STA303/1002 - Week 6 R Markdown

February 14, 2018

Case Study 3: The Data

Get the data (from R library):

```r
#load Sleuth3 R data library; see case2001
library(Sleuth3)
#Donner party survival data
donner = case2001
str(donner)
```

```
## 'data.frame':    45 obs. of  3 variables:
##  $ Age   : int  23 40 40 30 28 40 45 62 65 45 ...
##  $ Sex   : Factor w/ 2 levels "Female","Male": 2 1 2 2 2 2 1 2 2 1 ...
##  $ Status: Factor w/ 2 levels "Died","Survived": 1 2 2 1 1 1 1 1 1 1 ...
```

```r
attach(donner)
head(donner)
```

```
##   Age    Sex   Status
## 1  23   Male     Died     0
## 2  40 Female Survived     1
## 3  40   Male Survived     1
## 4  30   Male     Died     0
## 5  28   Male     Died     0
## 6  40   Male     Died     0
```

Case Study 3: Summarizing the data

```r
#two-way contingency table for status by sex
#check that cell counts>0
xtabs(~Status+Sex, data=donner)
```

```
##           Sex
## Status     Female Male
##   Died          5   20
##   Survived     10   10
```

```r
summary(Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    15.0    24.0    28.0    31.8    40.0    65.0
```

Case Study 3: Marginal Mean Ages

```r
tapply(Age, Status, mean)
```

```
##     Died Survived
##    35.48    27.20
```

```r
tapply(Age, Sex, mean)
```

```
##   Female     Male
## 31.06667 32.16667
```

```r
fita<-glm(Status~Age+Sex, family=binomial, data=donner)
```

Case Study 2: Additve model summary

```
##
## 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)  3.23041    1.38686   2.329   0.0198 *
## Age         -0.07820    0.03728  -2.097   0.0359 *
## SexMale     -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
```

Case Study 3: ANOVA table

```r
anova(fita)
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Status
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev
## NULL                    44     61.827
## Age   1   5.5358        43     56.291
## Sex   1   5.0344        42     51.256
```

Case Study 3: Modelling "Died"

```
status=relevel(Status, ref="Survived")
fitad<-glm(status~Age+Sex, family=binomial, data=donner)
summary(fitad)
```

```
##
## Call:
## glm(formula = status ~ Age + Sex, family = binomial, data = donner)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0472  -0.8877   0.3029   1.0441   1.7445
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.23041    1.38686  -2.329   0.0198 *
## Age          0.07820    0.03728   2.097   0.0359 *
## SexMale      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
##
```

Case Study 3: Sex Reference group as "Male"

```
sex=relevel(Sex, ref="Male")
fitadf<-glm(status~Age+sex, family=binomial, data=donner)
summary(fitadf)
```

```
##
## Call:
## glm(formula = status ~ Age + sex, family = binomial, data = donner)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0472  -0.8877   0.3029   1.0441   1.7445
##
## 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
##
```

Model 1

Case Study 3: Additive model for Survived

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

$p=2$

$\text{logit}(\pi_S) = 1.63 + (-0.078)\,\text{Age} + 1.59\, 1_F$

```
##
## 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
```

$1_F = \begin{cases} 1 & F \\ 0 & M \end{cases}$

$-2 \log L_{\text{Model 1}}$

$p=2$

$51.256 + 2(p+1)$

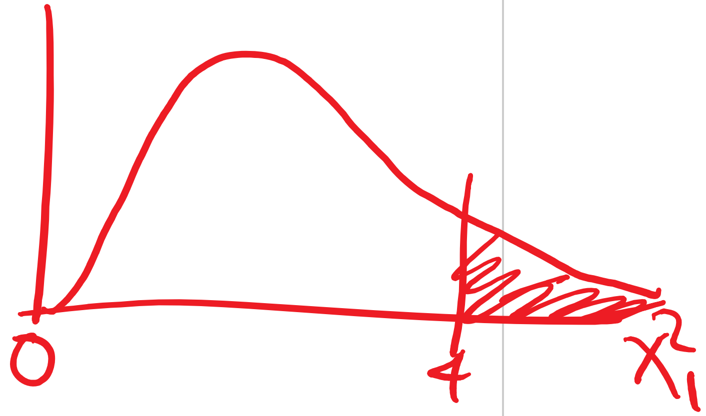
Case Study 3: Additive model for Survived

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

```
##
## Call:
## glm(formula = Status ~ sex + Age, 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
## sexFemale    1.59729    0.75547   2.114   0.0345 *
## Age         -0.07820    0.03728  -2.097   0.0359 *
## ---
## 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
```

$$z = \frac{-0.078}{0.0373}$$

$$z^2 = (-2.097)^2$$



Case Study 3: More ANOVA tables

```r
anova(fitasf)
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Status
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev
## NULL                    44     61.827
## Age   1   5.5358        43     56.291
## sex   1   5.0344        42     51.256
```

