# Generalized Linear Modeling - Logistic Regression

- Binary outcomes
- The logit and inverse logit
- interpreting coefficients and odds ratios
- Maximum likelihood estimation
- Problem of separation
- Evaluating predictive ability
- Multiple levels of outcome - Ordered and nominal logistic regression
- The proportional odds assumption
- Multinomial regression - generalized logit model

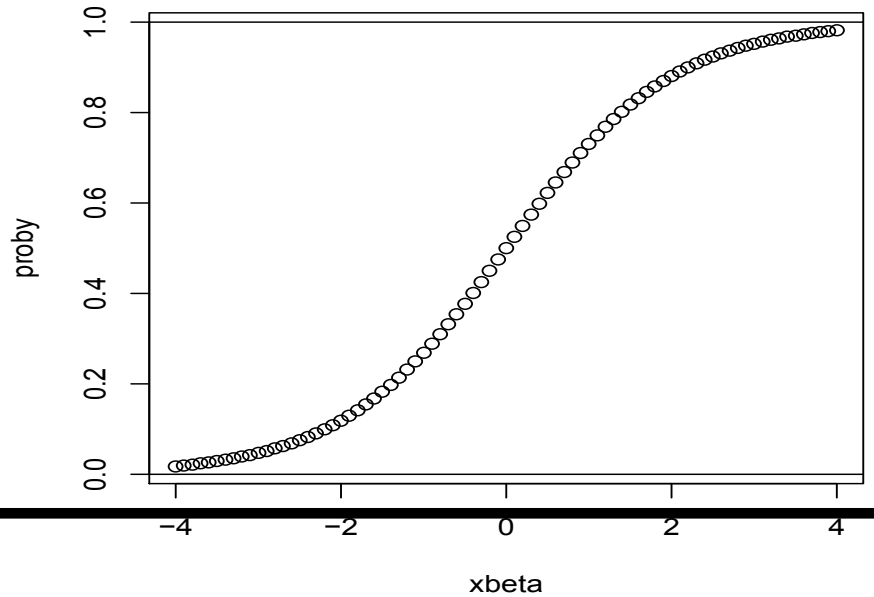
# The logistic function

How is a mother's gestational weight gain related to the the probability of the baby being born with a birthweight considered clinically in the High range (i.e. $>$ 4000 grams or $>$ 8.8 pounds).

The outcome variable takes on values of 0 or 1. Rather than fitting $Y = X\beta + \epsilon$ where we would be modeling $E(Y|X) = X\beta$, we instead model the $E(Y|X) = Pr(Y = 1|X)$ with the following **nonlinear** function called the logistic function

$$Pr(Y_i = 1|X_i) = \frac{exp(X_i\beta)}{1 + exp(X_i\beta)}.$$

```
xbeta <-seq(-4,4,by=.1)
proby <- exp(xbeta)/(1+exp(xbeta))
```



# The logistic model and deriving the logit link

Notice that since $Y_i$ is 0-1 we can model it with a Binomial distribution with parameter $\pi_i$. So we have

$$Y_i|X_i \sim Bin(1, \pi_i)$$
$$\pi_i = \frac{exp(X_i\beta)}{1 + exp(X_i\beta)}$$

For the ease of estimation we want to rewrite the relationship between $\pi$ and $X\beta$ so that $X\beta$ is on a side by itself equal to a nonlinear function of $\pi$. This is accomplished by finding the inverse function for the logistic.

Derive $X\beta = log\left(\frac{\pi}{1-\pi}\right)$...

# The logit link

Define $logit(\pi) = log\left(\frac{\pi}{1-\pi}\right)$. So rewriting the Binomial model we have

$$Y_i|X_i \sim Bin(1, \pi_i)$$
$$logit(\pi_i) = X_i\beta$$

The logit function is said to be the *canonical link* for binomial data within the generalized linear modeling framework since it is the function of the $E(Y|X)$ for which the predictors are linear.

What do you notice about the logit and its relation to the ODDS?

