Introduction to Logistic Regression (Thanks to Dr. Cunningham for the Example)

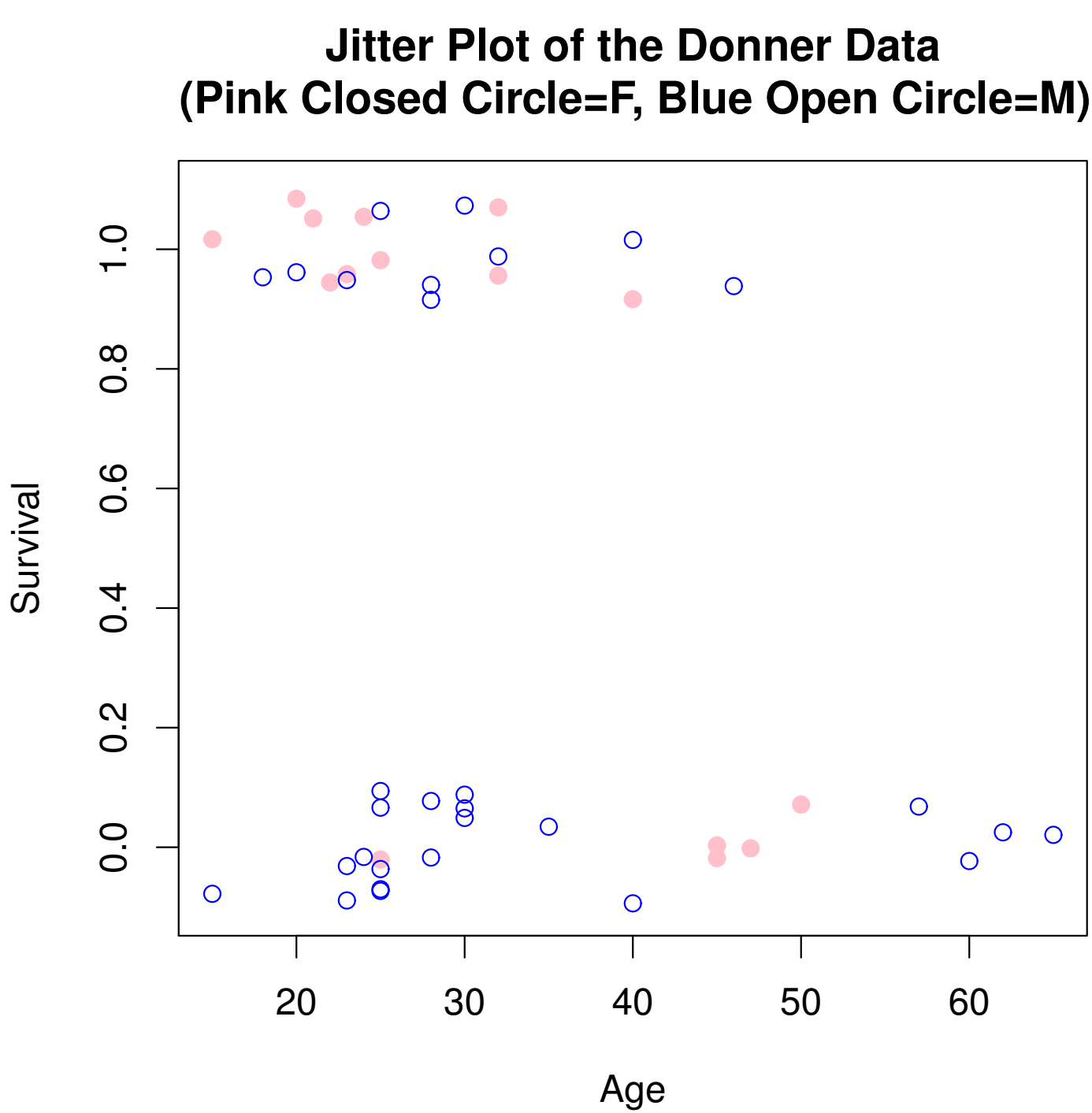
- In 1846 the Donner and Reed families left Springfield, Illinois, for California by covered wagon. In July, the Donner Party, as it became known, reached Fort Bridger, Wyoming. There its leaders decided to attempt a new and untested route to the Sacramento Valley. The group became stranded in the eastern Sierra Nevada mountains when the region was hit by heavy snows in late October. By the time the last survivor was rescued on April 21, 1847, 40 of the 87 members had died from famine and exposure to extreme cold.

