# BST 210
# Applied Regression Analysis


"I checked the serial number of your laptop. It's a waffle iron."


"Tech support says the problem is located somewhere between the keyboard and my chair."

*(In light of last Thursday's technical difficulties…)*

# Lecture 15
## Plan for Today

- Logistic Regression Modeling
  - Recall: model selection process
  - Recall: calibration through goodness-of-fit tests
  - Next: discrimination through *c*-statistics and ROC curves

- Ordinal Logistic Regression
  (more than binary outcome, ordered)
- Multinomial Logistic Regression
  (more than binary outcome, not ordered)

# Review: Logistic Regression Model Selection

- We have been modeling binary outcomes as:

$$\textbf{logit}(p_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$$

always assuming

(1) we have the correctly specified model to relate $Y_i$ with $X_{ij}$

(2) $Y_i$ are independent

# Review: Logistic Regression Model Selection

- **Our modeling process for logistic regression has involved:**

  - Objective for model building; purposeful selection of covariates
  - Checking for potential confounding
  - Evaluation of possible effect modification
  - Reducing/aggregating covariates based on collinearity concerns
  - Univariable screening; scatterplot smooths
  - Assessing possible interaction effects
  - Modeling potential nonlinear terms
  - Metrics for assessing which models are 'best'
  - Subject matter knowledge and/or automated procedure for discerning 'best' model(s)
  - Validating fit/appropriateness of final model(s)
  - Keeping parsimony and hierarchy in focus

# Review: Goodness of Fit (GOF)

- At the validation/assessing-fit stage of the modeling process, the goodness of fit (GOF) of a logistic regression model is usually assessed via:

- (1) **Calibration** (covered last Thursday)

  (want model predictions to look close to the real data)

  - based on J covariate patterns
  - used less for model selection; more for GOF

- (2) **Discrimination** (will cover today)

  (want model separating cases from controls)

- Note that a *saturated model* is one that has as many parameters as there are possible covariate patterns

# Review: Calibration

- There are two common methods used to assess the **calibration** of a logistic regression model:

(a) **Pearson chi-square goodness-of-fit test**

(for smaller number of covariate patterns)

$$X^2_{Pearson} = \sum_{j=1}^{J} \frac{(O_j - E_j)^2}{V_j} \sim \chi^2_{J-(p+1)}$$

(b) **Hosmer-Lemeshow goodness-of-fit test**

(for larger number of covariate patterns)

$$X^2_{HL} = \sum_{j=1}^{G} \frac{\left(O_j - E_j\right)^2}{n_j \bar{p}_j (1 - \bar{p}_j)} \sim \chi^2_{G-2}$$

# Review: Calibration

- Let's look more closely at the hypothesis test for each -

**(a) <u>Pearson chi-square goodness-of-fit test</u>**

(for smaller number of covariate patterns)

* Hypothesis: $H_0$: the model has adequate fit

vs

$H_a$: the model does not have adequate fit

* Test Statistic: $X^2_{Pearson} = \sum_{j=1}^{J} \frac{(O_j - E_j)^2}{V_j} \sim \chi^2_{J-(p+1)}$

(df based on theory; need $V_j \neq 0$)

* Intuitively: How similar are model's predicted probabilities to truth?

(using observations from J possible covariate patterns)

# Review: Calibration

- Let's look more closely at the hypothesis test for each -

**(b) <u>Hosmer-Lemeshow goodness-of-fit test</u>**

(for larger number of covariate patterns)

* Hypothesis: $H_0$: the model has adequate fit

vs

$H_a$: the model does not have adequate fit

* Test Statistic: $$X^2_{HL} = \sum_{j=1}^{G} \frac{\left(O_j - E_j\right)^2}{n_j \bar{p}_j (1 - \bar{p}_j)} \sim \chi^2_{G-2}$$

df based on simulation

* Intuitively: How similar are model's predicted probabilities to truth?

