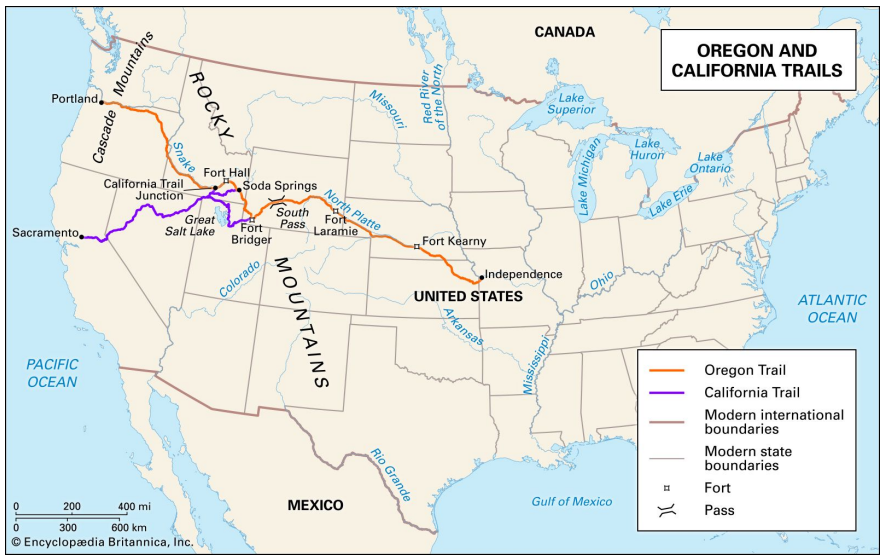
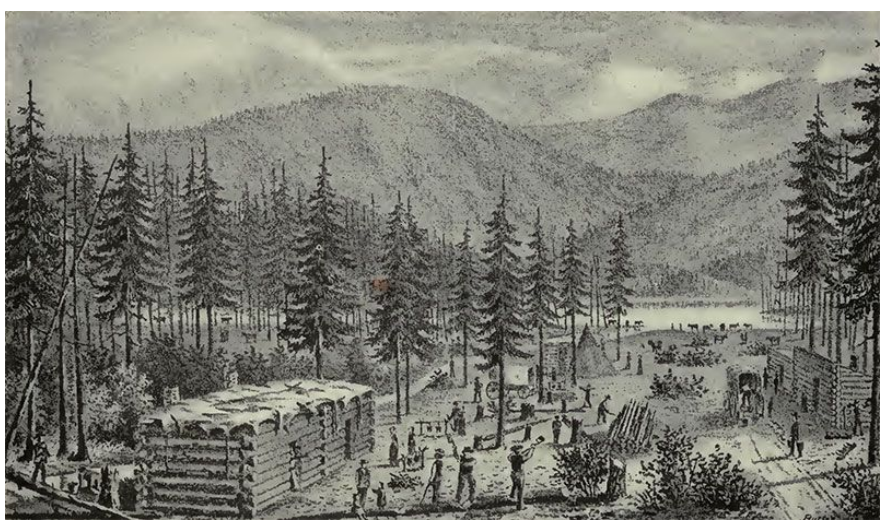
Logistic Regression

Week 5

The Donner Party

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 rote to the Sacramento Valley. Having reached its full size of 89 people and 20 wagons, the party was delayed by a difficult crossing of the Wasatch Range and again in the crossing of the desert west of the Great Salt Lake. 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, **41 of the 89 members had died** from famine and exposure to extreme cold.





adapted from Ramsey, F.L. and Schafer, D.W. (2002).

The Donner Party

> **41 of the 89 members died**

- what characteristics of those who died were different from those who survived?
  - sex?
  - age?
- can we build a model to predict who would survive based on these characteristics?

The Donner Party

▶ Code

```
# A tibble: 6 × 3
    Age Sex    Status
  <dbl> <fct>  <fct>
1    23 Male   Died
2    40 Female Survived
3    40 Male   Survived
4    30 Male   Died
5    28 Male   Died
6    40 Male   Died
```

```
Rows: 43
Columns: 3
$ Age    <dbl> 23, 40, 40, 30, 28, 40, 45, 62, 65, 45, 25, 28, 28, 23, 22, 23,…
$ Sex    <fct> Male, Female, Male, Male, Male, Male, Female, Male, Male, Femal…
$ Status <fct> Died, Survived, Survived, Died, Died, Died, Died, Died, Died, D…
```

