STA 303/1002: Class 9- Case Study III Inference

Binary Logistic Regression Example

- Case Study III: The Donner Party Example
  - Confidence interval for Odds Ratio
  - Testing β's $\rightarrow$ Higher-order Models
- Joke: *"I asked a statistician for her phone number... and she gave me an estimate."*(www.workjoke.com)

Logistic Regression: Testing whether single β's are zero

WALD CHI-SQUARE PROCEDURES

- Hypotheses: $H_0 : \beta_j = 0$ (X_j has no effect on log-odds)
  $H_a : \beta_j \neq 0$
- Test Statistic: $\boxed{z = \dfrac{\widehat{\beta}_j}{SE(\widehat{\beta}_j)}}$

  where
  - $\widehat{\beta}_j$- maximum likelihood (ML) estimate and
  - $SE(\widehat{\beta}_j)$- estimated standard error from the numerical procedure that generated the MLE.
  - By standard large-sample results, MLE's are normally distributed. Thus, for large n, under H_0, z is an observation from an approx. $\mathcal{N}(0, 1)$ distribution.
- 95% Confidence interval: $\boxed{\widehat{\beta}_j \pm 1.96 SE(\widehat{\beta}_j)}$

$\text{logit}(\pi) = X\beta$
$= \beta_0 + \beta_1 X_1 + \ldots$

$Est \pm Z_{\alpha/2}\, SE(Est)$.

Examples: Testing whether single β's are zero

Using R output ('Coefficients'):

	Age	*Sex*
Test statistic	$(-0.078/0.0373)^2$	
P-value	0.036	
95% CI for β	$-0.078 \pm 1.96(0.0373)$ $=$(-0.15, -0.0055)	
CI for Odds ratio	$(e^{-0.15}, e^{-0.0055})$=(0.86,0.995)	
Conclusion	For the same sex, the odds ratio for a 1-year increase in age is between .86 and 0.995.	

In R Output

$\hat{\beta}_j \pm z_{\alpha/2}\, SE(\hat{\beta}_j)$

Recall relationship between $\mathcal{N}(0,1)$ and Chi-square distribution:

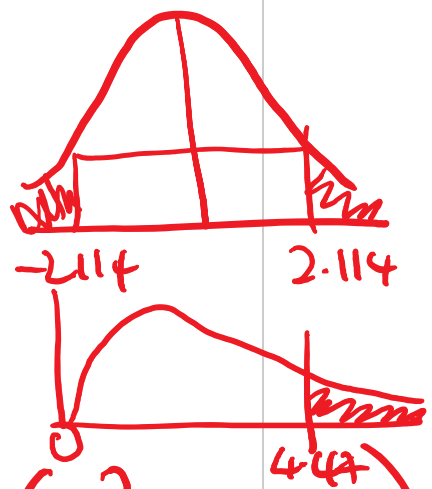
Z $\qquad$ Z^2

Examples: Testing whether β's are zero

Using R output:

	Age	*Sex*
Test statistic	$(-0.078/0.0373)^2$	4.47 $= 2.114$ (Refer to Additive model)
P-value	0.036	0.0345
95% CI for β	$-0.078 \pm 1.96(0.0373)$ =(-0.15, -0.0055)	(0.117, 3.078)
CI for Odds ratio	$(e^{-0.15}, e^{-0.0055})$=(0.86,0.995)	(1.124, 21.72)
Conclusion	For the same sex, the odds ratio for a 1-year increase in age is between .86 and 0.995.	

