSS$
- This suggests the following generalization of residuals for GLM:

1 Deviance residual:

$$\hat{\epsilon}_i^D \equiv \text{sign}(Y_i - \hat{\mu}_i) \sqrt{d_i}$$

where d_i is the deviance for the i th observation

2 Pearson residual:

$$\hat{\epsilon}_i^P \equiv \frac{\sqrt{\omega_i}(Y_i - \hat{\mu}_i)}{\sqrt{b''(\hat{\theta}_i)}}$$

- These residuals have approximately the same properties as OLS residuals when N is large
- Thus, most regression diagnostics for linear models also work for GLMs:
 - Plotting standardized and studentized residuals
 - Analyze influence points and outliers
 - Added-variable plots, component-residual plots, etc.

Residuals and Model Checking

- In normal linear regression, $D = \mathcal{X}^2 = \sum_{i=1}^n \hat{\epsilon}_i^2 = RSS$
- This suggests the following generalization of residuals for GLM:

1 Deviance residual:

$$\hat{\epsilon}_i^D \equiv \text{sign}(Y_i - \hat{\mu}_i) \sqrt{d_i}$$

where d_i is the deviance for the i th observation

2 Pearson residual:

$$\hat{\epsilon}_i^P \equiv \frac{\sqrt{\omega_i}(Y_i - \hat{\mu}_i)}{\sqrt{b''(\hat{\theta}_i)}}$$

- These residuals have approximately the same properties as OLS residuals when N is large
- Thus, most **regression diagnostics** for linear models also work for GLMs:
 - Plotting standardized and studentized residuals
 - Analyze influence points and outliers
 - Added-variable plots, component-residual plots, etc.

Residuals and Model Checking

- In normal linear regression, $D = \mathcal{X}^2 = \sum_{i=1}^n \hat{\epsilon}_i^2 = RSS$
- This suggests the following generalization of residuals for GLM:

1 Deviance residual:

$$\hat{\epsilon}_i^D \equiv \text{sign}(Y_i - \hat{\mu}_i) \sqrt{d_i}$$

where d_i is the deviance for the i th observation

2 Pearson residual:

$$\hat{\epsilon}_i^P \equiv \frac{\sqrt{\omega_i}(Y_i - \hat{\mu}_i)}{\sqrt{b''(\hat{\theta}_i)}}$$

- These residuals have approximately the same properties as OLS residuals when N is large
- Thus, most **regression diagnostics** for linear models also work for GLMs:
 - Plotting standardized and studentized residuals
 - Analyze influence points and outliers
 - Added-variable plots, component-residual plots, etc.

Residuals and Model Checking

- In normal linear regression, $D = \mathcal{X}^2 = \sum_{i=1}^n \hat{\epsilon}_i^2 = RSS$
- This suggests the following generalization of residuals for GLM:

1 Deviance residual:

$$\hat{\epsilon}_i^D \equiv \text{sign}(Y_i - \hat{\mu}_i) \sqrt{d_i}$$

where d_i is the deviance for the i th observation

2 Pearson residual:

$$\hat{\epsilon}_i^P \equiv \frac{\sqrt{\omega_i}(Y_i - \hat{\mu}_i)}{\sqrt{b''(\hat{\theta}_i)}}$$

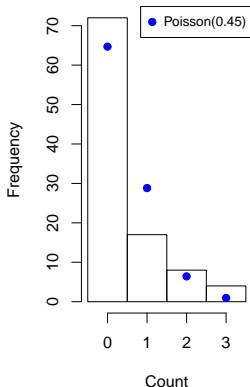
- These residuals have approximately the same properties as OLS residuals when N is large
- Thus, most **regression diagnostics** for linear models also work for GLMs:
 - Plotting standardized and studentized residuals
 - Analyze influence points and outliers
 - Added-variable plots, component-residual plots, etc.

Example: Democracy and War Involvement

Benoit (1996):

- Y_i : # of involvement in international wars, 1960–80
- X_i : democracy (Freedom House score), population, military capacity, economic interdependence

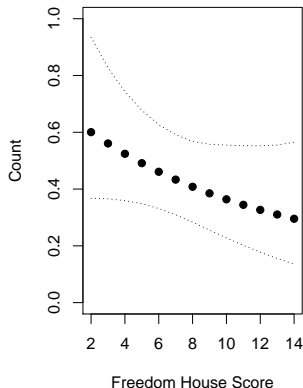
N. of Wars, 1960–80



Coefficients:

	Est.	s.e.	
(Int.)	-3.97	1.62	*
fh73	-0.06	0.04	
lpopln70	0.62	0.30	*
lmilwp70	1.29	0.45	**
ecintdep	-1.28	1.11	
	z	p	
	-2.45	0.014	
	-1.49	0.136	
	2.07	0.039	
	2.85	0.004	
	-1.16	0.247	

Estimated Mean Count of War

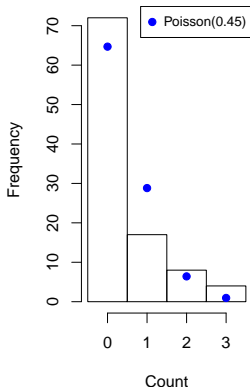


Example: Democracy and War Involvement

Benoit (1996):

- Y_i : # of involvement in international wars, 1960–80
- X_i : democracy (Freedom House score), population, military capacity, economic interdependence

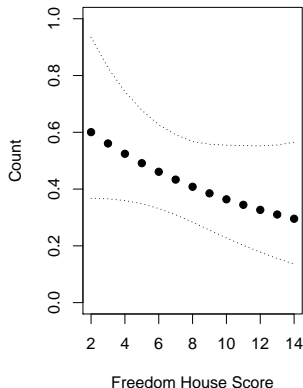
N. of Wars, 1960–80



Coefficients:

	Est.	s.e.	
(Int.)	-3.97	1.62	*
fh73	-0.06	0.04	
lpopln70	0.62	0.30	*
lmilwp70	1.29	0.45	**
ecintdep	-1.28	1.11	
	z	p	
	-2.45	0.014	
	-1.49	0.136	
	2.07	0.039	
	2.85	0.004	
	-1.16	0.247	

Estimated Mean Count of War

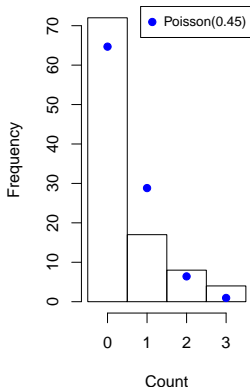


Example: Democracy and War Involvement

Benoit (1996):

- Y_i : # of involvement in international wars, 1960–80
- X_i : democracy (Freedom House score), population, military capacity, economic interdependence

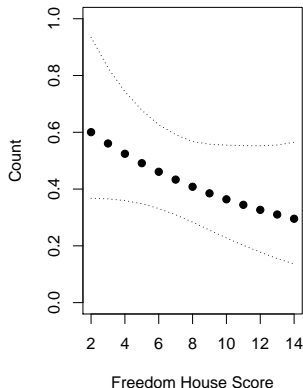
N. of Wars, 1960–80



Coefficients:

	Est.	s.e.	
(Int.)	-3.97	1.62	*
fh73	-0.06	0.04	
lpopln70	0.62	0.30	*
lmilwp70	1.29	0.45	**
ecintdep	-1.28	1.11	
	z	p	
	-2.45	0.014	
	-1.49	0.136	
	2.07	0.039	
	2.85	0.004	
	-1.16	0.247	

Estimated Mean Count of War

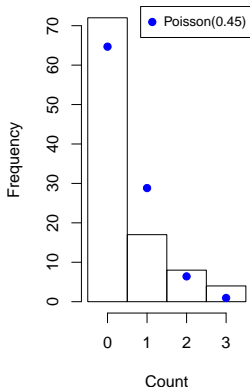


Example: Democracy and War Involvement

Benoit (1996):

- Y_i : # of involvement in international wars, 1960–80
- X_i : democracy (Freedom House score), population, military capacity, economic interdependence

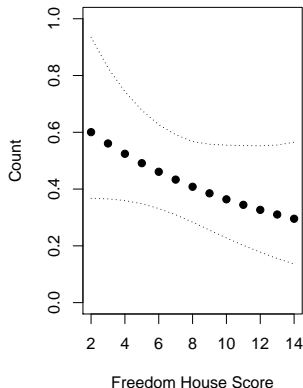
N. of Wars, 1960–80



Coefficients:

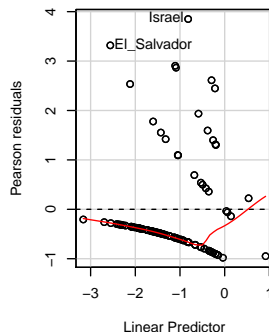
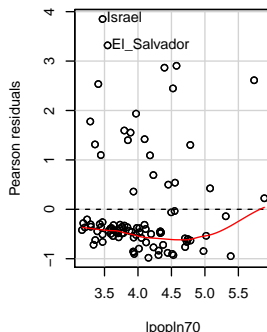
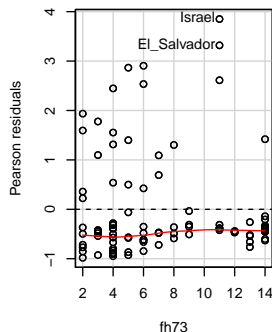
	Est.	s.e.	
(Int.)	-3.97	1.62	*
fh73	-0.06	0.04	
lpopln70	0.62	0.30	*
lmilwp70	1.29	0.45	**
ecintdep	-1.28	1.11	
	z	p	
	-2.45	0.014	
	-1.49	0.136	
	2.07	0.039	
	2.85	0.004	
	-1.16	0.247	