# Why exponentiating coefficients leads to an Odds Ratio

Consider what happens when X is increased by 1 unit...

$$log\left(\frac{P(Y=1|X)}{P(Y=0|X)}\right) = X\beta$$

$$log\left(\frac{P(Y=1|X+1)}{P(Y=0|X+1)}\right) = (X+1)\beta$$

So taking the difference we have,

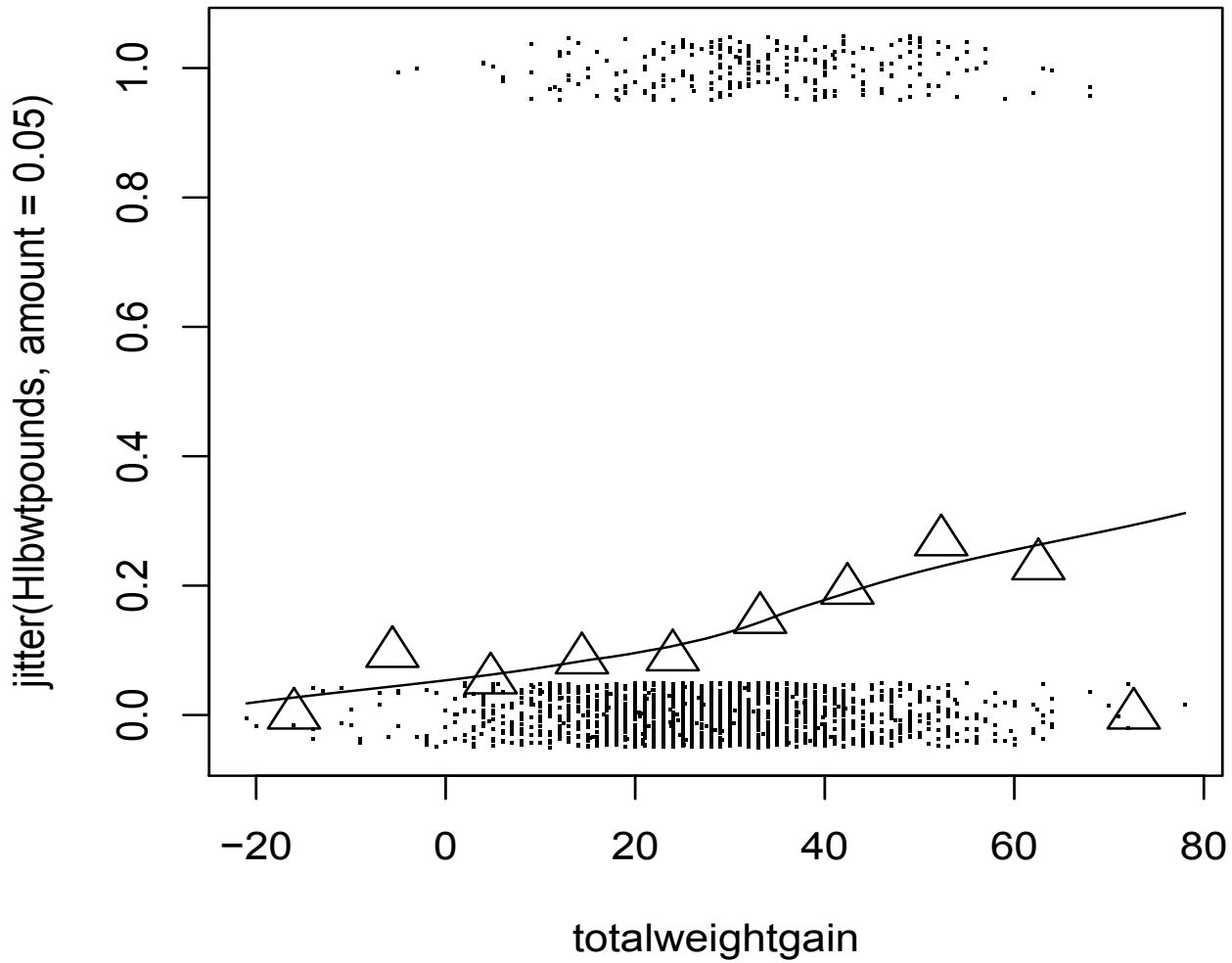
$$\beta = log\left(\frac{P(Y=1|X+1)}{P(Y=0|X+1)}\right) - log\left(\frac{P(Y=1|X)}{P(Y=0|X)}\right)$$

$$= log\left(\frac{\text{odds}(Y|X+1)}{\text{odds}(Y|X)}\right)$$

$$= log(\text{odds ratio of Y given one unit increase in X})$$

Hence, if we take $exp(\beta)$ we have odds ratio of Y given one unit increase in X.

