# Assessing Model Fit

# Announcements

- There seems to be an issue with Race in Visit 10 – Hispanic/Latinx individuals have no data in this visit??

Project Announcements from Sakai

1. Screening data should be included in the merged dataset. Education should be used from here
2. Smoking status should be taken from Visit 10 and should be binary.
3. The three variables you choose should come from Visit 10 OR not cause major numerical issues.
4. Everyone in your group will present a part of your presentations.
5. **Proposal is due Monday, May 29th at 11pm.**

# Class 14 Learning Objectives

1. Use test statistics of goodness-of-fit to determine if our preliminary final model fits the data well
   - Using Pearson residual statistic ($X^2$) ~ $\chi^2_{J-(p+1)}$
   - Using Hosmer and Lemeshow goodness-of-fit statistic ($\hat{C}$) ~ $\chi^2_{g-2}$

2. Apply ROC AUC to determine how well model predicts binary outcome

3. Apply AIC and BIC as a summary measure to make additional comparisons between potential models

# Overview (I)

- Once the preliminary final model has been determined, we need to assess the fit of the model

- Variable selection is no longer our focus at this stage
  - We want to find answer to whether the model fits the data adequately

- Assessing the Goodness of Fit or Assessing model fit
  - Assess how well our fitted logistic regression model predicts/estimates the observed outcomes
  - Comparison: fitted/estimated outcome vs. observed outcome

# Overview (II)

- The model building strategies we have discussed so far only assess the importance of covariates
  - It *did not* assess model fit
- Previous in model building, we made relative comparisons between models
  - Our conclusions were limited to: Model 1 (full model) fits data better than Model 2 (reduced model)
- Assessing goodness of fit is
  - Not a relative comparison
  - It is an **absolute** comparison
    - To compare the fitted model to the largest possible model (saturated model)
  - Model adequacy vs. Model comparison

# Poll Everywhere Question 1

relative comparisons

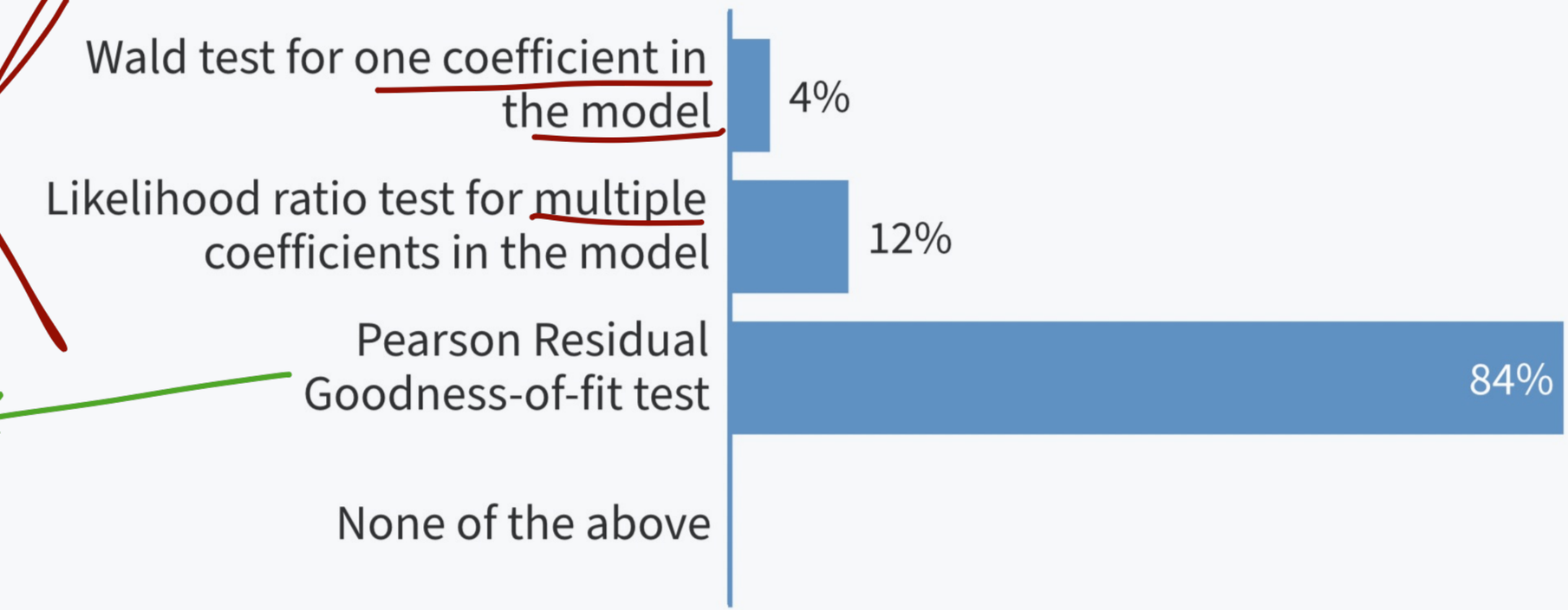
absolute comp.

Which of the following is an absolute comparison of a fitted model?

https://www.polleverywhere.com/multiple_choice_polls/9sf0DDvkVDhQgcXfCwTwC

Respond at **PollEv.com/nickywakim275**

## Which of the following is an absolute comparison of a fitted model?




Powered by Poll Everywhere

# Components to Assess Model Fit (II)

- The model fits the data if
  - Summary measures of the distance between the predicted/estimated/fitted and observed Y ~~are small~~
  - The contribution of each pair (predicted and observed) to these summary measures is unsystematic and is small relative to the error structure of the model
- It is possible to see a "good" summary measure of the distance between predicted and observed Y with some substantial deviation from fit for a few subjects