(using G user-defined groups of observations)

# Review: Calibration

- **Final notes on calibration**:
  - Unlike in most hypothesis tests, we want p-value to be large (which means fit is good)!
  - Fit is limited to what variables were actually collected (there could be other great covariates that were not)
  - For use in model selection:
    - **Only helpful if**, while using it to assess GOF, one of the models we are comparing does not fit well (ie big Chi-square test statistic; small pvalue), giving us strong reason *not* to select it as our final model.
    - **Not helpful if** all models we are comparing have adequate fit (ie smaller Chi-square test statistics; larger pvalue), giving us no statistical reason to exclude any particular model from our selection of the final model. (Best approach is to undergo model selection based on substantive knowledge, hypothesis testing (LRT), and metrics such as AIC, *THEN* do GOF test on final model as a 'check.')

# GOF: Discrimination

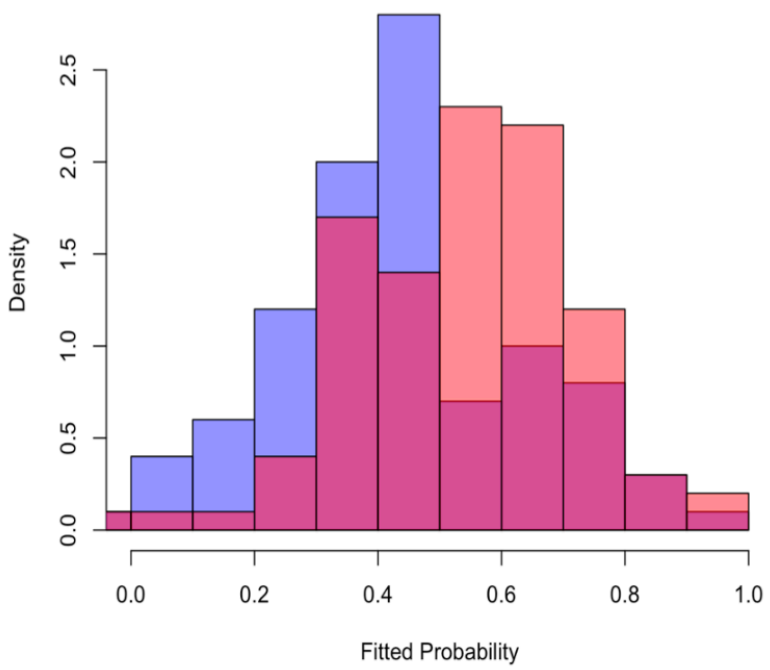
- Next we consider our 2nd approach to GOF testing in logistic regression; an alternative to calibration -