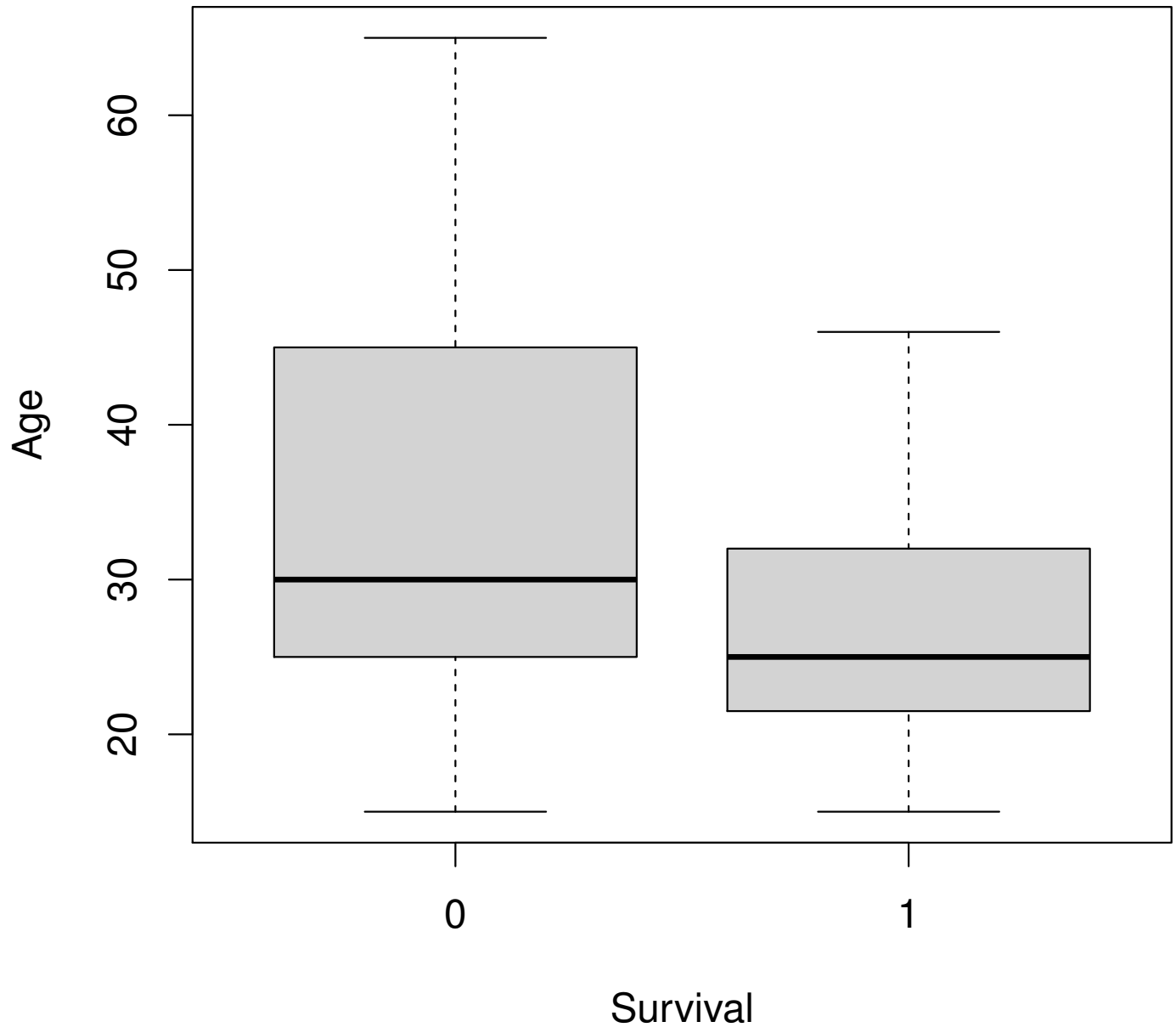
- Objective of this exercise: model probability of survival as a function of age and sex.

The Data

```
> don=read.csv('.../donner.csv',header=T)
> head(don)
  Age Sex Survival
1  23   1        0
2  40   0        1
3  40   1        1
4  30   1        0
5  28   1        0
6  40   1        0
```

- Data file `donner.csv` on sakai
- 45 individuals, Age as given, Sex=1 for male, 0 for female, Survival=1 for alive, 0 for dead


Jitter Plot of the Donner Data
(Pink Closed Circle=F, Blue Open Circle=M)
Survival
0.0
0.2
0.4
0.6
0.8
1.0
20
30
40
50
60
Age


Boxplot of Ages by Survival Status
Age
20
30
40
50
60
0
1
Survival

Initial Summaries

- Women: 10 survive, 5 die
- Men: 10 survive, 20 die
- Women do better than men
- Also, it's clearly advantageous to be young
- For the next part of the presentation, we will focus on the Age factor and look at Sex only later

Regression of Survival on Age

- $y_i = 1$ if person i is alive, 0 if dead
- x_i is the age of person i
- Why not do a simple regression of y_i on x_i?
  - Well you *could* do this, it's not a terrible idea, but ...
  - No guarantee that predicted value $\widehat{y}_i$ is between 0 and 1, so how to interpret?
  - y_is are not normally distributed, that was one of our assumptions (but it's not essential)
  - If $y_i = 1$ with probability p_i, 0 with probability $1 - p_i$, the variance of y_i is $p_i(1 - p_i)$, so it's not constant, and that *is* an important assumption
  - So, do something different ...

Logistic Regression

- $p_i = \Pr\{y_i = 1\}$
- Define $\eta_i = \log\left(\frac{p_i}{1-p_i}\right)$, called *logit* of p_i
- Inverse relationship: $p_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$
- We can allow $-\infty < \eta_i < \infty$, always gives $0 < p_i < 1$ (limits 0 or 1 as $\eta_i \to -\infty$ or $+\infty$)
- Suppose $\eta_i = \beta_0 + \beta_1 x_i$
- Extension to p covariates: $\eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$
- Fit by `glm` function in R with *family=binomial*. Estimation is by *method of maximum likelihood* but we won't go into details about that (come to STOR 557 if you want to learn)