- Note: Both marginal p-values are less that 0.05 and the confidence intervals for the odds ratios do not include 1.
- Hence, we have moderate evidence that both *Age* and *Sex* have an effect on survival over and above each other.
- Recall: If $Z \sim \mathcal{N}(0,1)$, then $Z^2 \sim \chi_1$.



$$\text{P-value} = 2P(Z \geq |2.114|) = P(\chi^2_1 > 4.47).$$

In R: 2(1-pnorm(2.114)) , 1-pchisq(4.47,df=1)

Additional CI Examples

Using R output:

- Q: Find a 95% CI for the change in odds of survival for a 40-yr old to 20-yr old of the same sex.
- A:
  - The log odds change by -0.078*(40-20)=-1.56.
  - 95% CI for the change in log odds is 20*(-0.15, -0.0055) =(-3.0, -0.11).
  - 95% CI for the odds ratio is (0.05, 0.896). $=(e^{-3.0}, e^{-0.11})$
  - The odds of survival of a 40-yr old woman were $e^{-1.56} = 0.21$ times the odds of survival for a 20-yr old.

- Note that it is not appropriate to compute CI for π since $0 \leq \pi \leq 1$ and it is not normally distributed.

Model Assumptions for Binary Logistic Regression

1. Underlying probability model for response is Bernoulli. → Var is not constant
2. Observations are independent.
3. The form of the model is correct.
   - Linear relationship between logits and explanatory variables
   - All relevant variables are included; irrelevant ones excluded
4. Sample size is large enough for valid inference-tests and CIs. (Recall large-sample properties of MLEs.)

Binary Logistic Regression vs Linear Regression

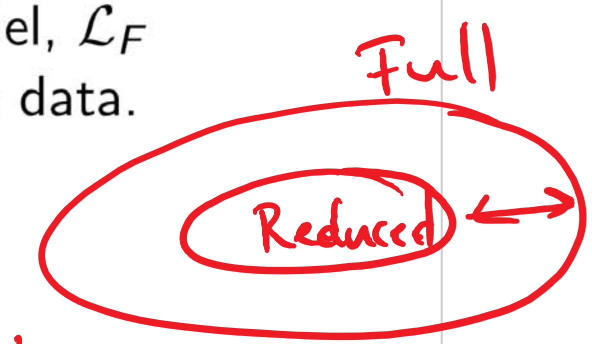
- Both utilize MLE's for the β's
- *Less assumptions to check for than in linear (least squares) regression*
  - No need to check for outliers since Y is either 0 or 1.
  - No residual plots; No meaning can be inferred from residuals
  - Variance is not constant

Case Study III: Testing model assumptions

- Independence: We know that there were families within Donner's party, so we have concerns that the observations were not independent!
- Form of the model: Test higher-order terms such as
  - Age^2- non-linear (quadratic) in X
  - $Sex * Age$ interaction, and
  - $Age^2 * Sex$ interaction.

Comparing models: Likelihood Ratio Test

- **Idea**: Compare likelihood of data under FULL (F) model, $\mathcal{L}_F$ to likelihood under REDUCED (R) model, $\mathcal{L}_R$ of same data.



$$\text{Likelihood ratio} : \frac{\mathcal{L}_R}{\mathcal{L}_F}, \quad \text{where } \mathcal{L}_R \leq \mathcal{L}_F$$

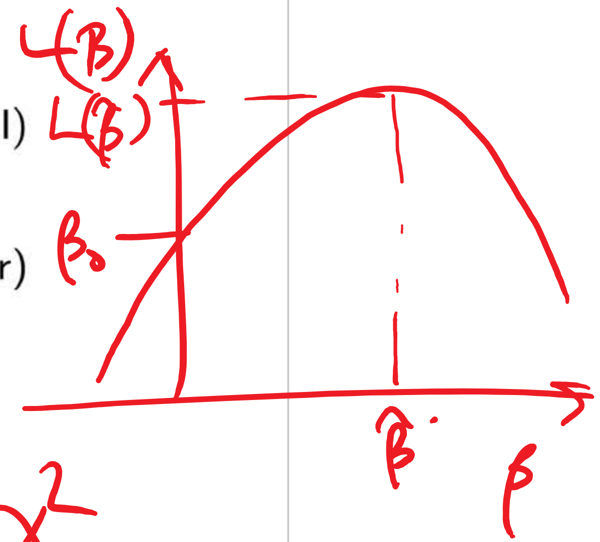
$-2L_R \geq -2L_F$

- **Hypotheses**: $H_0 : \beta_1 = \cdots = \beta_k = 0$

  (Reduced model is appropriate; fits data as well as Full model)

  H_a: at least one $\beta_1, \cdots, \beta_k \neq 0$

  (Full model is better)



- **Test Statistic**: Deviance (residual),

$$G^2 = \boxed{-2\log\mathcal{L}_R} - (\boxed{-2\log\mathcal{L}_F}) = -2\log\left(\frac{\mathcal{L}_R}{\mathcal{L}_F}\right) \sim \chi^2_k$$

- For large n, under H_0, G^2 is an observation from a chi-square distribution with k df.