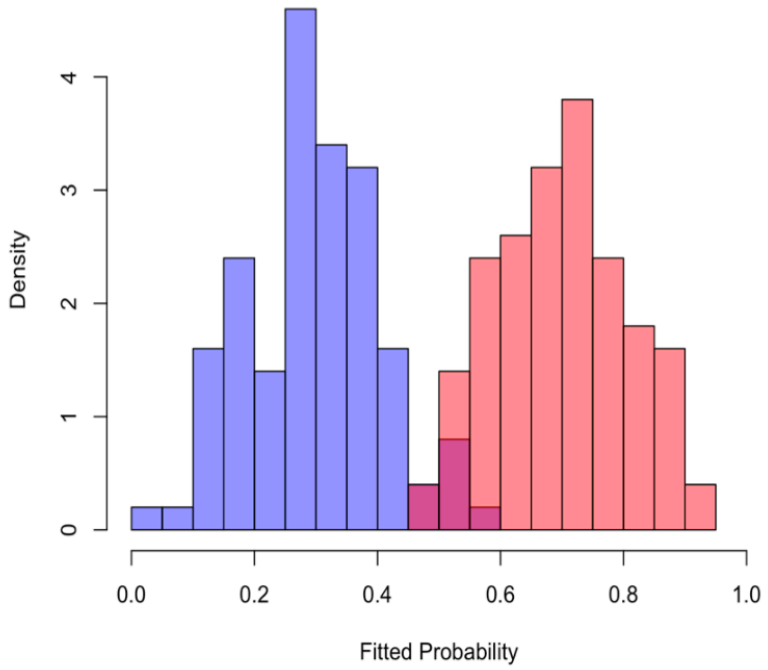
- **Discrimination**

  - A model has *good discrimination* if we can discriminate readily between the cases and controls.
  - The distribution of predicted probabilities of the outcome for cases (with the outcome) and controls (without the outcome) separate out well.
  - Cases tend to have higher probabilities of the outcome, controls tend to have lower probabilities, and there is not too much overlap between the two distributions.
  - A performance measure for a classification problem at various threshold settings

# GOF: Discrimination

(Best way is to visualize it!)



# GOF: Discrimination

- We need the concordance statistic, or ***c* statistic** to assess discrimination which can be described as follows:

- Suppose we have two subjects A and B, where A has had an outcome event (a "success" or $Y_i = 1$) and B has not had an outcome event (a "failure" or $Y_i = 0$)

# GOF: Discrimination

- Let $\hat{p}_A$ and $\hat{p}_B$ be the predicted probabilities of an event for subjects A and B, based on the logistic regression model, and define the ***c* statistic**

$$\boldsymbol{c} = \mathbf{P}(\hat{p}_A > \hat{p}_B)$$

- If the variables in the model give no information about the outcome, then $c = 0.5$ -- in that case, we might as well flip a coin to decide who gets the outcome (and we may as well not fit this model).

# GOF: Discrimination

- If the predicted probabilities of all subjects with a success outcome are higher than the predicted probabilities of all subjects with a failure outcome, then $c = 1.0$

- We say that the model predicts the outcome perfectly, though that will rarely occur

# GOF: Discrimination

- The *c* statistic is usually estimated by constructing an ***ROC (receiver-operator characteristic) curve***

- Suppose we select a cut point $p_c$ such that all subjects with predicted probability of an outcome

- $\hat{p}_i \geq p_c$ will be classified as high risk or test positive subjects.

- All subjects with $\hat{p}_i < p_c$ will be low risk or "test negative" subjects

# GOF: Discrimination

- Using the cut point $p_c$, the *sensitivity* of the test is the probability a subject has a positive test given that he or she has the outcome event (True Positive)

$$\Pr(\hat{p}_i \geq p_c \,|\, \text{case})$$

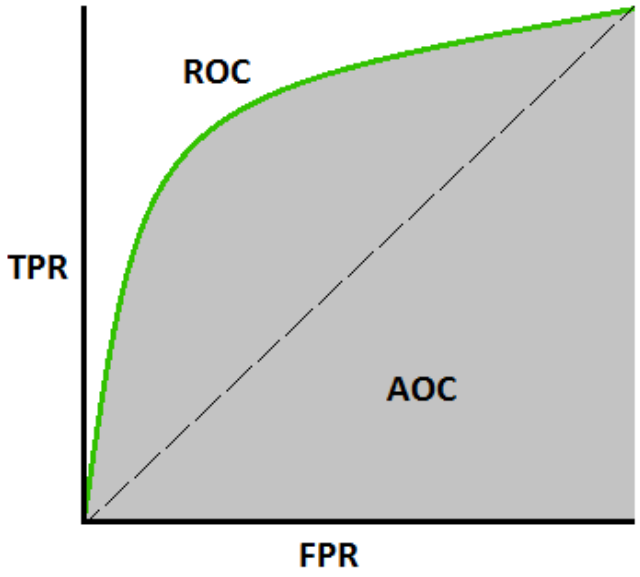
- Using the same cut point $p_c$, the *specificity* of the test is probability a subject has a negative test given that he or she does not have the outcome event (True Negative, and note 1-TN = FP)

$$\Pr(\hat{p}_i < p_c \,|\, \text{control})$$

# GOF: Discrimination - ROC Curve

- The cut point $p_c$ is arbitrary – if we increase $p_c$ the sensitivity will decrease but the specificity will increase

- An **ROC curve** is a plot of sensitivity on the *y*-axis versus [1 – specificity] on the *x*-axis, for all possible choices of the cut point $p_c$ . In other words, a plot of the True Positive rate vs the False Positive rate.