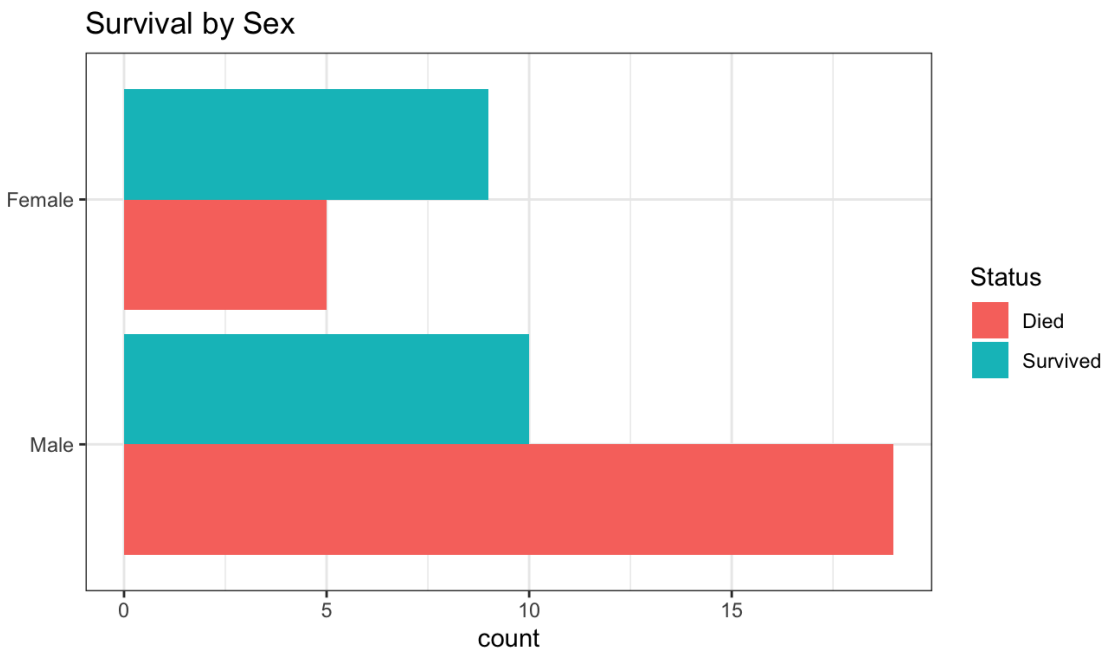
- data for “adults”, i.e. people who were over 15 years old

The Donner Party

▶ Code

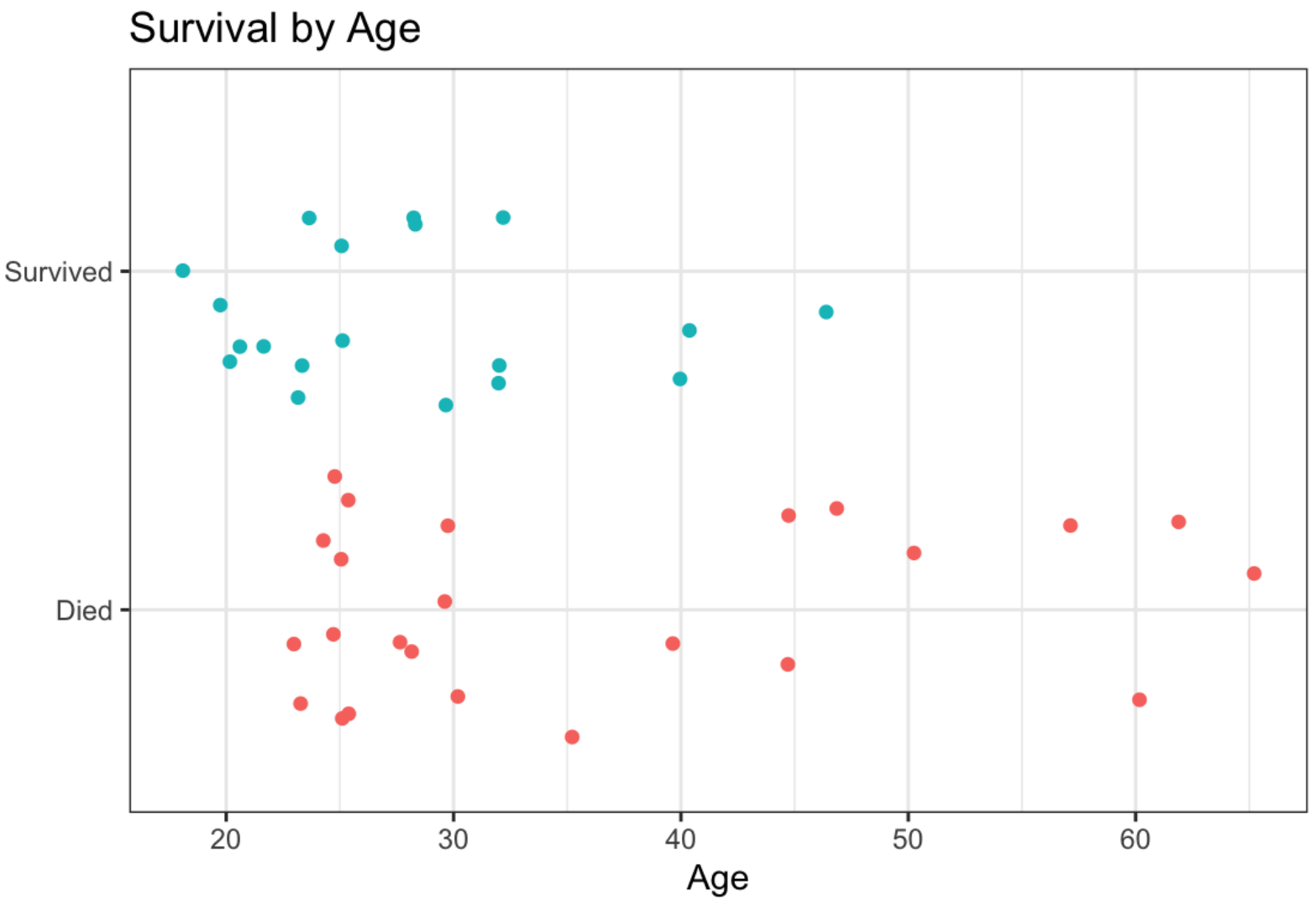
```
# A tibble: 4 × 3
# Groups:   Status [2]
  Status   Sex    count
  <fct>    <fct>  <int>
1 Died     Male      19
2 Died     Female     5
3 Survived Male      10
4 Survived Female     9
```