Case Study III Exercise: Comparing models

Using R output,
Q: Determine whether a model with the 3 higher-order polynomial terms and/or interaction terms is an improvement over the additive model.

- Hypotheses:
- Test Statistic:
- Distribution of TS:
- P-value:
- Conclusion:

Case Study 3: Higher Order Model with 3 higher order/interaction terms

```
fitfull<-glm(Status~Age+sex+Age:sex+I(Age^2)+I(Age^2):sex, family=binomial, dat
summary(fitfull)
```

```
##
## Call:
## glm(formula = Status ~ Age + sex + Age:sex + I(Age^2) + I(Age^2):sex,
##     family = binomial, data = donner)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3396  -0.9757  -0.3438   0.5269   1.5901
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -3.318484   3.940184  -0.842    0.400
## Age                  0.183031   0.226632   0.808    0.419
## sexFemale            0.265286  10.455222   0.025    0.980
## I(Age^2)            -0.002803   0.002985  -0.939    0.348
## Age:sexFemale        0.299877   0.696050   0.431    0.667
## sexFemale:I(Age^2) -0.007356   0.010689  -0.688    0.491
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 45.361  on 39  degrees of freedom
## AIC: 57.361
```

Testing β's: Wald versus LRT test

	Wald	LRT
Testing whether a single β=0		
Comparing nested models		
Small to moderate sample sizes		
β near boundary of parameter space		

Case Study III Exercise: Comparing models

Using R output,
Q: Determine whether the effect of *Age* on the odds of survival differ with *Sex*.

$\Rightarrow \beta_3 = 0$

- Hypotheses: H_0: $\text{logit}(\pi) = \beta_0 + \beta_1 Age + \beta_2 1_F$
  H_a: $\text{logit}(\pi) = \beta_0 + \beta_1 Age + \beta_2 1_F + \beta_3 (Age \times 1_F)$
- Test Statistic: $G^2 = D_0 - D_A = 51.256 - 47.346 = 3.91$
- Distribution of TS: $G^2 = 3.91 \sim \chi^2_1$
- P-value: $P(\chi^2_1 > 3.91) = 0.048$
- Conclusion:
  Suggestive but not conclusive evidence of interaction

Case Study 3: Interaction Model, Age*Sex

```
fitas<-glm(Status~Age*sex, family=binomial, data=donner)
summary(fitas)
```

```
##
## Call:
## glm(formula = Status ~ Age * sex, family = binomial, data = donner)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2279  -0.9388  -0.5550   0.7794   1.6998
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.31834    1.13103   0.281   0.7784
## Age             -0.03248    0.03527  -0.921   0.3571
## sexFemale        6.92805    3.39887   2.038   0.0415 *
## Age:sexFemale   -0.16160    0.09426  -1.714   0.0865 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 47.346  on 41  degrees of freedom
## AIC: 55.346
##
```

Comparing models: 'Global' LRT

- Idea: Compares Fitted model to NULL [$\text{logit}(\pi) = \beta_0$] model
- Hypotheses: $H_0 : \beta_1 = \cdots = \beta_p = 0$

  (NULL model is appropriate)

  H_a: at least one $\beta_1, \cdots, \beta_p \neq 0$

  (Fitted model is better)