Estimated Mean Count of War



Example: Democracy and War Involvement

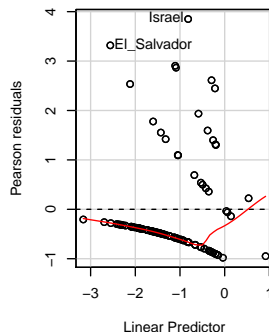
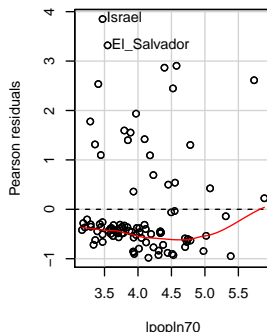
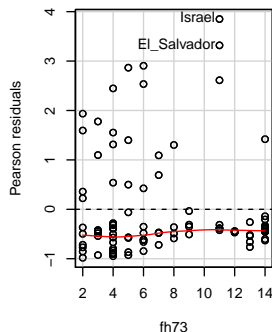
Plotting $\hat{\epsilon}_i^P$ against X_{ij} and $X_i^\top \hat{\beta}$:



- Evidence of mild nonlinearity (quadratic) for log population
- Heavy skew to the right:
 - Nonnegativity of the outcome variable + small sample size
 - Potentially alleviated by a “zero inflation” model

Example: Democracy and War Involvement

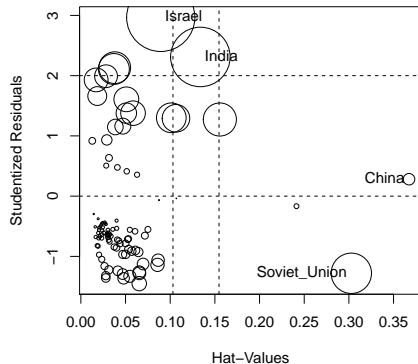
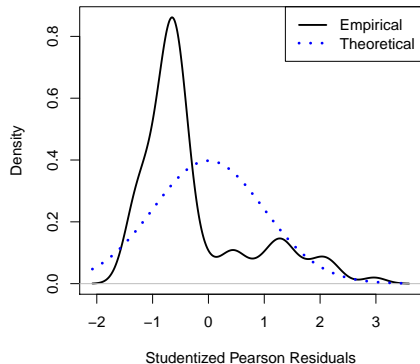
Plotting $\hat{\epsilon}_i^P$ against X_{ij} and $X_i^\top \hat{\beta}$:



- Evidence of mild nonlinearity (quadratic) for log population
- Heavy skew to the right:
 - Nonnegativity of the outcome variable + small sample size
 - Potentially alleviated by a “zero inflation” model

Example: Democracy and War Involvement

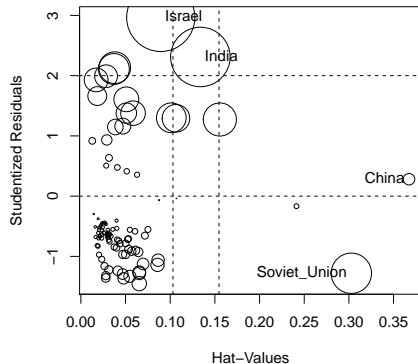
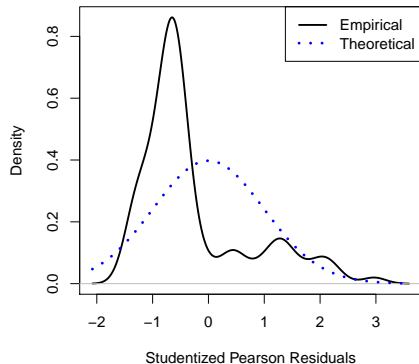
Studentized residuals and hat values:



- The density of studentized residuals confirms the right skew
- No obvious outliers (except perhaps Israel)

Example: Democracy and War Involvement

Studentized residuals and hat values:



- The density of studentized residuals confirms the right skew
- No obvious outliers (except perhaps Israel)

Model fit