# High Birthweight example




```
HIbwtpounds<-as.numeric(I(bwtpounds>8.8))
plot(totalweightgain,jitter(HIbwtpounds,amount=.05),pch=".")
lines(loess.smooth(totalweightgain,HIbwtpounds,span=.9))

ctotalwtgn<-factor(cut(totalweightgain,10),ordered=T)
proportionhibwt<-aggregate(HIbwtpounds,by=list(ctotalwtgn),mean)
midpttotalwtgn<-aggregate(totalweightgain,by=list(ctotalwtgn),mean)
points(midpttotalwtgn[,2],proportionhibwt[,2],pch=2,cex=2)
```

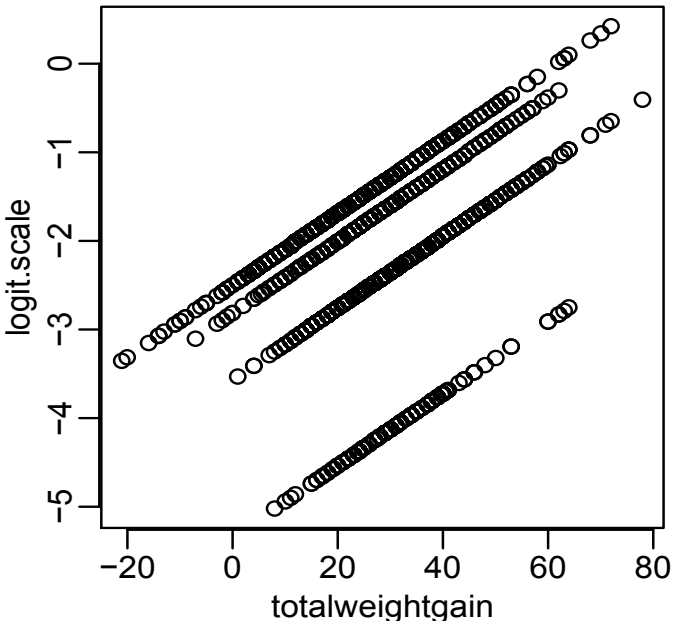
SEE HANDOUT FOR LOGISTIC REGRESSION
IN SAS and R.

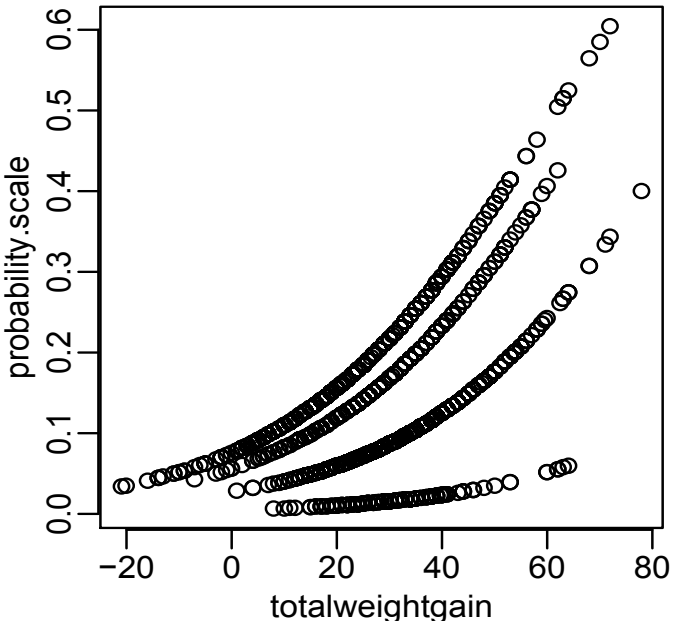
# Fitted values on the linear link and inverse link scale