# Components to Assess Model Fit (I)

1. Computation and evaluation of **overall measures of fit**

2. Examination of the **individual components** of the summary statistics (often through picture)

3. Examination of other measures of the difference or **distance between the observed and fitted values**

# Components to Assess Model Fit (I)

1. Computation and evaluation of **overall measures of fit** TODAY

2. Examination of the **individual components** of the summary statistics (often through picture)

3. Examination of other measures of the difference or **distance between the observed and fitted values**

# Summary Measures of Goodness of Fit

- Aka overall measure of fit
- Need to define what the fitted outcome is
- Need to calculate how close the fitted outcome is to the observed outcome
- Summarize across all observations (or individuals' data)

# Class 14 Learning Objectives

1. Use test statistics of goodness-of-fit to determine if our preliminary final model fits the data well
   - Using Pearson residual statistic ($X^2$) ~ $\chi^2_{J-(p+1)}$
   - Using Hosmer and Lemeshow goodness-of-fit statistic ($\hat{C}$) ~ $\chi^2_{g-2}$
2. Apply ROC AUC to determine how well model predicts binary outcome
3. Apply AIC and BIC as a summary measure to make additional comparisons between potential models

# Comparing **fitted outcome** to observed outcome

- In logistic regression model, we estimate $\pi(\boldsymbol{x}) = \mathrm{P}(Y_i = 1|\boldsymbol{x})$
  - Predicted value, $\hat{\pi}(\boldsymbol{x})$, is between 0 and 1 for each subject
- However, we always observe $Y_i = 1$ or $Y_i = 0$
  - Not an observed $\pi(\boldsymbol{x})$
- We can determine the fitted outcome
  - $\hat{Y}_i \sim \mathrm{bernoulli}\left(\hat{\pi}(\boldsymbol{x})\right)$ → $p(1-p) = \hat{\pi}(\boldsymbol{x})(1-\hat{\pi}(\boldsymbol{x}))$
- If there are groups of individuals that share the same covariate observations, then we can use the same $\hat{\pi}(\boldsymbol{x})$
  - $\sum_j \hat{Y}_i \sim \mathrm{binomial}\left(\hat{\pi}_j(\boldsymbol{x})\right)$
- In stead of comparing the expected vs. observed at individual level, we can compare them at "group" level

# Number of Covariate Patterns (I)

- When the logistic regression model contains only categorical covariates, we can think of the number of covariate patterns
- **For example:** model contains two categorical covariates (sex assigned at birth and smoking status), there will be **4 unique combination** of these factors
  - This model has 4 ***covariate patterns***
  - Subjects can be divided into 4 groups based on the covariates' values
- We can then compute the predicted number of individuals with Y=1 in each group, and compare that with the actual observed number of individuals with Y=1 in that group

# Number of Covariate Patterns (II)

- Let $m_j$ **be the number of patients in jth covariate pattern**, j = 1,…, J, $\sum m_j = n$.
  ex = J = 4
- The fitted values in logistic regression are computed for each covariate pattern and depends on the estimated probability of the covariate pattern:

$$\hat{y}_j = m_j \hat{\pi}_j = m_j \frac{e^{\hat{g}(x_j)}}{1 + e^{\hat{g}(x_j)}}$$

# of pp in cov pat

# Pearson Residual (I)

- In logistic regression model, can use Pearson residual for summary measure of goodness-of-fit
  - Uses the $\hat{y}_j$ fitted value from previous slide

- Pearson residual for jth covariate pattern is:

$$r(y_j, \hat{\pi}_j) = \frac{(y_j - m_j\hat{\pi}_j)}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}} = \frac{(y_j - \hat{y}_i)}{\sqrt{\hat{y}_i(1-\hat{\pi}_j)}}$$

$y_j = \sum_j y_i$

$m_j \hat{\pi}_j$

- The summary statistics of Pearson residual is thus:

$$X^2 = \sum_{j=1}^{J} r(y_j, \hat{\pi}_j)^2$$

# Pearson Residual (II)

- $X^2$ is the Pearson's Chi-square statistic.
  - NOT the Pearson's Chi-square statistic we learnt for testing no association between two categorical variables
- A small $X^2$ (and thus a large P value from the test) suggests the model fits well

- There was a package that used to run this test (package was "LogisticDx") but now it does not work in my R
  - Still looking if there is function within R to run this calculation for us

# Issues with Pearson Residuals

- Assume current model has p covariates…
  - then $X^2$ (Pearson residual) follows $\mathrm{X}^2_{J-(p+1)}$
    - under the null hypothesis based on *large sample theory*
  - Appropriate if the **number of covariate patterns is less than the number of observations**

- When the logistic regression model contains one or more continuous covariates, it is likely that the **number of covariate patterns equals to the sample size n**

- We should **not use Pearson Residuals** to evaluate goodness-of-fit test **when the fitted model contains one or more continuous variables**

# Class 14 Learning Objectives

1. Use test statistics of goodness-of-fit to determine if our preliminary final model fits the data well
   - Using Pearson residual statistic $(X^2) \sim \chi^2_{J-(p+1)}$
   - Using Hosmer and Lemeshow goodness-of-fit statistic $(\hat{C}) \sim \chi^2_{g-2}$
2. Apply ROC AUC to determine how well model predicts binary outcome
3. Apply AIC and BIC as a summary measure to make additional comparisons between potential models

# The Hosmer-Lemeshow Test (I)

- If number of covariate patterns is roughly same as the number of observations
  - Whenever you include a continuous variable in your model
  - **Hosmer-Lemeshow (HL) goodness-of-fit test** should be used instead
- However, HL test does not work well if the number of covariate patterns is small
  - **HL test should not be used if the number of covariate patterns ≤ 6**
  - $X^2$ should be used when the number of covariate patterns is small
- A large p-value from HL test suggests the model fits well