Case Study 3: More ANOVA tables

```
anova(fitsaf)
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Status
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev
## NULL                   44     61.827
## sex   1   4.5403       43     57.286
## Age   1   6.0300       42     51.256
```

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

Model 2

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

Age^2

\# of predictors $= p = 5$

```
##
## 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
```

$\hat{\beta}_j$ $SE(\hat{\beta}_j)$

$45.361 + 2(5+1)$

$-2 \log L_M$

Case Study 3: Interaction Model, Age*Sex

Model 3

$p=3$

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

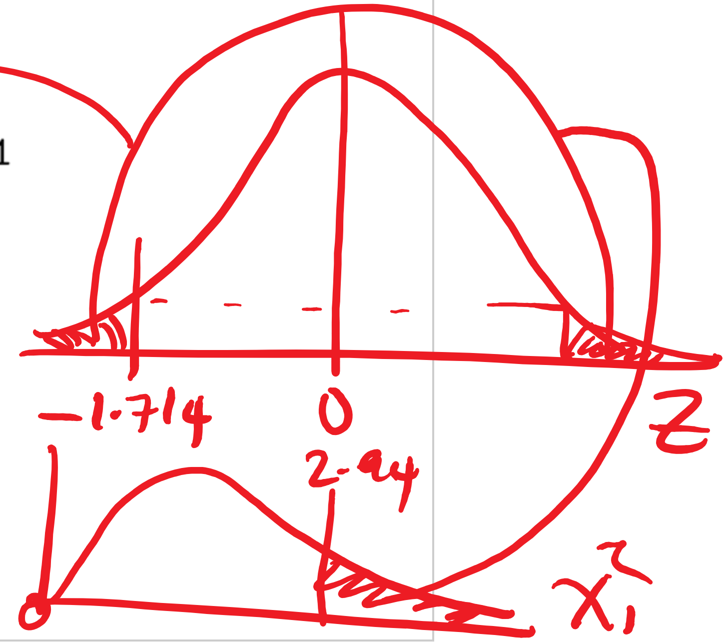
$\text{logit}(\hat{\pi}) = \hat{\beta}_0 + \hat{\beta}_1 Age + \hat{\beta}_2 1_F + \hat{\beta}_3 (Age)(1_F)$

```
##
## 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
##
```

$z^2 = 2.94$

$z = -1.714 = \dfrac{-0.1616}{0.09426}$

$\sim N(0,1)$



$-2\log L_{M_2}$

$47.346 + 2(4)$

Case 3: Deviance test and Estimated Var-Cov of β

```
anova(fitasf, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Status
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                   44     61.827
## Age   1   5.5358       43     56.291  0.01863 *
## sex   1   5.0344       42     51.256  0.02485 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
print(vcov(fitasf,digits=3))
```

Additive

```
##             (Intercept)          Age   sexFemale
## (Intercept)  1.23250837 -0.038472741  0.06007099
## Age         -0.03847274  0.001390134 -0.00823197
## sexFemale    0.06007099 -0.008231970  0.57073339
```

$\rightarrow se(\hat{\beta})$.

Eg, $se(\hat{\beta}_2) = 0.7557$

$Var(\hat{\beta}_2) = 0.7557^2 = 0.57$

Case 3: Confidence Intervals for β's

coef

```
cbind(bhat=coef(fitasf), confint.default(fitasf))  # 95% CI for betas
```

```
##                    bhat      2.5 %       97.5 %
## (Intercept)  1.63312031 -0.5428002  3.809040837
## Age         -0.07820407 -0.1512803 -0.005127799
## sexFemale    1.59729350  0.1166015  3.077985503
```

```
exp(coef(fitasf)) # exponentiate estimated betas, get odds ratios
```

```
## (Intercept)         Age   sexFemale
##   5.1198252   0.9247757   4.9396452
```

```
exp(cbind(OR=coef(fitasf), confint.default(fitasf))) #CI for odds ratio
```

```
##                    OR     2.5 %     97.5 %
## (Intercept) 5.1198252 0.5811187 45.1071530
## Age         0.9247757 0.8596067  0.9948853
## sexFemale   4.9396452 1.1236716 21.7146143
```

Case 3: Wald tests in R

Computes Wald chi-squared test for 1 or more β coefficients

- R package: aod (Analysis of Overdispersed Data)
- Syntax wald.test(Sigma, b, Terms)
- Sigma: var-cov matrix, extracted from the glm function
- b: coefficients (coef(glm()))
- Terms: specifies which terms in the models are to be tested

```
library(aod) # Analysis of Overdispersed Data
wald.test(Sigma=vcov(fitasf), b=coef(fitasf), Terms=2:3)
```

Do not utilize Wald for more than 1 β.