- $\widehat{logit(\pi)} = \widehat{log\frac{\pi_i}{1-\pi_i}} = \mathbf{X}_i\hat{\boldsymbol{\beta}} \leftarrow$ on the logit scale
- $\hat{\pi}_i = logit^{-1}(\mathbf{X}_i\hat{\boldsymbol{\beta}}) = \frac{exp(X_i\hat{\beta})}{1+exp(X_i\hat{\beta})} \leftarrow$ probability scale

Recall the high birthweight example. We regressed high birthweight on both mother's total weight gain AND mother's baseline BMI category.

$$logit(\pi) = -3.57 + 0.0405 * totwtgain - 1.776 * underwt + 0.755 * overwt + 1.072 * obese$$




Compare the differences between what a change in the predictors means on the two different scales.

# Interpreting the intercept

$$logit(\pi) = -3.57 + 0.0405 * totwtgain - 1.776 * underwt + 0.755 * overwt + 1.072 * obese$$

What does the intercept represent? Think about back transforming it.

```
CODE for plots in R on previous page.....

\includegraphics[width=3.5in,height=1.9in,angle=0]
{Pics/bwtlogitprobscale.eps}

#### Obtains the predicted values for all observations on the logit scale
logit.scale<-fitlogistic2$linear.predictors

#### Back transforms the logit predicted values onto the original probability scale
probability.scale<-fitlogistic2$fitted.values

par(mfrow=c(2,2))
par(mar=c(2.5,2.5,1,1),mgp=c(1.6,.5,.05))
plot(totalweightgain,logit.scale)
plot(totalweightgain,probability.scale)
```

# Examining Odds Ratio, Risk Ratio and Risk Difference

| Without risk factor | | With Risk Factor | | Summary measure | | |
|---|---|---|---|---|---|---|
| Probability | Odds | Odds | Probability | OR | RR | Rdiff |
| .05 | .0526 | .1052 | 0.117 | 2 | 2.35 | .067 |
| .2 | .25 | .5 | .33 | 2 | 1.65 | .13 |
| .5 | 1 | 2 | .67 | 2 | 1.34 | .17 |
| .8 | 4 | 8 | .89 | 2 | 1.11 | .09 |
| .9 | 9 | 18 | .95 | 2 | 1.06 | .05 |
| .98 | 49 | 98 | .99 | 2 | 1.01 | .01 |

- odds = prob/(1-prob), prob = odds/(odds + 1)
- OR = odds given risk factor/ (odds given no risk factor)
- RR = prob given risk factor/ (prob given no risk factor)
- Rdiff = prob given risk factor - prob given no risk factor

Compare the summary measures across the different probabilities. How does the RR (relative risk) differ from the OR (oods ratio) across the different probabilities? How does the Rdiff (Risk difference) differ from the RR?

# Examining Odds Ratio, Risk Ratio and Risk Difference

from Chaprter 10 of Harrell F (2001) *Regression Modeling Strategies With applications to linear models, logistic regression and survival analysis.*