# Poll Everywhere Question 2

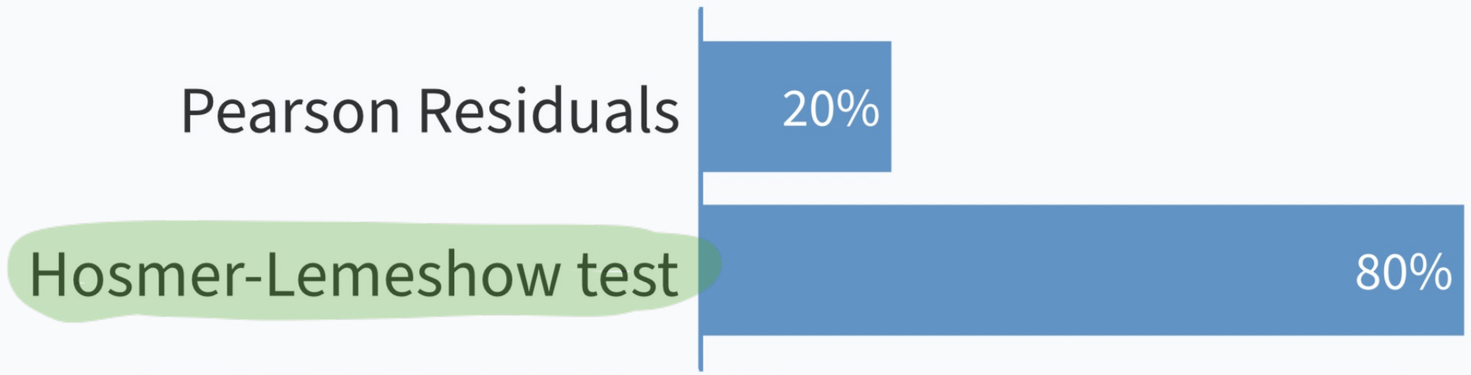
```
##
## Call:
## glm(formula = fracture ~ age + height + priorfrac + momfrac +
##     armassist + raterisk2 + age * priorfrac + momfrac * armassist,
##     family = binomial, data = glow2)
##
```

Which goodness of fit test should I use to test if the following model fits the data well for the preliminary GLOW model we found in class 12?

https://www.polleverywhere.com/multiple_choice_polls/jsgC6njR7u62JcfADn4GY

Respond at **PollEv.com/nickywakim275**

## Which goodness of fit test should I use to test if the following model fits the data well for the preliminary GLOW model we found in class 12?



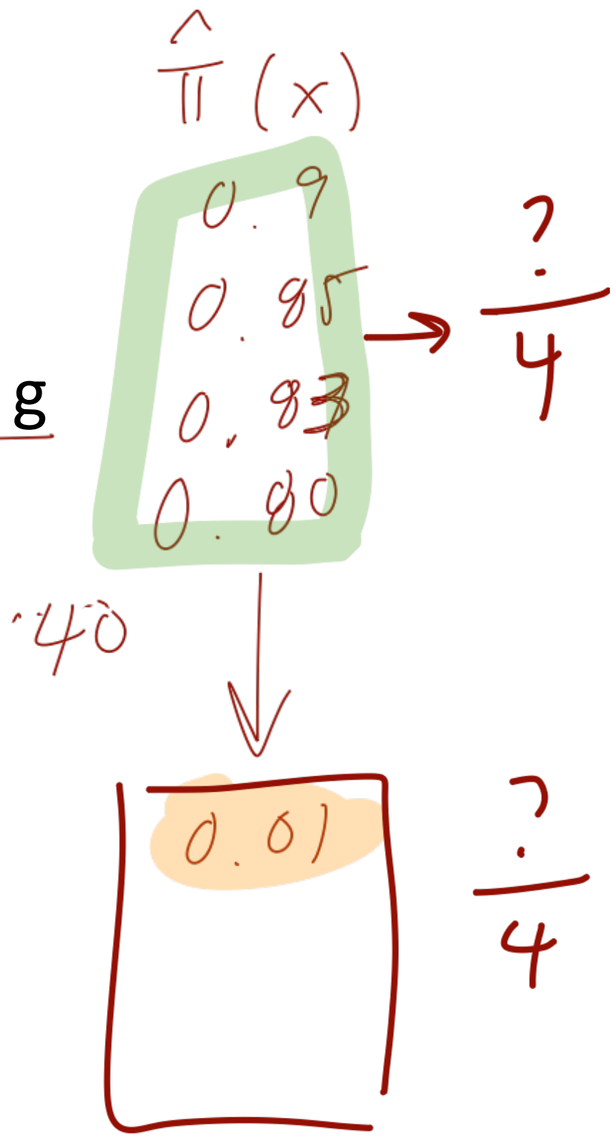

2 / 5 | Instructions | Responses | More | EXIT

Powered by Poll Everywhere

# The Hosmer-Lemeshow Test (II)

Steps to compute HL test statistic:

1. Compute estimated probability ($\hat{\pi}(x_i)$) for all n subjects ($i = 1, \ldots, n$)
2. Order $\hat{\pi}(x_i)$ from largest to smallest values
3. Divide ordered values into g percentile grouping (usually g = 10 based on H-L's suggestion)
4. Form table of observed and expected counts
5. Calculate HL test statistic from table
6. Compare HL test statistic to $\chi^2_{g-2}$



# The Hosmer-Lemeshow Test (III)

- The test statistic of Hosmer-Lemeshow goodness-of-fit test is denoted by $\hat{C}$, which is obtained by calculating the Pearson chi-squared statistic from the g X 2 table of observed and estimated expected frequencies

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n'_k \bar{\pi}_k)^2}{n'_k \bar{\pi}_k (1 - \bar{\pi}_k)}$$

where $n'_k$ is the total number of subjects in the kth group.

- Let $c_k$ be the number of covariate patterns in the kth decile, $o_k = \sum_{j=1}^{c_k} y_j$, and $\bar{\pi}_k = \sum_{j=1}^{c_k} \frac{m_j \hat{\pi}_j}{n'_k}$.