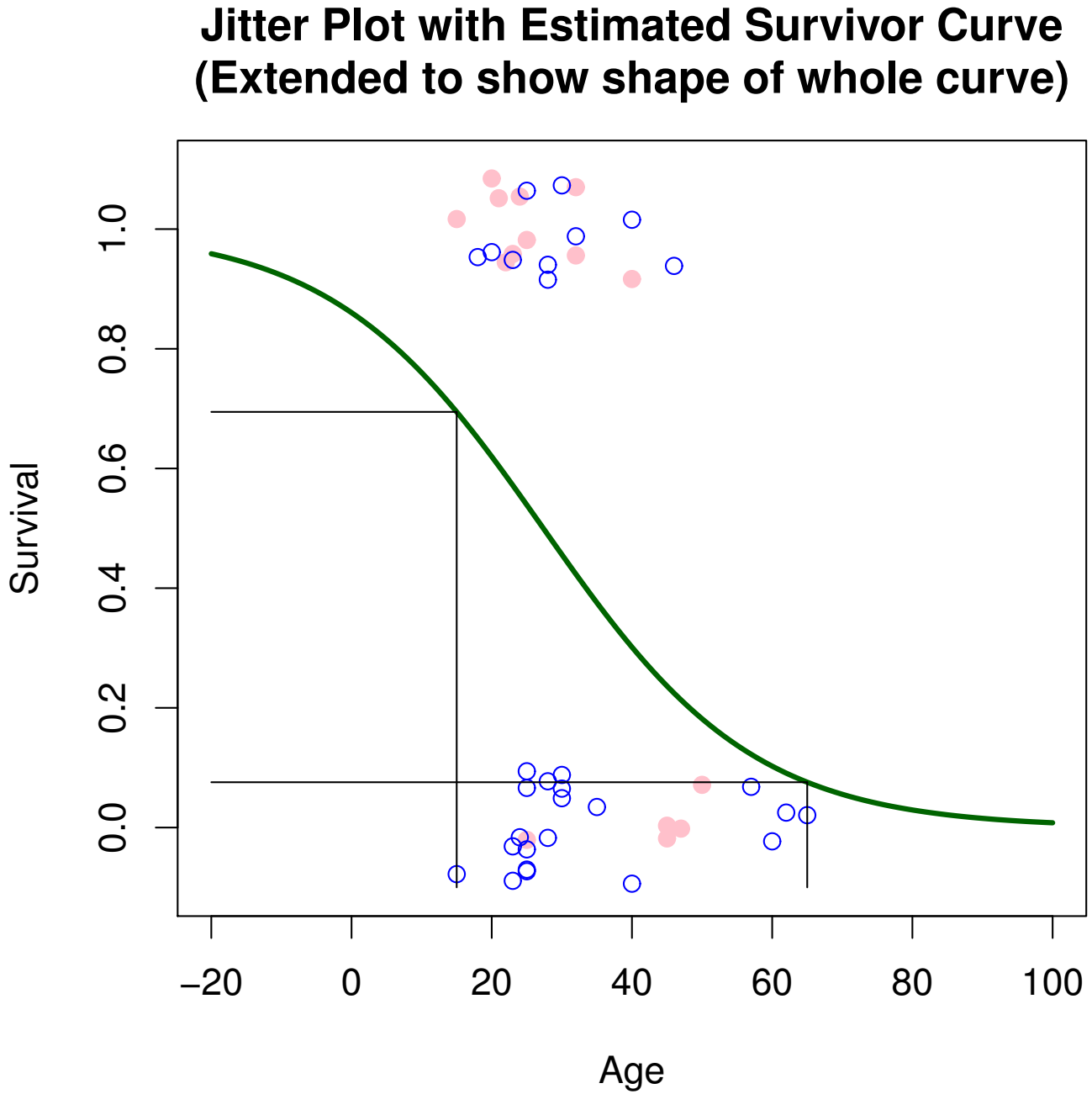
Fitting in R

```
> m1=glm(Survival~Age,family=binomial,don)
> summary(m1)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5401  -1.1594  -0.4651   1.0842   1.7283

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.81852    0.99937   1.820   0.0688 .
Age         -0.06647    0.03222  -2.063   0.0391 *
---
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: 56.291  on 43  degrees of freedom
AIC: 60.291
Number of Fisher Scoring iterations: 4
```

- For an individual of age x, the estimated probability of survival is $\frac{e^{1.81852-0.06647x}}{1+e^{1.81852-0.06647x}}$



Probability of survival is 0.69 at age 15, 0.08 at age 65

Uncertainties of the Estimated Probabilities

Consider probabilities for individuals aged 15 and 65

```
predict(m1,type='response',newdata=data.frame(Age=c(15,65)),se.fit=T)
# $fit
# 0.6945470 0.0757146
# $se.fit
# 0.11938933 0.08330818
```

Estimated probabilities 0.695, 0.076, standard errors 0.119, 0.083

Here's a better way to do it (calculate confidence intervals on logit scale and back-transform to probabilities):

```
p1=predict(m1,type='link',newdata=data.frame(Age=c(15,65)),se.fit=T)
eta=p1$fit;exp(eta)/(1+exp(eta))
eta=p1$fit+1.96*p1$se.fit;exp(eta)/(1+exp(eta))
eta=p1$fit-1.96*p1$se.fit;exp(eta)/(1+exp(eta))
# 0.6945470 0.0757146
# 0.8726354 0.4578956
# 0.430077229 0.007881829
```

95% confidence intervals for probability of survival are (0.43,0.873) at age 15 and (0.008, 0.458) at age 65

Logistic Regression with Multiple Covariates

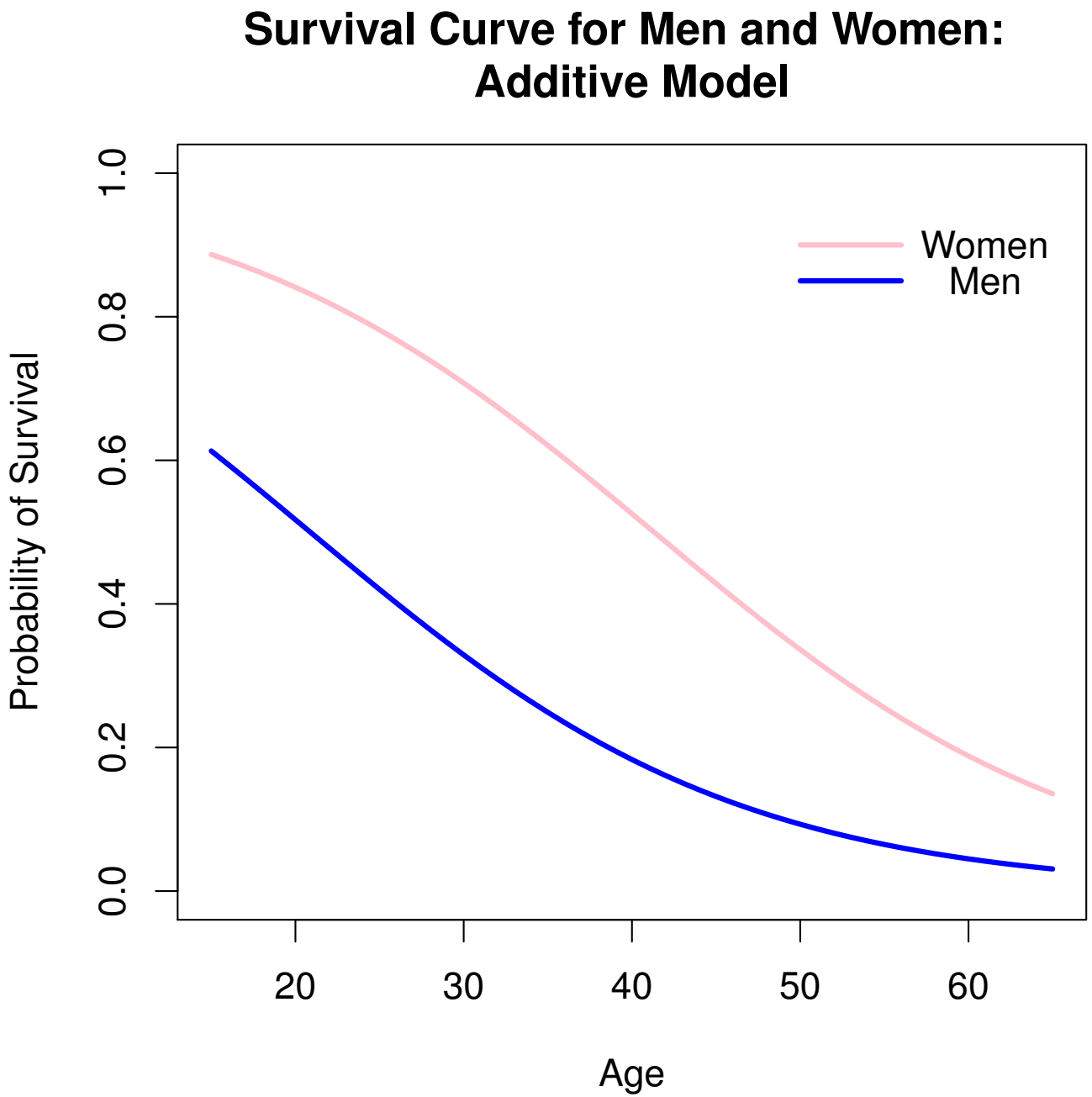
- We can extend the `glm` function to include as many covariates as we like, e.g. with 3 covariates `glm(y~x1+x2+x3,family=binomial)` etc.

```
> m2=glm(Survival~Age+Sex,family=binomial,don)
> summary(m2)
...
            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 *
Sex         -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
```