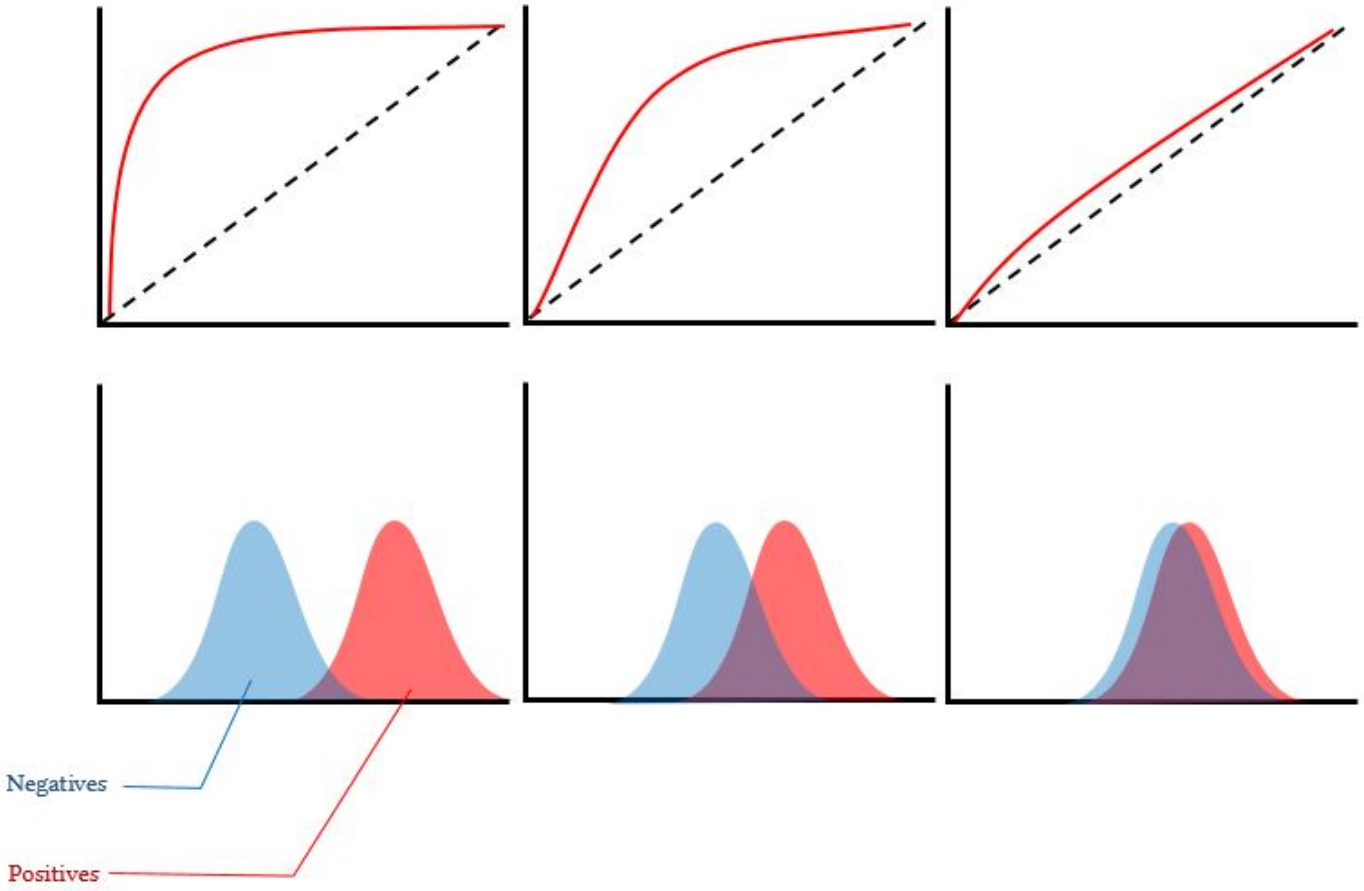


# GOF: Discrimination - ROC Curve

- The ROC is also the plot of Power P(reject $H_0 | H_1$ true) vs Type I error rate P(reject $H_0 | H_0$ true)

- If the model does not help to predict the outcome event, then the curve will be a straight (diagonal) line going through the points (0,0) and (1,1)

# GOF: Discrimination - ROC Curve



# GOF: Discrimination - ROC Curve

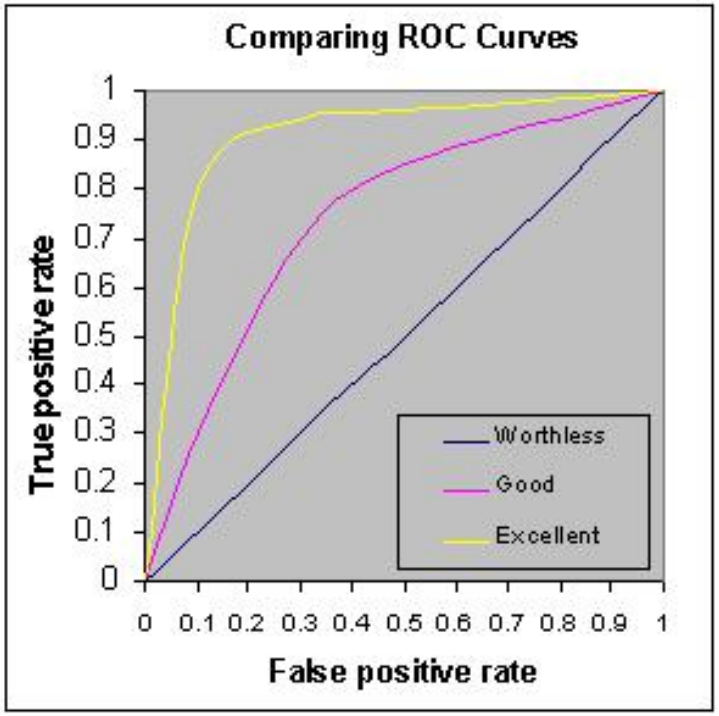
- It can be shown that the **area under the ROC curve (AUC)** is equal to the *c* statistic

- **AUC** = P(randomly selected case will have a higher predicted probability of the outcome than a randomly selected control)
  = percent chance that model will be able to classify correctly

- Higher the AUC, the better the model is at predicting 0s as 0s and 1s as 1s (or patients with disease and no disease)

# GOF: Discrimination - ROC Curve

- Higher the AUC, the better the model is at predicting 0s as 0s and 1s as 1s (or patients with disease and no disease)



# Example: Discrimination - ROC Curve

- After running a logistic regression model, we specify:

```
. lroc
```

to calculate the area under the ROC curve (AUC)