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 6.9, df = 2, P(> X2) = 0.032
```

Case 3: Wald tests in R

```r
# Testing interaction, Refer to interaction model
# summary(fitas)
# Testing a single beta
wald.test(Sigma=vcov(fitas), b=coef(fitas), Terms=4)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 2.9, df = 1, P(> X2) = 0.086
```

Case 3: Estimated probability of survival

π_i – observed

$\hat{\pi}_i$ – estimate

$n=45$

```
phats<-predict.glm(fitasf, type="response") # predicted probability of survival
phats[1:5]
```

```
##         1         2         3         4         5
## 0.4587010 0.5255405 0.1831661 0.3289359 0.3643360
```

$\hat{\pi}_i$

π_i 0 1 1 0 0

$$\hat{\pi}_i = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 Age+\hat{\beta}_2 1_F}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 Age+\hat{\beta}_2 1_F}}$$

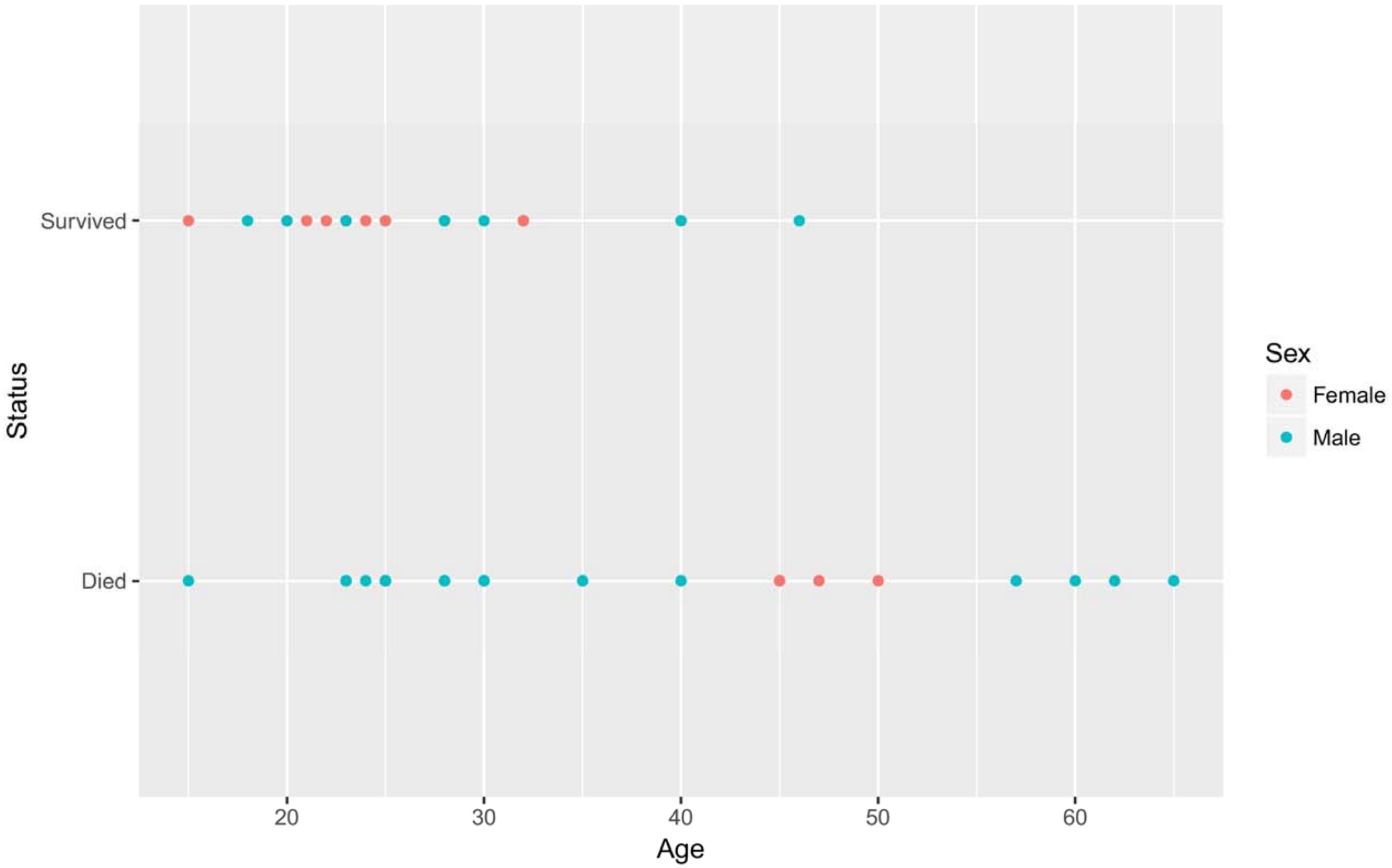
$$\text{logit}(\pi) = \beta_0+\beta_1 Age+\beta_2 1_F$$

$$\frac{\pi}{1-\pi} = e^{\beta_0+\beta_1 Age+\beta_2 1_F}$$

$$\pi = \frac{e^{M}}{1+e^{M}}$$

Case 3 Plots: Of Data

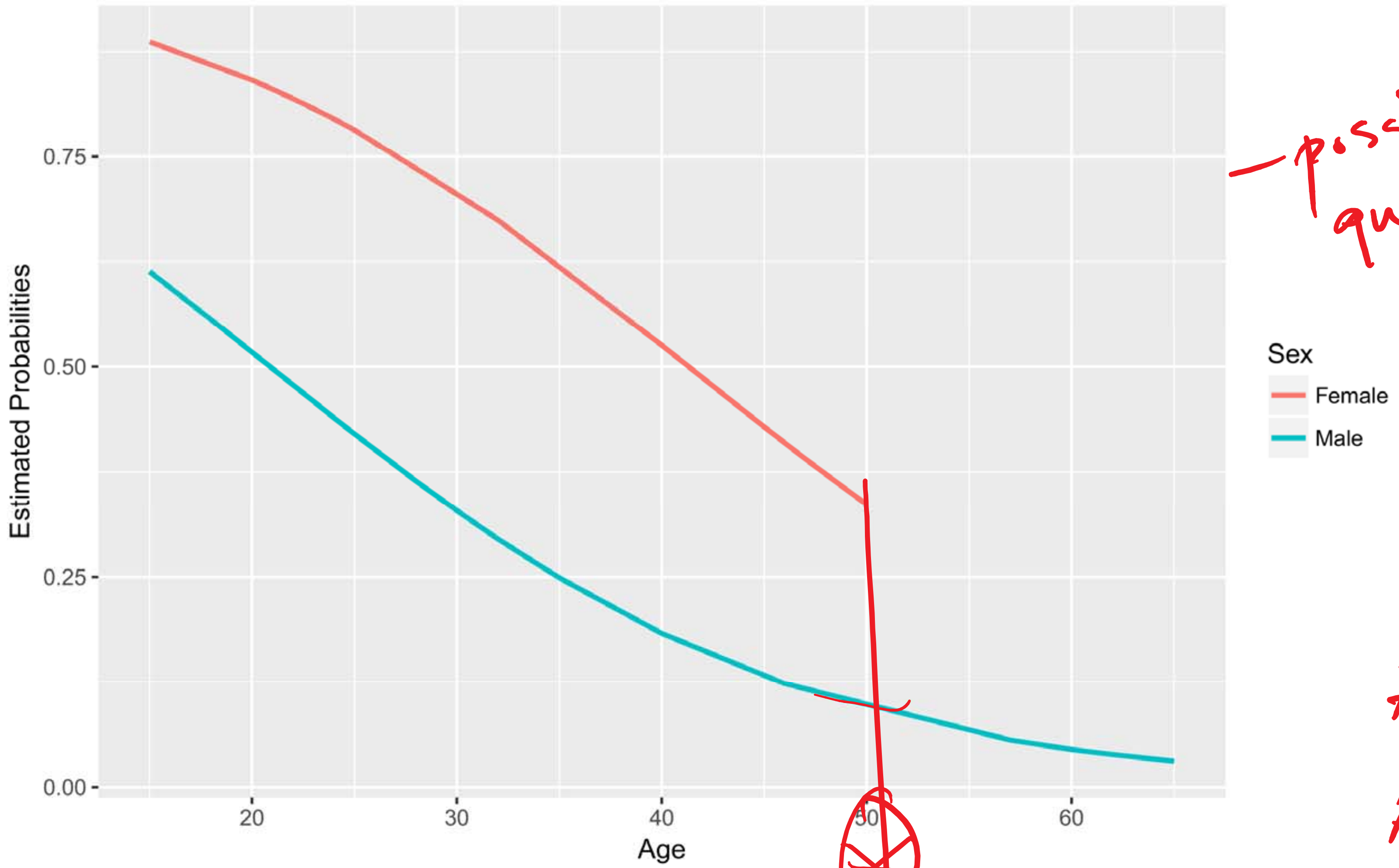
```
#contrasts(Status)
library(ggplot2)
ggplot(donner, aes(x=Age, y=Status, color=Sex))+geom_point()
```



Case 3 Plots: Additive Logistic Regression Model

```
ggplot(donner,aes(x=Age, y=phats))+ylab("Estimated Probabilities")+
  geom_line(aes(color=Sex), size=1)
```

$\hat{\pi}_i$



– possible quadratic effect of Age.

Age^2

$Age^2 * Sex$