Interpretation

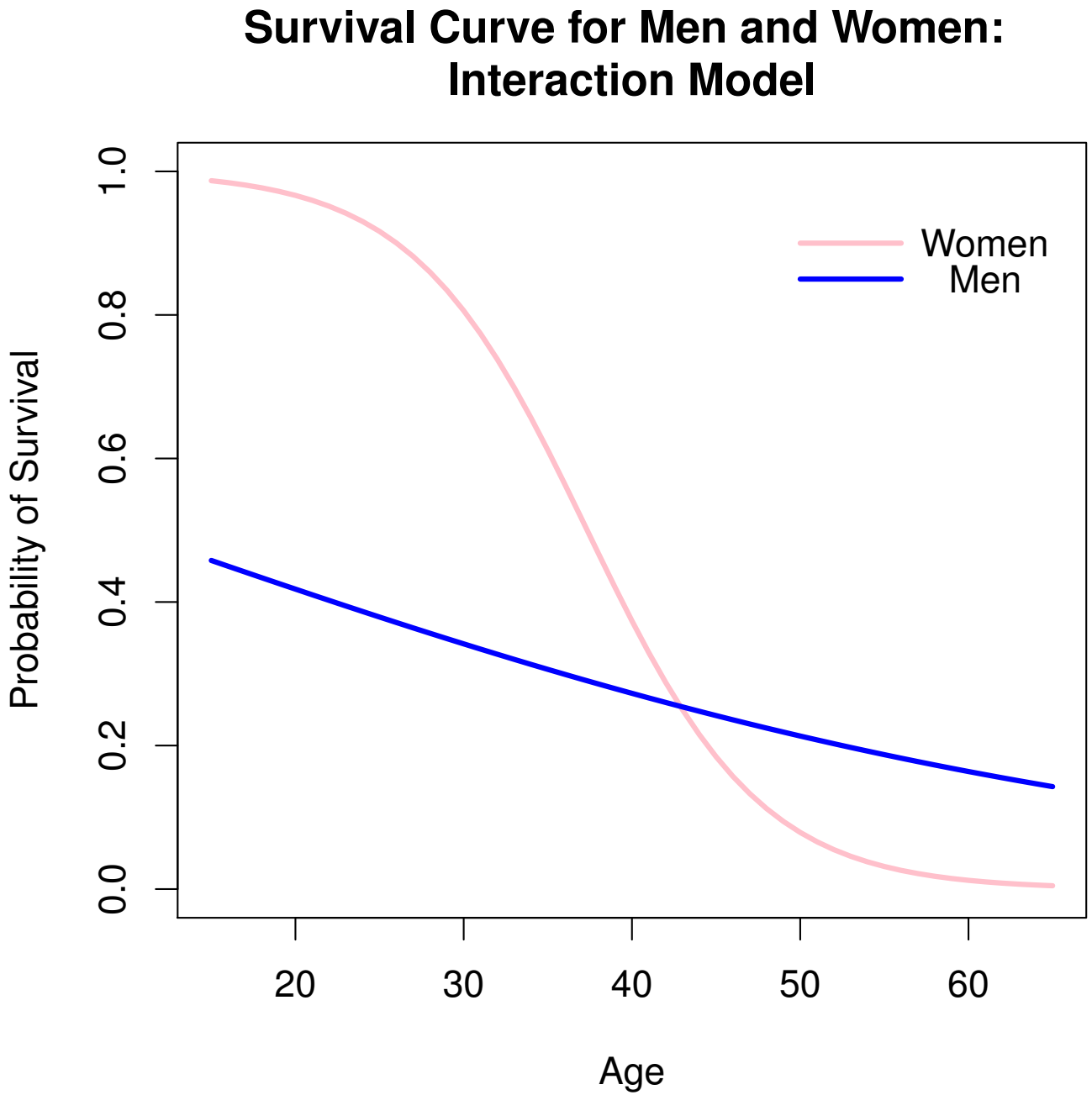
- Adding one year of age changes your odds of survival (probability that you survive divided by probability that you do not survive) to $e^{-0.07820} \approx 0.925$ of its previous value (a reduction of 7.5% in your odds of survival)
- Being male (compared with female) means your odds of survival are multiplied by $e^{-1.59729} \approx 0.2$ (an 80% reduction in *odds*, but that's not the same as probability)
- We can calculate separate survival curves for men and women (next plot)
- For either men or women of any given age, we could compute a probability of survival (with either a standard error or a confidence interval) by using the `predict` command, same as in the one-variable case


Survival Curve for Men and Women:
Additive Model
Probability of Survival
0.0
0.2
0.4
0.6
0.8
1.0
20
30
40
50
60
Age
Women
Men

Comparing Multiple Models

```
m0=glm(Survival~1,family=binomial,don)
m3=glm(Survival~Age*Sex,family=binomial,don)
anova(m0,m1,test='Chi')
anova(m1,m2,test='Chi')
anova(m2,m3,test='Chi')
```