- The option `nograph` can be added

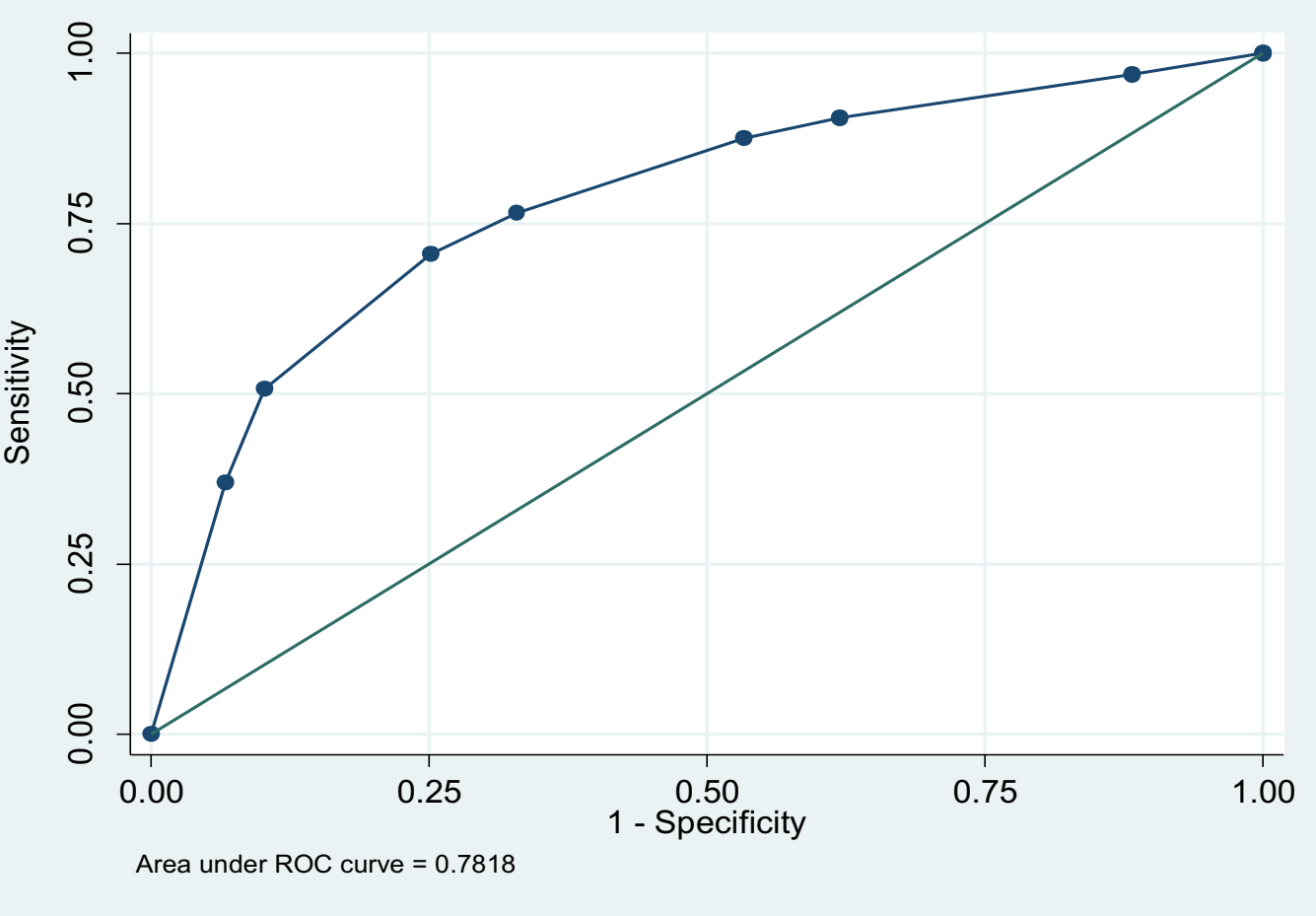
# Example: Discrimination - ROC Curve

```
. lroc

Logistic model for death

number of observations =      5629
area under ROC curve   =    0.7818
```

$c = 0.78$

In general, a *c* statistic of 0.7 is considered good while 0.8 is excellent, so *the discrimination of this model is quite good*.