Case Study III Exercise: ‘Global’ LRT

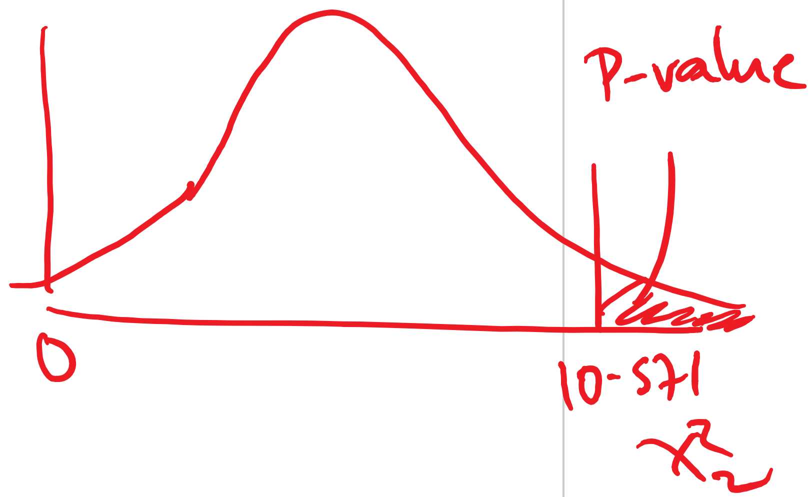
$\text{logit}(\pi) = \beta_0 + \beta_1 \text{Age} + \beta_2 1_F$

Using R output,
Q: Determine whether or not the additive model fits better than the Null model.

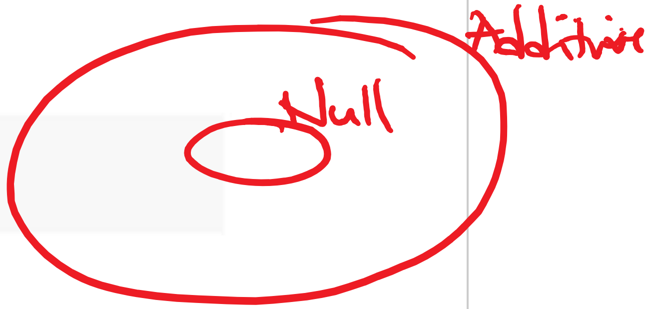
$\text{logit}(\pi) = \beta_0$

- Hypotheses: H_0: (Null) $\beta_1 = \beta_2 = 0$ vs H_a: (Additive model is better) at least 1 of the β's not zero
- Test Statistic: $G^2 = D_N - D_A = 61.827 - 51.256$
- Distribution of TS: $= 10.571 \sim \chi^2_2$
- P-value: $P(\chi^2_2 > 10.571) = 0.005$
- Conclusion:

The additive model is better. IOW, Age and Sex are relevant for the log odds of survival.



Case Study 3: Additive model for Survived




```r
fitasf<-glm(Status~Age+sex, family=binomial, data=donner)
summary(fitasf)
```

```
##
## Call:
## glm(formula = Status ~ Age + sex, family = binomial, data = donner)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.7445  -1.0441  -0.3029   0.8877   2.0472
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.63312    1.11018   1.471   0.1413
## Age         -0.07820    0.03728  -2.097   0.0359 *
## sexFemale    1.59729    0.75547   2.114   0.0345 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 51.256  on 42  degrees of freedom
## AIC: 57.256
##
## Number of Fisher Scoring iterations: 4
```

$\hat{\beta}_j$ $SE(\hat{\beta}_j)$ $\hat{\beta}_j / SE(\hat{\beta}_j)$ $2P(Z \geq |z|)$

$\hat{\beta}_1$ — Age, $\hat{\beta}_2$ — sexFemale

$H_0: \beta_j = 0$

$H_a: \beta_j \neq 0$

R : $\text{Deviance}_R = -2 \log L_R$

F : $\text{Deviance}_F = -2 \log L_F$

$G^2 = \text{Deviance}_R - \text{Deviance}_F$
$= 61.827 - 51.256$

Week 4 R functions

- Create factor: `as.factor()`
- Cross Tabulations: `xtabs()`
- Specifying the reference level: `relevel()`
- Generalized Linear Models: `glm()`
- Find deviance: `deviance()`