- `m0` included as baseline model (intercept but no covariates); `m3` includes product of Age and Sex (interaction model)
- `anova` carries out an ANOVA test for nested models; `test='Chi'` computes a p-value using a chi-squared test. This is similar to an F test for a linear model, but less precise because the chi-squared test is only approximate (the F test is exact when the normal-theory assumptions are satisfied)
- In this case the three p-values are 0.019, 0.035, 0.048 — all are < 0.05, suggesting that each model is significantly better than the previous one. However, this is harder to interpret, because the `m3` model produces a counter-intuitive pair of survival curves

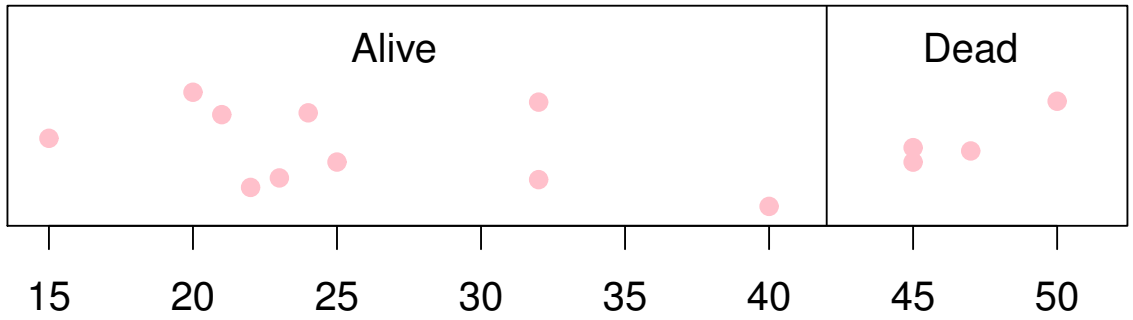


What does it mean that the two curves cross over?

More on the Interaction Model

- The interaction model is equivalent to fitting separate survivor curves to the men and women
- There was one 25-year-old woman who died (row 11). If we omit her, the plot for the rest of the women looks like this:

Jitter Plot for Women only Omitting Row 11



- All the women of ages ≤ 40 lived, those ≥ 45 died
- If we fitted the logistic regression curve just to them, the slope would in effect be infinite — a well-known problem with logistic regression when the data can be split into two disjoint groups
- The actual curve doesn't look like this because of that one 25-year-old who died, but it's still probably a much steeper slope than it should be
- The additive model avoids this artifact and is probably more realistic

Validating the Model: A Thought Experiment

- Same idea as cross-validation
- Omit each observation in turn
- Refit both models (`m2` and `m3`) to the other 44 data points: if the predictive probability for the omitted datapoint is > 0.5 we guess "survived"; otherwise "did not survive"
- The proportion of correct guesses is 71% under model `m2`, 73% under model `m3`
- Either model is a significant improvement on random guessing!

Summary of Logistic Regression

- Logistic regression is appropriate when the responses are all 0 or 1 (binary data)
- The model fits $\eta_i = \log \frac{p_i}{1-p_i}$ as a linear combination of covariates, where $p_i = \Pr\{y_i = 1\}$
- We can estimate parameters, standard errors, etc., very similarly to standard linear models, but using `glm` with `family=binomial`
- We showed how to use `anova` to compare nested models. For non-nested, can use AIC, BIC, etc. similarly to linear models
- We can estimate probabilities (with standard errors or confidence intervals) for new data using the `predict` command
- We can also use cross-validation
- For more detail and other kinds of glms, come to STOR 557!

Topics I Wanted to Cover but Didn't

- Diagnostics for Outliers and Influence (Chapter 5 of text)
- Analysis of Variance Models (not to be confused with Analysis of Variance Tests)

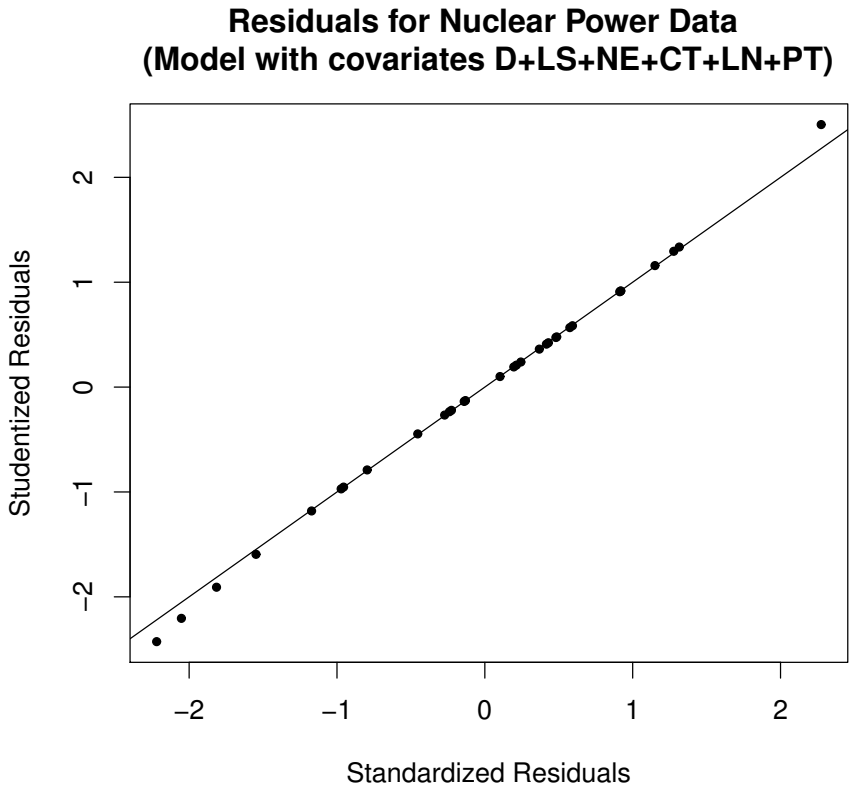
Diagnostics for Outliers and Influence

- Studentized residuals
- Hat values and Leverage
- Influence: Cook's D statistic

Studentized Residuals (pp. 353–354 of course text)