# The Hosmer-Lemeshow Test: key info

- Hypotheses:
  - $H_0$: model fits data well
  - $H_1$: model does not fit data well

- Test statistic for HL test

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n'_k \bar{\pi}_k)^2}{n'_k \bar{\pi}_k (1 - \bar{\pi}_k)}$$

- HL test statistic follows a chi-squared distribution

$$\hat{C} \sim \chi^2_{df = g-2}$$

# 10 minute break?

# GLOW Study

- From Class 12, at the end of Step 6, we had a preliminary final model

- Variables/interactions included:
  - Age
  - Height
  - Prior fracture
  - Mom fracture
  - Arm Assist
  - Rated risk of fracture
  - Age x Prior fracture
  - Mom fracture x Arm Assist

## Step 6

- **Preliminary final model** with fixed effects and interactions

```
final = int
summary(final)
```

```
##
## Call:
## glm(formula = fracture ~ age + height + priorfrac + momfrac +
##     armassist + raterisk2 + age * priorfrac + momfrac * armassist,
##     family = binomial, data = glow2)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.65767  -0.74860  -0.53179   0.06819   2.38111
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 1.71724    3.32173   0.517 0.605175
## age                         0.05731    0.01650   3.473 0.000515 ***
## height                     -0.04674    0.01833  -2.550 0.010781 *
## priorfracYes                4.61229    1.88018   2.453 0.014163 *
## momfracYes                  1.24664    0.39296   3.172 0.001512 **
## armassistYes                0.64416    0.25193   2.557 0.010561 *
## raterisk2Greater            0.46897    0.24078   1.948 0.051449 .
## age:priorfracYes           -0.05527    0.02593  -2.132 0.033044 *
## momfracYes:armassistYes -1.28054    0.62299  -2.055 0.039834 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

May 17, 2023 54

# GLOW Study: HL Test (I)

- Now we can move to Step 7 in model building
- Use HL test to help use determine if our preliminary final model fits the data well
- Hypotheses:
  - $H_0$: GLOW preliminary final model fits data well
  - $H_1$: GLOW preliminary final model does not fit data well

```
prelim_final = glm(fracture ~ age + height + priorfrac + momfrac + armassist + raterisk2 +
                   age*priorfrac + momfrac*armassist,
                      data = glow2, family = binomial)
summary(prelim_final)
```

```
##
## Call:
## glm(formula = fracture ~ age + height + priorfrac + momfrac +
##     armassist + raterisk2 + age * priorfrac + momfrac * armassist,
##     family = binomial, data = glow2)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.65767  -0.74860  -0.53179   0.06819   2.38111
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)                1.71724    3.32173   0.517 0.605175
## age                        0.05731    0.01650   3.473 0.000515 ***
## height                    -0.04674    0.01833  -2.550 0.010781 *
## priorfracYes               4.61229    1.88018   2.453 0.014163 *
## momfracYes                 1.24664    0.39296   3.172 0.001512 **
## armassistYes               0.64416    0.25193   2.557 0.010561 *
## raterisk2Greater           0.46897    0.24078   1.948 0.051449 .
## age:priorfracYes          -0.05527    0.02593  -2.132 0.033044 *
## momfracYes:armassistYes -1.28054    0.62299  -2.055 0.039834 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

# GLOW Study: HL Test (II)

- Use HL test in R
- Part of “ResourceSelection” package

```
glow2 = glow2 %>% mutate(frac_num = as.numeric(fracture)-1)

library(ResourceSelection)
```

```
## ResourceSelection 0.3-5   2019-07-22
```

```
hoslem.test(glow2$frac_num, fitted(prelim_final), g = 10)
```