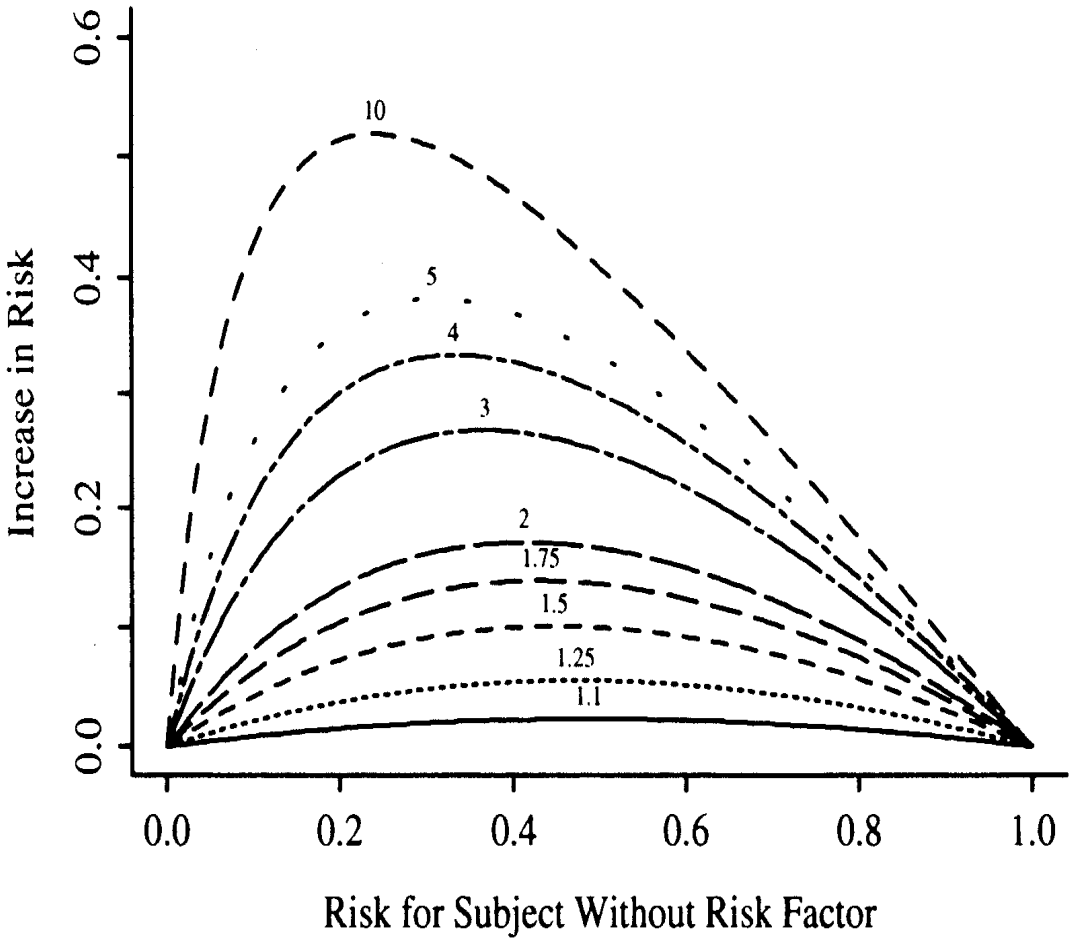



FIGURE 10.2: Absolute benefit as a function of risk of the event in a control subject and the relative effect (odds ratio) of the risk factor. The odds ratios are given for each curve.

# Estimation by Maximum Likelihood

- Given independent data $Y = Y_1, \ldots Y_n$ and $X = \mathbf{X}_1 \ldots \mathbf{X}_n$, where $Y$ is the outcome of interest and $\mathbf{X}$ are predictors, and given a parametric model for $Y_i|X_i$, we can form the likelihood function.
- Generally we can write the model for $Y_i|X_i$ as $Y_i|X_i \sim Distr(\mathbf{\Theta}, X_i)$ where $\mathbf{\Theta}$ represents a set of unknown parameters and $Dist$ represents some specific distribution family, e.g. normal, binomial, Poisson, gamma.
- The likelihood is the joint distribution of the observations viewed as a function of the parameters,

$$\text{Likelihood} L(\mathbf{\Theta}|Y; X) = \prod_{i=1}^{n} f(Y_i|X_i; \mathbf{\Theta})$$

$$\text{Log Likelihood} \ell(\mathbf{\Theta}|Y; X) = \sum_{i=1}^{n} f(Y_i|X_i; \mathbf{\Theta})$$

- The goal is to find $\mathbf{\Theta}$ which maximizes this (log)likelihood function since intuitively that value would be the value of the parametric distribution most likely to have been the one that generated the data.

# Maximizing the likelihood

- This goal of maximizing the likelihood is accomplished using calculus which provides tools for maximizing functions. The derivative of the log likelihood is taken with respect to the parameter vector $\boldsymbol{\Theta}$ and set equal to 0. The derivative of the log likelihood is called the **score function**.
- The **maximum likelihood estimates** are found by solving the score function which will yield the values that maximize the likelihood assuming the likelihood is unimodal. In general this solution must be found numerically (no closed form).
- Problems can occur when likelihood function is multimodal (only find local maximum rather than global maximum) or when the maximum is found along the boundary of the parameter space.
- We use the hat notation, $\hat{\boldsymbol{\Theta}}$, to indicate the MLEs of $\boldsymbol{\Theta}$.
- The second derivative of the log likelihood is called the **information** and is used in creating standard errors.