# Example: Discrimination - ROC Curve



# GOF: $R^2$ in Logistic Regression?

- In linear regression, $R^2$ -- called the *coefficient of determination* -- represents the percent of the total variation in the dependent variable that is explained by the independent variables in the least squares model

- This does not make sense for logistic regression, as the method of least squares is not used to fit the model, and we don't have normally distributed outcomes, so percent of variation may not be useful

# GOF: *pseudo-R*$^2$ in Logistic Regression

- The concept of a pseudo-$R^2$ based on the log-likelihood of the fitted model is used by some computer programs

$$\text{pseudo } R^2 = \frac{(-2\log\ L)_{\text{constant only model}} - (-2\ \log L)_{\textit{fitted}\ \text{model}}}{(-2\ \log L)_{\textit{constant only model}}}$$

$$= \frac{(\log\ L)_{\textit{fitted}\ \text{model}} - (\log\ L)_{\textit{constant only model}}}{(\log\ L)_{\textit{constant only model}}}$$

# Example: *pseudo-$R^2$*

```
. logistic death surfactant bwt2 bwt3 bwt4

Logistic regression                               Number of obs   =       5629
                                                  LR chi2(4)      =    1134.94
                                                  Prob > chi2     =     0.0000
Log likelihood =  -2468.517                       Pseudo R2       =     0.1869

------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  surfactant |   .7005633    .056131    -4.44   0.000     .5987518    .8196867
        bwt2 |   .2294394   .0203125   -16.63   0.000     .1928902     .272914
        bwt3 |   .0959767   .0096688   -23.26   0.000     .0787798    .1169275
        bwt4 |   .0494499   .0055618   -26.73   0.000     .0396668    .0616457
------------------------------------------------------------------------------
```

# Example: *pseudo-$R^2$*

- Here is the constant only model

```
. logistic death

Logistic regression                               Number of obs   =       5629
                                                  LR chi2(0)      =      -0.00
                                                  Prob > chi2     =          .
Log likelihood = -3035.9852                       Pseudo R2       =    -0.0000

------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
------------------------------------------------------------------------------
```

- Here ***pseudo-$R^2$*** = (-2468.517+3035.9852)/3035.9852 = 0.1869

# Example: *pseudo-$R^2$*

- The pseudo-$R^2$ (0.1869 in this example) can only be very loosely interpreted as the percent of variation explained, but it does not have the same meaning for a binary outcome as it does for a continuous outcome, and *I do not recommend its use or presentation in a report or manuscript*

- In fact, there have been a number of *other* pseudo-$R^2$ values proposed in the literature, different packages may report different values, and none have the simple interpretation of that in the linear regression context

# Next:

- **Categorical Logistic Regression with >2 Outcomes:**

- Now we extend concepts we've learned for (binary) logistic regression to outcome variables that have more than 2 possible levels:

  - Multinomial logistic regression for unordered outcomes
  - Ordinal logistic regression for ordered outcomes

# Recall: Simple Logistic Regression Model

- For logistic regression with a single covariate $X$ to predict a binary outcome $Y$, we use the model:

$$\textbf{logit}(p) = \log[p / (1 - p)] = \beta_0 + \beta_1 x$$

- Here $p = P(Y = 1 \mid X = x)$ and we assume that the relationship between logit($p$) and $x$ is linear

- Solving for $p$, we obtain:

$$p = P(Y=1) = \frac{\exp^{(\alpha + \beta X)}}{1 + \exp^{(\alpha + \beta X)}}$$

and

$$1 - p = P(Y = 0) = \frac{1}{1 + \exp^{(\alpha + \beta X)}}$$

# New: Multinomial Logistic Regression

- Sometimes we have a categorical outcome variable that can assume ***more than two categories*** (not ordered)
- In this situation, we can use *multinomial* or *polychotomous logistic regression*
- We now need proportions $p_1$, $p_2$, $p_3$,…$p_c$ instead of simply $p_1$ and $p_2$ (all which still need to sum to 1)
- AND subjects still must be independent of each other