observed outcome: make sure numeric

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  glow2$frac_num, fitted(prelim_final)
## X-squared = 6.3919, df = 8, p-value = 0.6034
```

# Poll Everywhere Question 3

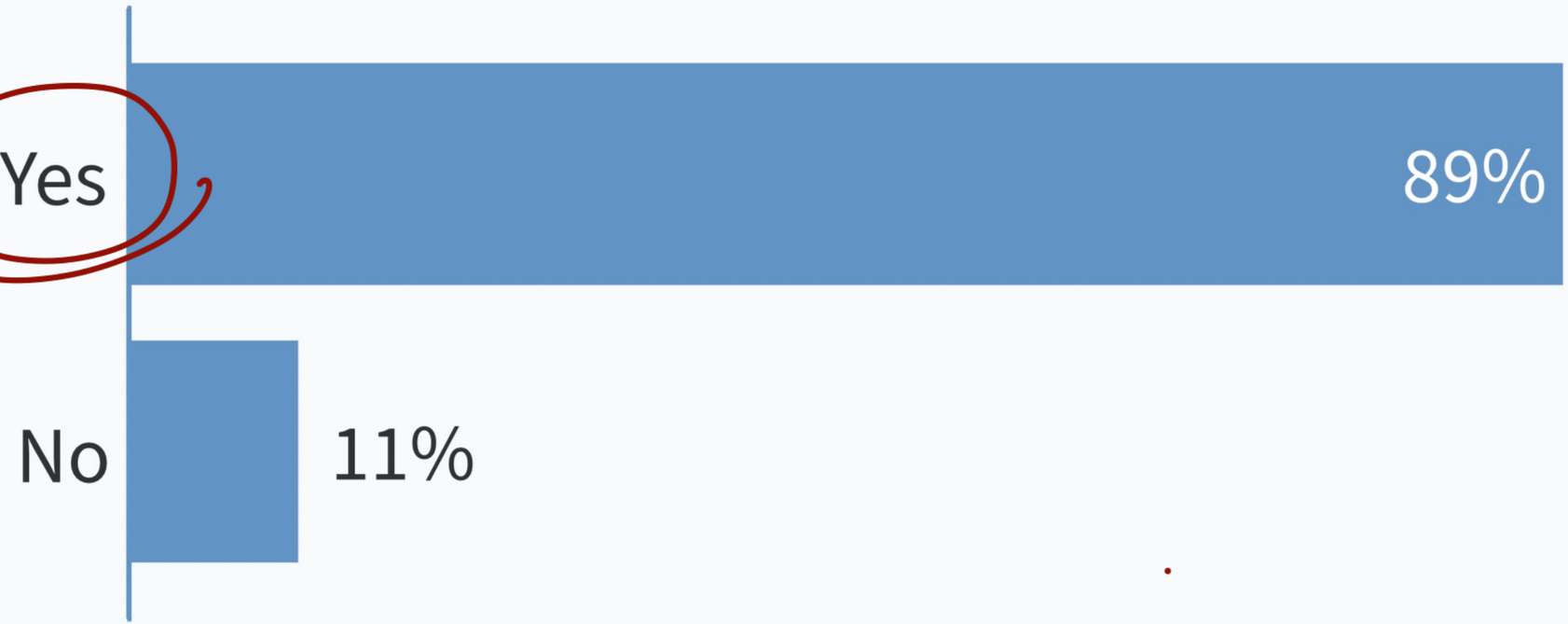
$H_0$: model fits data well

Are you confident that our selected model fits the GLOW study data well?

https://www.polleverywhere.com/multiple_choice_polls/yDlHkx8yaJ4cDomOxCLEF

Respond at **PollEv.com/nickywakim275**

## Are you confident that our selected model fits the GLOW study data well?




Powered by Poll Everywhere

# GLOW Study: HL Test (III)

- **Conclusion:** The p-value is 0.6034, so we fail to reject the null hypothesis that the model fits the data well. Thus, the preliminary final model for the GLOW dataset fits the data well

- Don’t forget that we still need to check individual observations

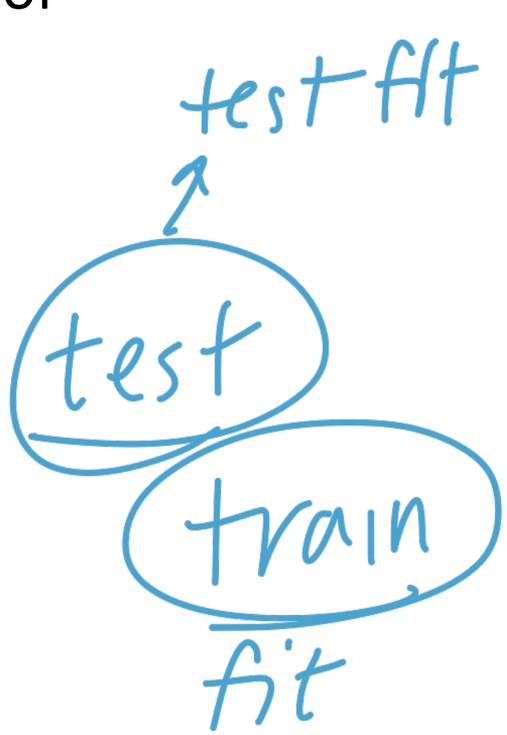
- R may give results for the HL test even if it is not appropriate to use it!
  - **If number of covariate patterns ≤ 6, do not use HL test**

# Big Data Issue in Goodness-of-fit Test

- When the sample size is too big (> 1000), it is much more likely to find the H-L reject the model fit (even when the expected vs. observed in each decile categories looks pretty similar)
- This is due to "too much" power in hypothesis testing.
  - Paul et al. (2012) for samples sizes from 1000 to 25,000, the number of groups $g$ should be equal to

$$g = \max\left(10, \min\left\{\frac{n_1}{2}, \frac{n-n_1}{2}, 2+8\left(\frac{n}{1000}\right)^2\right\}\right)$$