# The likelihood for logistic regression

Given the model

$$Y_i|X_i \sim Bin(1, \pi_i)$$
$$\pi_i = \frac{exp(X_i\boldsymbol{\beta})}{1 + exp(X_i\boldsymbol{\beta})}$$

and given n independent observations $(Y_i, \mathbf{X}_i)$

$$\begin{aligned} L(\boldsymbol{\beta}|Y, X) &= \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i} \\ &= \prod_{i=1}^{n} \frac{exp(X_i\boldsymbol{\beta})}{1 + exp(X_i\boldsymbol{\beta})}^{Y_i} \frac{1}{1 + exp(X_i\boldsymbol{\beta})}^{1-Y_i} \\ \ell(\boldsymbol{\beta}|Y, X) &= \sum_{i=1}^{n} Y_i log\left(\frac{exp(X_i\boldsymbol{\beta})}{1 + exp(X_i\boldsymbol{\beta})}\right) + (1-Y_i) log\left(\frac{1}{1 + exp(X_i\boldsymbol{\beta})}\right) \end{aligned}$$

Take derivative of this function w.r.t $\boldsymbol{\beta}$ set equal to zero and solve in order to obtain MLE's for $\boldsymbol{\beta}$, ie $hat\boldsymbol{\beta}$.

# Hypothesis testing from maximum likelihood theory

Given some hypothesis: $H_0 : \boldsymbol{\Theta} = \boldsymbol{\Theta}_0$

- **Likelihood ratio test** - ratio of the likelihood at the hypothesized parameter value (under the null) to the likelihood of the data at the MLEs. Typically the likelihood ratio is defined as -2 time log likelihood ratio, i.e.

$$\begin{aligned} LR &= -2log\frac{L_{\boldsymbol{\Theta}_0}}{L_{\hat{\boldsymbol{\Theta}}}} \\ &= -2\ell_{\boldsymbol{\Theta}_0} + 2\ell_{\hat{\boldsymbol{\Theta}}} \end{aligned}$$

- **Wald Test** - generalization of the Z or t statistics. It is a function of the difference between the MLE and the $\boldsymbol{\Theta}_0$ divided by some estimate of the standard error of the MLE.

$$W = \frac{\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0}{s.e.(\hat{\boldsymbol{\Theta}})}$$

- **Score Test** - measures how far away from zero the score function is when evaluated at the $H_0$. Typically it is standardized by the information.

See handout from Harrel (2002) Chapter 9 Which test statistic to use when

# Look back at confidence intervals from High Birthweight Example

- Notice difference in CI's from SAS and R
- SAS creates Wald confidence intervals by default. Estimate +- 1.96 * S.E.
- The confint() function in R creates Likelihood ratio based confidence intervals (done computationally no closed form)
- Adding the option CLodds = PL to the model statement in SAS will provide the "profile likelihood confidence intervals". These confidence intervals based on the likelihood ratio test
- Hauck and Donner (1977) Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association*, 72:851-863 notice that the Wald CI can be too large especially when there are strong effects.
- LR confidence intervals considered better. With larger samples they will be very similarly (asymptotically the same).

# Problem of Separation in Logistic Regression

- An identifiability problem that can arise in logistic regression, called separation, occurs when a predictor or a combination of predictors are perfectly aligned with the outcome such that y = 0 for ALL values of that predictor beyond some point and y = 1 for ALL values of that predictor less than some point.
- Often occurs in small or sparse samples with highly predictive covariates.
- Simples case is in the analysis of a $2 \times 2$ table with one zero cell count.
- For a continuous predictor, separation can be demonstrated by (Draw plot).
- For a categorical predictor separation means that in some category (or with multiple predictors, in some combination of categories) all individuals in that category either have a 1 or 0.

Leads to non-convergence of the likelihood and/or infinite parameter estimates.

# Solutions to the problem of Separation

**Classical solution** - Drop the predictor or somehow aggregate levels. Leave problematic predictors in but only report results for predictors without separation problem.