- Usual linear model — p covariates plus an intercept
- $T_i = \frac{y_i - \widehat{Y}_{(-i)}(x_{i1},\ldots,x_{ip})}{\widehat{\sigma}_{(-i)}/\sqrt{1-h_{ii}}}$ where
  - $\widehat{Y}_{(-i)}(x_{i1},\ldots,x_{ip})$ is the prediction of the i th observation *omitting the i th observation itself from the regression model fit*
  - $\widehat{\sigma}_{(-i)}$ is the estimated σ omitting the i th observation itself from the regression model fit
  - h_{ii} is the i th hat value (diagonal entry of $H = X(X^TX)^{-1}X^T$)
- Important fact: if all model assumptions (including independent normal errors) are satisfied, then $T_i \sim t_{n-p-2}$ (exactly)
- `rstudent` in R

Application to Nuclear Power Data



- 6-covariate model as in previous examples
- The studentized residuals are slightly more spread out than the standardized residuals — extremes at –2.43, 2.50 instead of –2.22, 2.28
- p-values `2*pt(-2.50,24)=0.02` not too extreme by Bonferroni correction — interpreted as a test for outliers, conclude no outliers

Hat Values and Leverage (Section 5.3 of course text)

- Defined just by the x values — often used to characterize which covariates are extreme

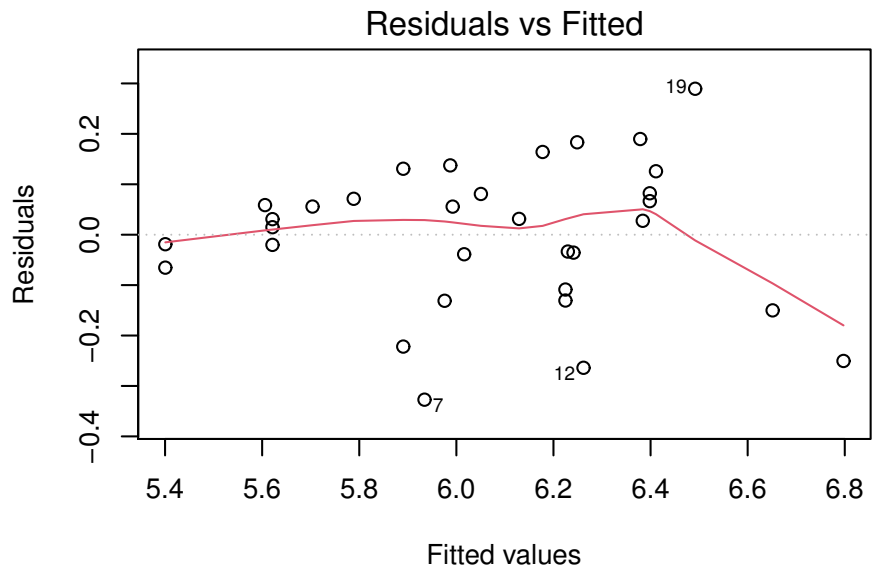
- $\sum_i h_{ii} = p + 1$ (assuming intercept) — mean is $\frac{p+1}{n}$ — a value greater than $\frac{2(p+1)}{n}$ is considered "worthy of further examination"

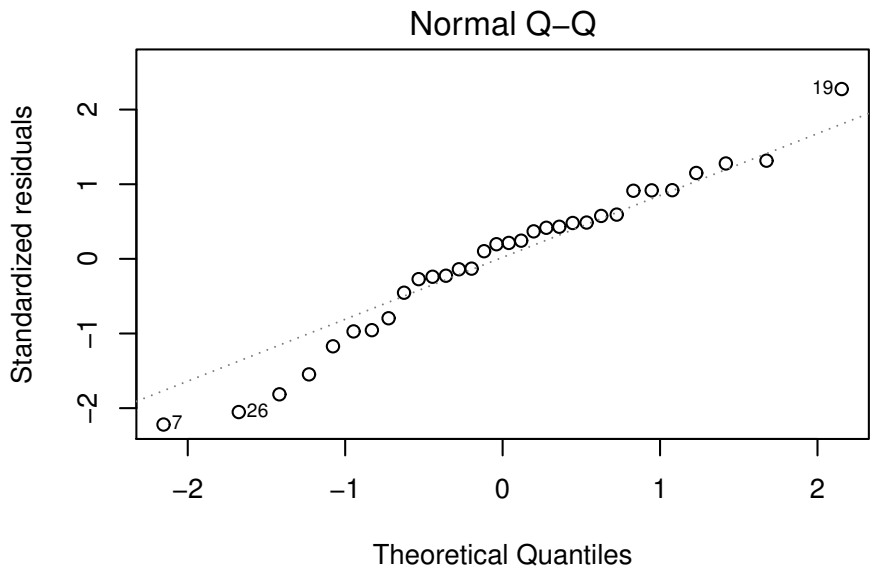
- `hatvalues` in R

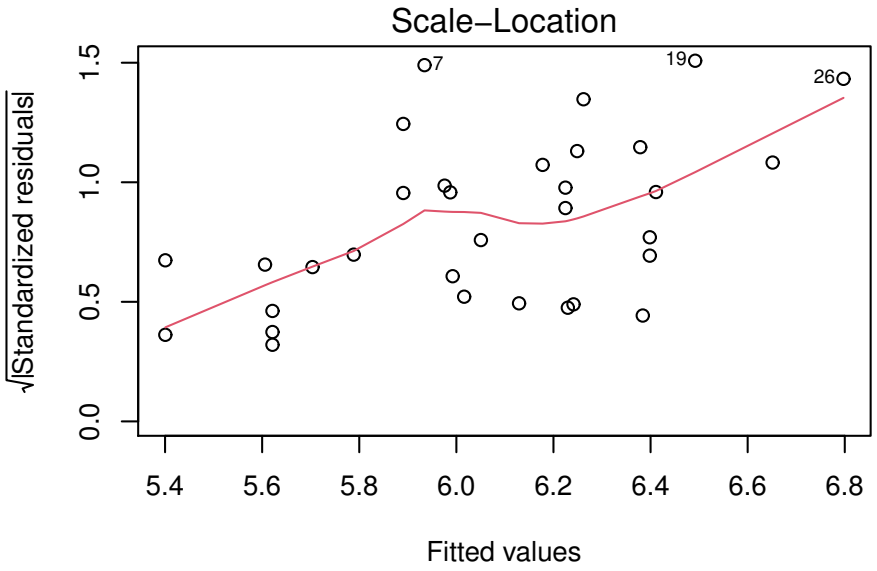
  Nuclear power data: largest value is 0.4126 (observation 26) but $\frac{2(p+1)}{n} = \frac{2\times 7}{32} = 0.4375$.