- For example, if one has a sample with $n = 10,000$ (sample size) and $n_1 = 1000$ (number of events) then $g = 500$ groups are suggested
- For n > 25000, other methods, such as partitioning data into a developmental data set (with smaller n) and a validation set



# Final Notes on Goodness-of-fit Test

- They should not be used for variable selection
  - The likelihood ratio tests for significance of coefficients are much more powerful and appropriate
- They are not for model comparison
  - One should not use the p-value from goodness of fit tests of different models to decide which model is better than the other
  - Something like AIC or BIC can be used

# Class 14 Learning Objectives

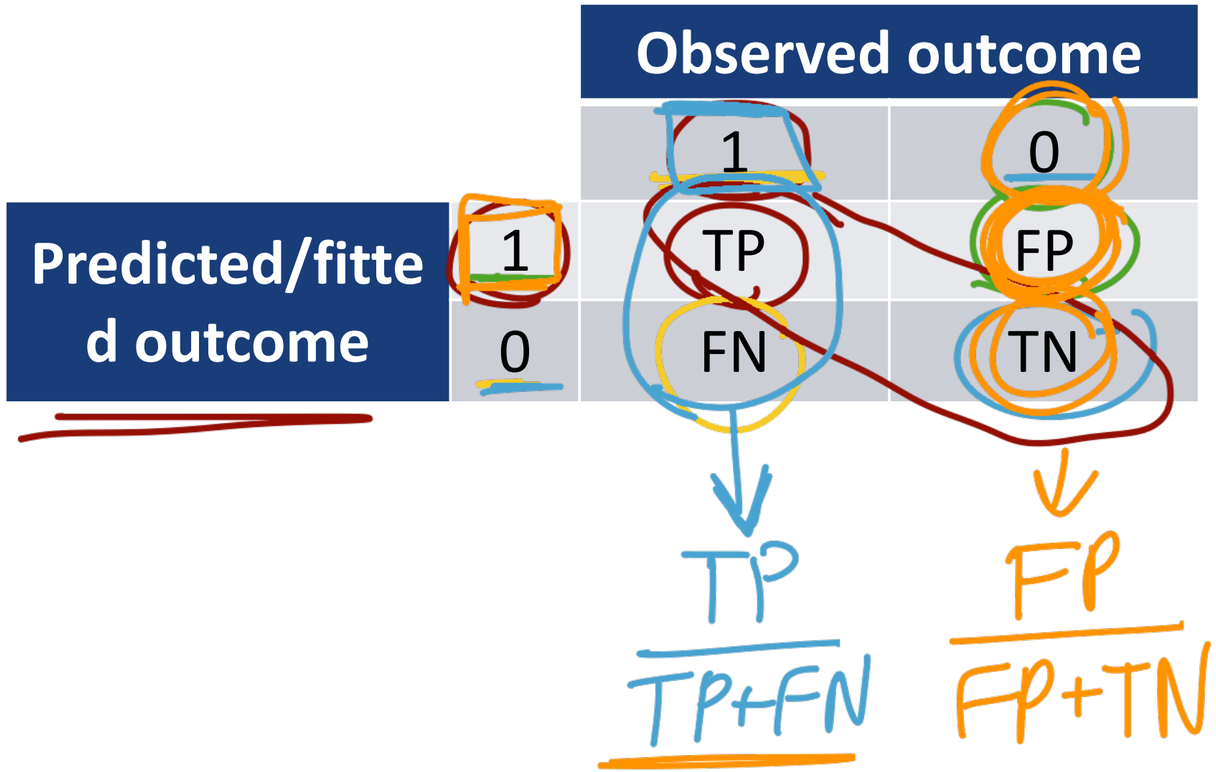
1. Use test statistics of goodness-of-fit to determine if our preliminary final model fits the data well
   - Using Pearson residual statistic ($X^2$) ~ $\chi^2_{J-(p+1)}$
   - Using Hosmer and Lemeshow goodness-of-fit statistic ($\hat{C}$) ~ $\chi^2_{g-2}$
2. Apply ROC AUC to determine how well model predicts binary outcome
3. Apply AIC and BIC as a summary measure to make additional comparisons between potential models

# ROC Curve and AUC (I)

- Receiver Operating Characteristics (ROC) curve is useful tool to quantify how good is our model predicting binary outcome

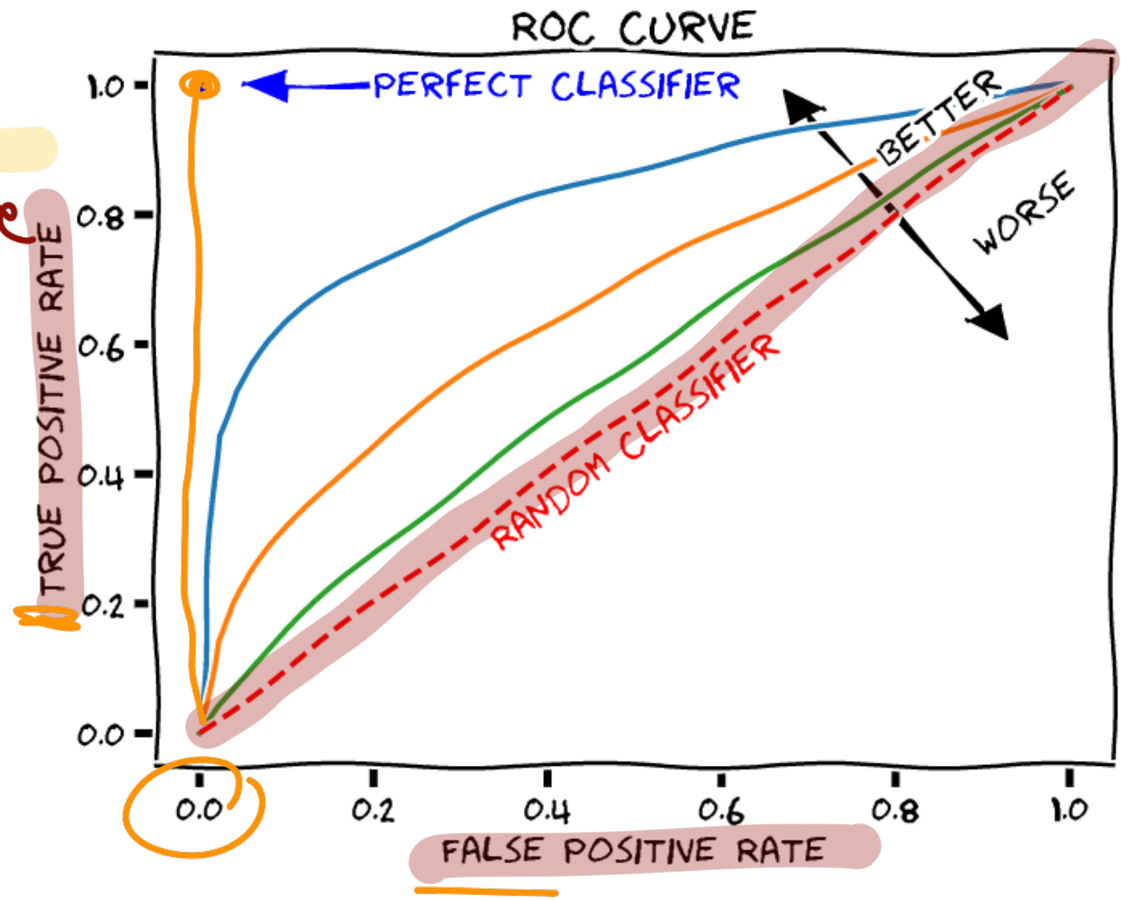
- It is a plot of sensitivity (true positive rate) versus (1-specificity) or false positive rate of fitted binary values