**Modern solution** -
See the website http://www.meduniwien.ac.at/msi/biometrie/programme/fl/
"Logistic regression using Firth's bias reduction: a solution to the problem of separation in logistic regression". Heinze and Ploner, 2004 put together a SAS MACRO (%fl) and also an R package (logistf()) that uses a **penalized maximum likelihood** method to obtain estimates.

We will try it out in the lab.

# Summarizing predictive ability in logistic regression

- An intuitive measure is the error rate - the proportion of cases for which the prediction of $\hat{y}_i$ is the same as $y_i$. Depends on the cutoff value chosen to define "positive" prediction.
- A natural choice is to take

$$\begin{cases} \hat{y}_i = 1 \text{ if } \hat{\pi}_i \geq \hat{p} \\ \hat{y}_i = 0 \text{ if } \hat{\pi}_i < \hat{p} \end{cases} \tag{1}$$

  where $\hat{p}$ is the overall proportion of 1s in the sample. That is, $\hat{p} = \bar{Y}$.
- Comparing $\hat{y}_i$ to $y_i$ yields a $2 \times 2$ table. The error rate is the proportion of observations on the off-diagonal
- To get this in SAS, use the ctable option after the model statement, can get error rate for any cutoff value
- To get this in R, calculate directly using predicted probabilities.

**See page from Harrell about why this simple measure should generally be avoided**

# Summarizing predictive ability in logistic regression

Better measures:

- $R^2$ or max-rescaled $R^2$ - function of the likelihood ratio test. Unlike linear regression it is not necessarily the case that more predictors lead to higher $R^2$ values. The maximum possible value of generalized R2 is not 1.0 as it is for linear regression. Max-rescaled R-Square divides by this maximum value to fix this so its maximum is 1.
- c index - rank correlation between the predicted probability of response under the fitted model and the actual response. It is equivalent to the area under a receiver operating characteristic (ROC) curve. The larger the area under this curve, the better the predictions. The maximum area is 1.0, and an area of 0.5 implies random predictions (i.e., a prediction of success is as likely whether success or failure is the truth). Harrell (1998) gives a guideline of C exceeding 0.80 as implying useful predictability of the model.
- AIC is **only useful as a comparative fit index** and is a penalized function of the log-likelihood, penalized by the the number of parameters in the model - when comparing two models, smaller values are better.

See handout for obtaining these in SAS and R

# Dealing with more than two outcome levels - Ordered categories

Examples: tumor stage (local, regional, distant), disability severity (none, mild, moderate severe), Likert items (strong disagree, disagree, agree, strongly agree)

- Dichotomize at some fixed level corresponding to a logical outcome of interest, e.g. maybe it is particularly of interest to distinguish between tumors detected at the regional stage and those at the distant stage, hence we could dichotomize the stages at that point.
- Could treat the ordered categories as a continuous variable. If it is reasonable to assume that a unit difference between one level and the next is constant, then this can be a reasonable approach. Often Likert items are simply treated as if they are continuous scores with unit increments 1,2,3,4.
- **Both above methods are suboptimal** since they either throw out information (dichotomizing) or make uncheckable assumptions (treating as continuous)
- A popular way to model the ordered categories directly is using an **ordered logistic regression**, also called ordinal or cumulative logistic regression and also called a "proportional odds model" which aptly states the model's main assumption

# Ordered logistic regression

Let $Y_i$ take on categories $1, 2, \ldots K$, the ordered logistic regression model is

$$Y_i \sim \text{Multinomial}(\pi_1, \pi_2, \ldots \pi_K)$$

$$log\left(\frac{\pi_{j+1} + \ldots + \pi_K}{\pi_1 + \pi_2 + \ldots \pi_j}\right) = log\left(\frac{Pr(Y_i > j)}{Pr(Y_i \leq j)}\right) = \beta_{0j} + \boldsymbol{\beta}\mathbf{X}$$

and $\beta_{01} \geq \beta_{02} \ldots \geq \beta_{0(K-1)}$

where $j = 1 \ldots K - 1$. Hence we are modeling the log odds of being greater than the cutoff value $j$ as compared to being less than it and a similar expression applies for $j$ at all $K - 1$ levels. For example, if K = 4 then we are modeling the odds of: 2,3,4 vs. 1; and 3,4 vs. 1,2; and 4 vs. 1,2,3

Note that the intercept parameter $\beta_{0j}$ is different for each $j$ allowing the jump in probability from one level to the next to differ, but that the $\boldsymbol{\beta}$ relating the predictor $\mathbf{X}$ to the logit of the outcome is constant across all $j$.