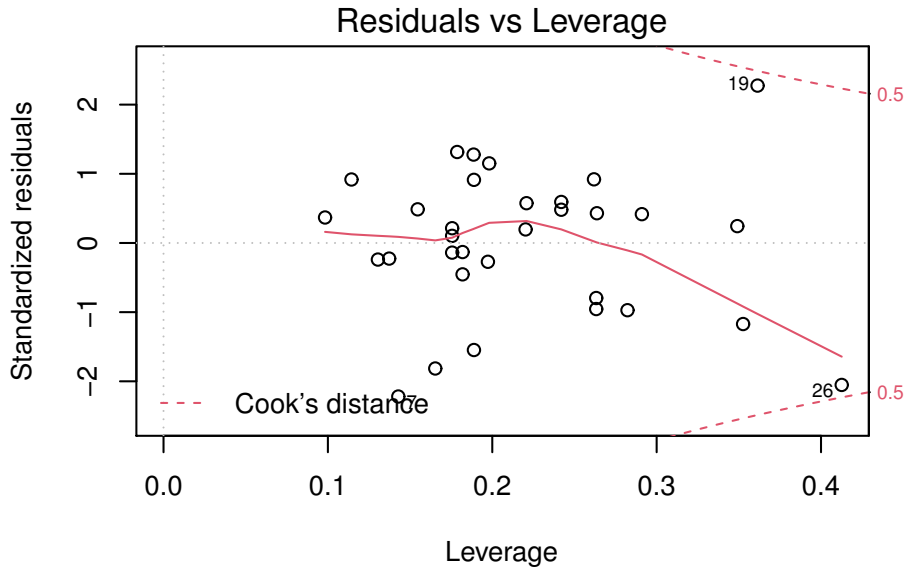
Cook's D Statistic
(Section 5.4 of course text)

- What it is (in words): a measure of how much the ith observation *influences* the vector of regression coefficients β
- Flags points that have *both* high leverage *and* are outliers as measured by the standardized or studentized residuals
- `cooks.distance(modelname)` in R, or use `plot(modelname)` for a series of diagnostic plots
- DFFITS is almost the same thing (not obvious)
- What are high values? Various rules of thumb in the literature, but R uses D=0.5 and D=1 as warning levels
- Nuclear power data: largest two values are 0.418 (observation 19) and 0.423 (observation 26)
- Donner data, interaction model: largest value is 0.401 (observation 11).