  - True Positive Rate $= \frac{TP}{TP+FN}$
  - False Positive Rate $= \frac{FP}{FP+TN}$

- The ROC curve shows the tradeoff between sensitivity and specificity



# ROC Curve and AUC (II)

- Area under the ROC curve (AUC ROC) is a reasonable summary of the overall ~~diagnostic~~ predictive accuracy of the test
  - Accuracy means how well the predicted value matches the observed value
- The closer the curve follows the left-hand border and top border of the ROC space, the more accurate the test
  - An AUC =1 represents 100% accuracy
- The closer the curve comes to the 45-degree diagonal line, the less accurate the test
  - An AUC = .5 represents a worthless test
    - Random predictions



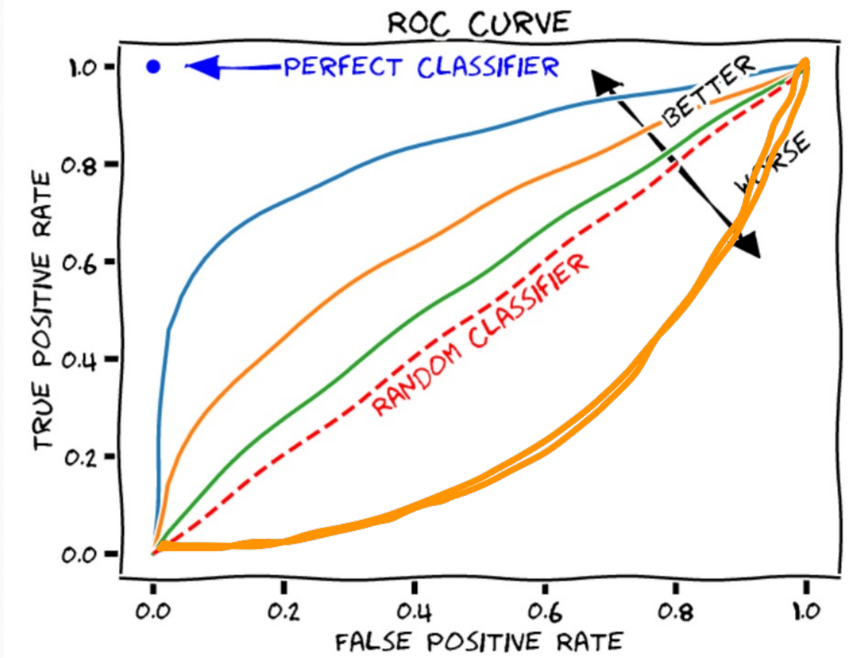
# Poll Everywhere Question 4

What does it mean if our ROC curve is below the random classifier line?

https://www.polleverywhere.com/free_text_polls/mVOI3EmlKoZOSTfE3dbnZ

Respond at **PollEv.com/nickywakim275**

## What does it mean if our ROC curve is below the random classifier line?




“ model doesn't predict better than chance ”

“ False positive is

Powered by Poll Everywhere

# ROC Curve and AUC (III)

- Often only report the AUC
- Textbook has some suggestions of how to interpret model fit through AUC values

| AUC Values | Fit |
|---|---|
| 0.5 | Uselesss |
| 0.5-0.7 | Poor |
| 0.7-0.8 | Acceptable |
| 0.8-0.9 | Excellent |
| 0.9-1 | Outstanding |

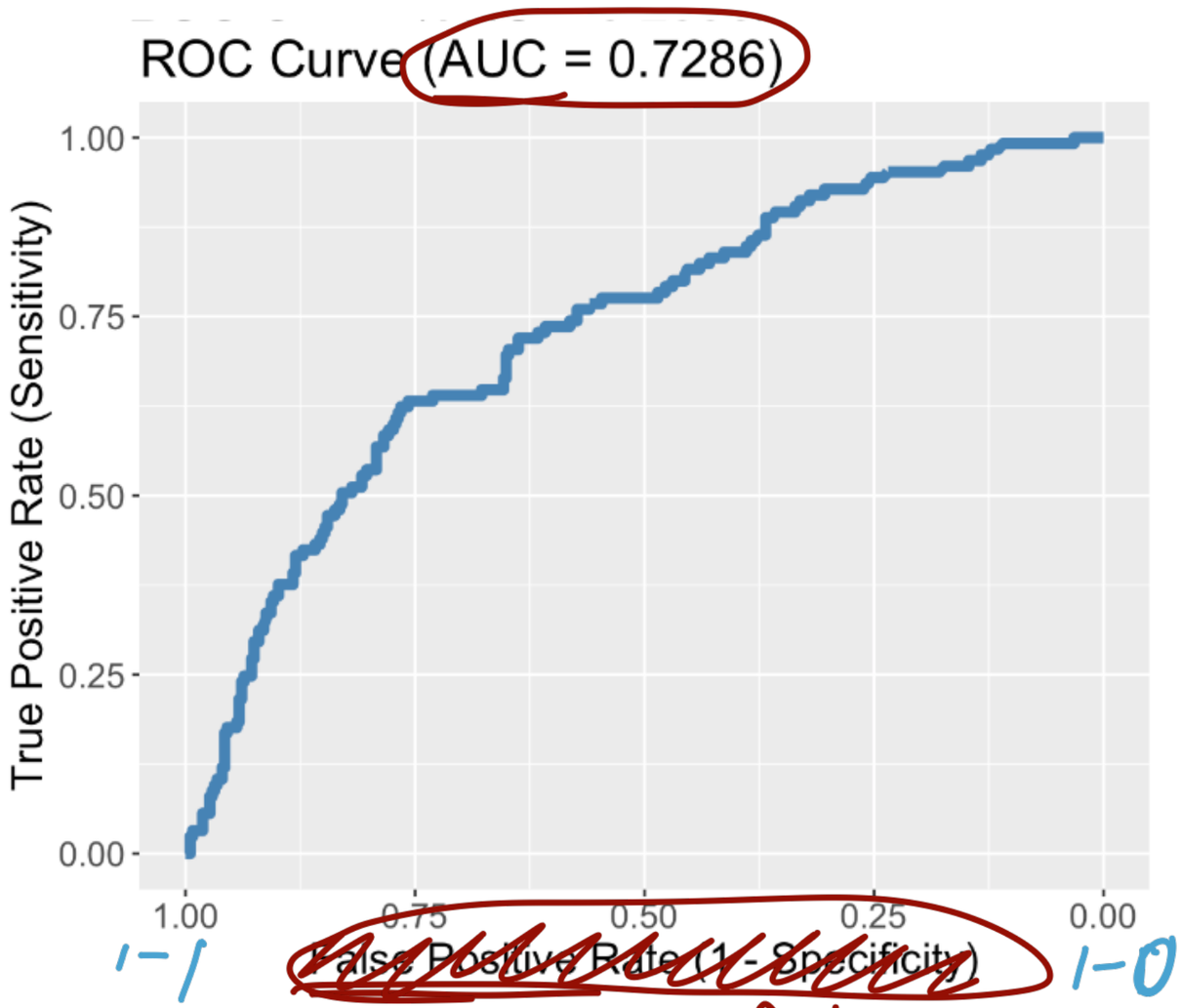
→ Random classifier

# GLOW Study: ROC of Preliminary Final Model