This **constant** $\boldsymbol{\beta}$ - interpreted as the "log odds ratio of being at a higher level compared to a lower level associated with a unit increase in $\mathbf{X}$" - is a strong assumption and is referred to as the "proportional odds" assumption and should be tested against the data.

# Proportional odds model in SAS and R

- In SAS: See "Using the proportional odds model for health-related outcomes: Why, When and How with Various SAS procedures by Marc Gameroff. **We will go through the example in that handout.** PROC LOGISTIC works.
- In R: Can use the **lrm()** function in the Design Package (see https://www.ats.ucl for an example). This is the same function that can be used to get the c-index and R-square for logistic regression. The proportional odds model can also be fit using **polr()** in the MASS Package, and the vglm() function in the VGAM Package.

See handout for "Fitting ordered logistic regression in SAS and R" where mother's baseline bmi category is regressed on age and parity.

## Assessing the proportional odds assumption

The ordered logistic regression model basically assumes that the way $\mathbf{X}$ is related to being at a higher compared to lower level of the outcome is the same across all levels of the outcome.

The **global test** for proportional odds considers a model

$$log\left(\frac{Pr(Y_i > j)}{Pr(Y_i \leq j)}\right) = \beta_{0j} + \boldsymbol{\beta}_j \mathbf{X}$$

and tests whether $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \ldots \boldsymbol{\beta}_{K-1}$ for all $p$ elements of $\boldsymbol{\beta}$ hence it is a test with $p * (K-2)$ degrees of freedom. This test is **known to be problematic** since it is "anti-conservative" (rejects more than it should) plus as a global test it does not tell us where the problem of non-proportionality is or how practically important it is.

Discuss in DETAIL the following paper in class: Bender R and Grouven U (1998) Using Binary Logistic Regression Models for Ordinal Data with Non-proportional Odds, *J Clin Epidemiology*, 51(10) 809-816.

- recommends considering separate tests for each covariate (from unadjusted models)
- recommends comparing slopes from separately fit logistic regression models
- discusses PPOM - partially proportional odds model and generalized logit models

# Dealing with more than two outcome levels - Nominal categories

Examples: consumer brand choice (Geico, State Farm, Acuity, Progressive), homeless sleeping situation (on street, with friend/family, hotel, shelter), parenting style (authorative, authoritarian, permissive, neglectful)

- Could run separate logistic regression models, one comparing each pair of outcomes. In fact this is quite similar to what the multinomial logistic regression model does.
- Could collapse categories so there were only two and then do a logistic regression, but this would lose information that may be of interest across categories
- Multinomial logistic or "generalized logit" models are a way to fit a nominal category outcome in a regression framework.

# Multinomial logistic model - Nominal categories

Let $Y_i$ take on categories $1, 2, \ldots K$, the general multinomial model is

$$Y_i \sim \text{Multinomial}(\pi_1, \pi_2, \ldots \pi_K)$$

$$log\left(\frac{\pi_j}{\pi_K}\right) = log\left(\frac{Pr(Y_i = j)}{Pr(Y_i = K)}\right) = \beta_{0jK} + \boldsymbol{\beta}_{jK}\mathbf{X}$$

where $j = 1 \ldots K - 1$ and $K$ is fixed as the reference group. Hence we are modeling the log odds of being at any particular level $j$ as compared to being in the reference class $K$ and this relationship is allowed to be different across the covariates. For example, if K = 4 then we are modeling the odds of: 1 vs. 4; and 2 vs. 4; and 3 vs. 4

- In SAS: use PROC LOGISTIC and add the /link=glogit option on the model statement
- In R: use multinom() in the nnet library of the MASS package, or vglm() in the VGAM package.