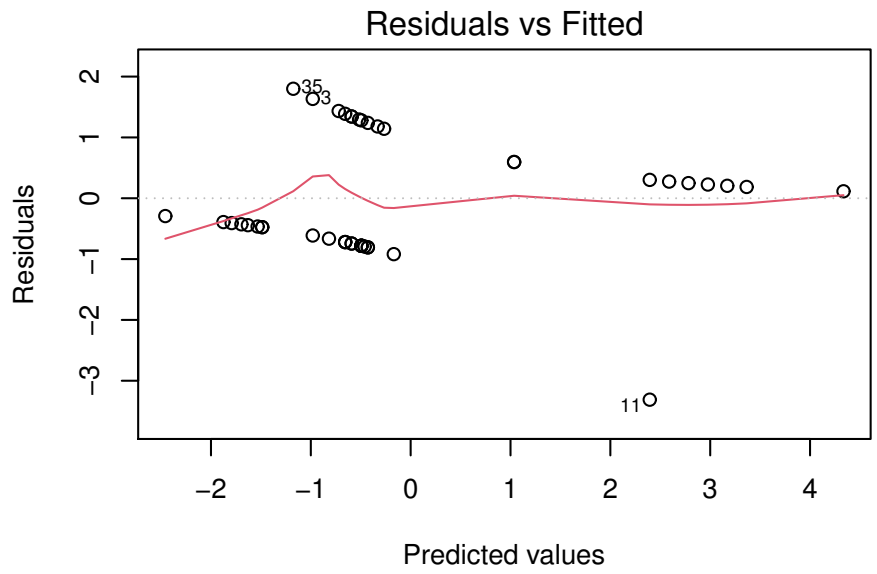


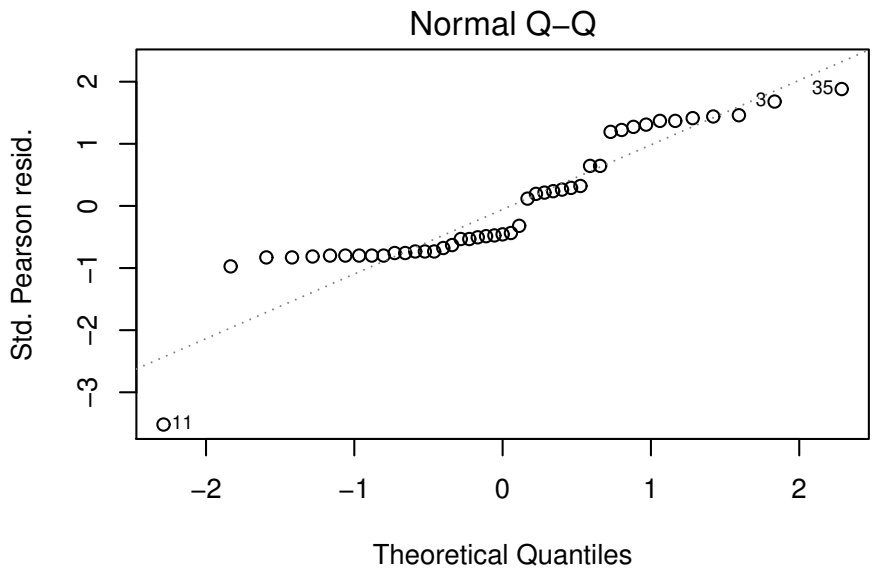


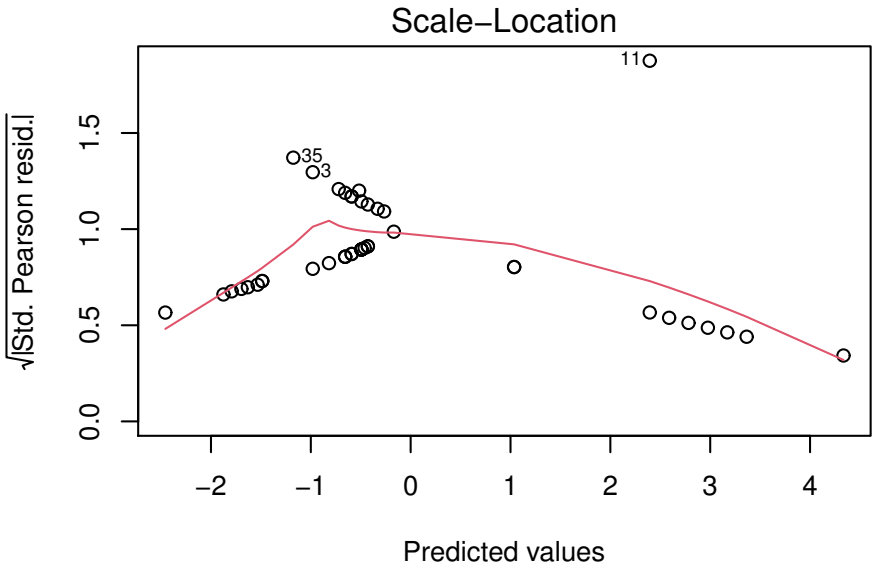


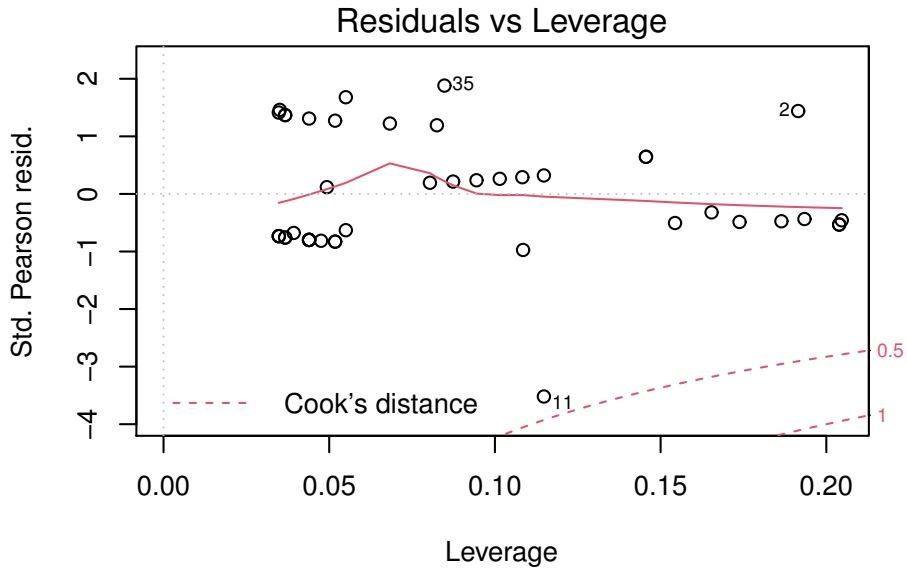


`plot(lm1)` for nuclear power data









`plot(m3)` for Donner data (Age*Sex model)

Analysis of Variance

Recall left-out part of HW3:

Comparing multiple means (Analysis of Variance)

What if we want to compare more than one team? In lecture, you learned about using an Analysis of Variance (ANOVA) F-test to check if more than two means are equal. We will use the function `aov( )` to find out if the big three North Carolina teams - UNC, Duke, and NC State - all tend to score the same number of points.

Question 10

a. Make a dataset called `nc_games` that includes only games for the North Carolina teams, and then alter the code below to create a box plot of the scores for the three North Carolina teams. Does it look like any of the means are significantly different?

b. Perform an ANOVA F-test on the means. Interpret the output. Is there evidence that the average scores of the three teams are not all equal?

R code

```
> bb=read.csv('basketball.csv',header=T)
> bb1=bb[bb$Team=='North Carolina'|bb$Team=='Duke'|
bb$Team=='North Carolina State',]
> a1=lm(Team.Score~Team,bb1)
> a0=lm(Team.Score~1,bb1)
> anova(a1,a0)
..
  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1     97 12986
2     99 13976 -2    -989.8 3.6966 0.02837 *
> # second way of doing the same thing
> aov(Team.Score~Team,bb1)
Sum of Squares    989.795 12986.315
Deg. of Freedom         2        97
> Fstat=(989.795/2)/(12986.315/97)
> 1-pf(Fstat,2,97)
[1] 0.0283682
> # third way
> summary(aov(Team.Score~Team,bb1))
            Df Sum Sq Mean Sq F value Pr(>F)
Team         2    990   494.9   3.697 0.0284 *
Residuals   97  12986   133.9
```

What's Going On Here?

- *Analysis of Variance* refers to a special class of linear models where the covariates are all factor variables
- After reducing the data frame, `Team` is a factor variable with just three values, `North Carolina, Duke` and `North Carolina State`.
- The command `a1=lm(Team.Score~Team,bb1)` fits the model where `Team.Score` is the response and `Team` the (sole) covariate
- However, because `Team` has three `levels`, and the model is still assumed to include an intercept, there are four parameters (the intercept, plus one term each for Duke, North Carolina and North Carolina State) but only three degrees of freedom to estimate them. R solves this problem by defaulting that the first level alphabetically (here, Duke) is automatically given the value 0 and the others expressed as difference from Duke.
- Interpretation: North Carolina scores on average 0.81 points per game more than Duke, and North Carolina State scores 6.24 points fewer
- The F test that all three means are equal has a p-value of 0.0284, which implies strong but not certain evidence against that hypothesis.

Two-way ANOVA

- `Team.Location` is also a factor variable with three levels, `Home`, `Away` and `Neutral`
- We can test both factors at once with a model like
  `a2=lm(Team.Score~Team+Team.Location,bb1)`
- We can also add an interaction, e.g.
  `a3=lm(Team.Score~Team*Team.Location,bb1)`
- Practical question: what does the interaction model test for, in language that any basketball fan would understand?
- We can use `anova` or `aov` to test model a2 against a1 or a3 against a2
- `Team.Location` is significant, but the interaction is not