```r
library(ggplot2)
library(pROC)


predicted <- predict(prelim_final, glow2, type="response")

#define object to plot and calculate AUC
rocobj <- roc(glow2$fracture, predicted)
auc <- round(auc(glow2$fracture, predicted),4)

#create ROC plot
ggroc(rocobj, colour = 'steelblue', size = 2) +
  ggtitle(paste0('ROC Curve ', '(AUC = ', auc, ')')) +
  theme(text = element_text(size = 16)) +
  xlab("False Positive Rate (1 - Specificity)") +
  ylab("True Positive Rate (Sensitivity)")
```

observed

predicted




1-1

1-0

specificity

1-specificity

# Another way to think about AUC

- GLOW Study: Consider the situation in which the fracture status of each individual is known
- Randomly pick one individual from fractured group and one from non-fractured outcome group
  - Based on their age, height, prior fracture, and all other covariates, we will correctly predict which is from fractured group
- The AUC is the percentage of randomly drawn pairs for which we predict the pair correctly
- Therefore, AUC represents the ability of our covariates to discriminate between individuals with the outcome (fracture) and those without the outcome

# Class 14 Learning Objectives

1. Use test statistics of goodness-of-fit to determine if our preliminary final model fits the data well
   - Using Pearson residual statistic ($X^2$) ~ $\chi^2_{J-(p+1)}$
   - Using Hosmer and Lemeshow goodness-of-fit statistic ($\hat{C}$) ~ $\chi^2_{g-2}$
2. Apply ROC AUC to determine how well model predicts binary outcome
3. Apply AIC and BIC as a summary measure to make additional comparisons between potential models

# AIC and BIC (I)

- Two widely used non-hypothesis testing based measurements that helps select a good model
  - Akaike Information Criterion (AIC)
  - Bayesian Information Criterion (BIC)

- Unlike likelihood ratio test which is only suitable for nested model, AIC and BIC are suitable for both nested and non-nested model

- There is no hypothesis/conclusion testing for the comparison between two models
  - So not the best for selecting covariates to include in model
  - BUT helpful if you have a few preliminary final models that you want to compare

# Poll Everywhere Question 5

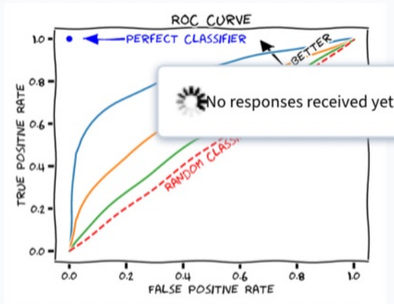
Can I compare these two models using AIC and BIC? $$logit(\pi(age_i)) = \beta_0 +\beta_1age_i$$ and $$logit(\pi(height_i)) = \beta_0 +\beta_1height_i$$

https://www.polleverywhere.com/multiple_choice_polls/ven54AVKtq8d2Ad5MiWvz

Class 14 | Share | Visual settings | Deactivate | Present

Respond at **PollEv.com/nickywakim275**

**What does it mean if our ROC curve is below the random classifier line?**




No responses received yet. They will appear here...

Powered by Poll Everywhere

4 / 5

Instructions | Responses

Clear responses

1. Configure | 2. Test | 3. Send

- How people can respond
- Audience restriction & identity
- Response settings
- Moderation

Edit | Response history | Delete

# AIC and BIC (II)

- Both AIC and BIC penalize a model for having many parameters

$$AIC = -2\log likelihood + 2q$$

$$BIC = -2loglikelihood + \mathrm{qlog(n)}$$

Where q is the number of parameters in the model and n is the sample size

- Both AIC and BIC can **only** be used to compare models fitting the **same data set**

- In comparing two models, **the model with smaller AIC and/or BIC is preferred**
  - When the difference in AIC between two models exceeds 3, the difference is viewed as “meaningful”

# AIC and BIC in R and SAS

- After fitting the logistic regression model, can calculate AIC and BIC

```
prelim_final = glm(fracture ~ age + height + priorfrac +
                     momfrac + armassist + raterisk2 +
                     age*priorfrac + momfrac*armassist,
                   data = glow2, family = binomial)
```

```
prelim_final2 = glm(fracture ~ age + height + priorfrac +
                      momfrac + armassist + raterisk2 +
                      age*height,
                    data = glow2, family = binomial)
```

```
AIC(prelim_final)
```

```
## [1] 518.4966
```

```
BIC(prelim_final)
```

```
## [1] 556.4281
```

```
AIC(cat_model)
```

```
## [1] 535.9395
```

```
BIC(cat_model)
```

```
## [1] 561.2272
```

# Class 14 Learning Objectives

1. Use test statistics of goodness-of-fit to determine if our preliminary final model fits the data well
   - Using Pearson residual statistic ($X^2$) ~ $\chi^2_{J-(p+1)}$
   - Using Hosmer and Lemeshow goodness-of-fit statistic ($\hat{C}$) ~ $\chi^2_{g-2}$

2. Apply ROC AUC to determine how well model predicts binary outcome

3. Apply AIC and BIC as a summary measure to make additional comparisons between potential models

# Wrap-up

- 4-minute exit ticket

- **Next class**
  - Assessing model fit part 2

## Class 14 Exit Ticket

https://forms.office.com/r/RwTA975tmg