▶ Code



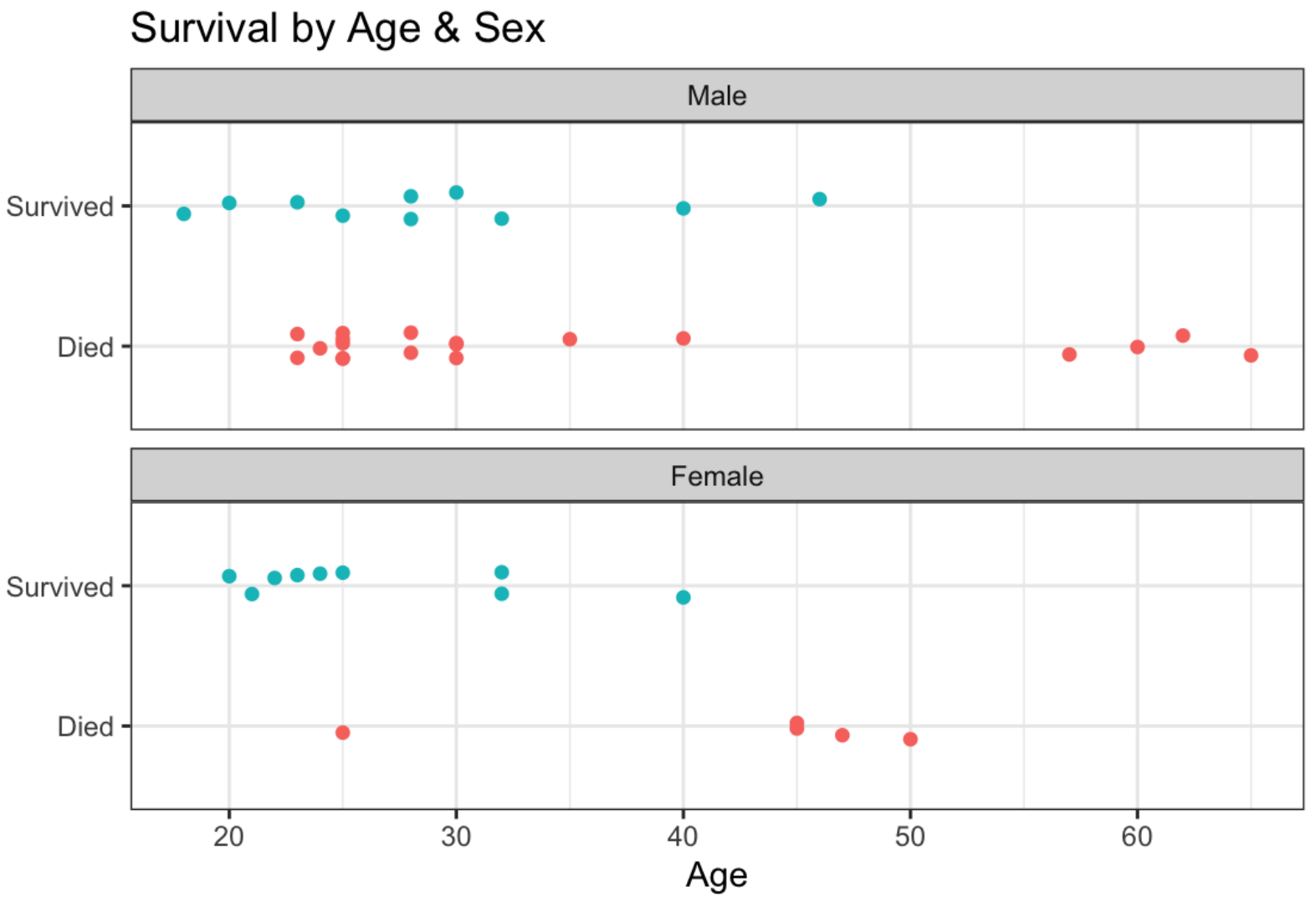
The Donner Party

▶ Code



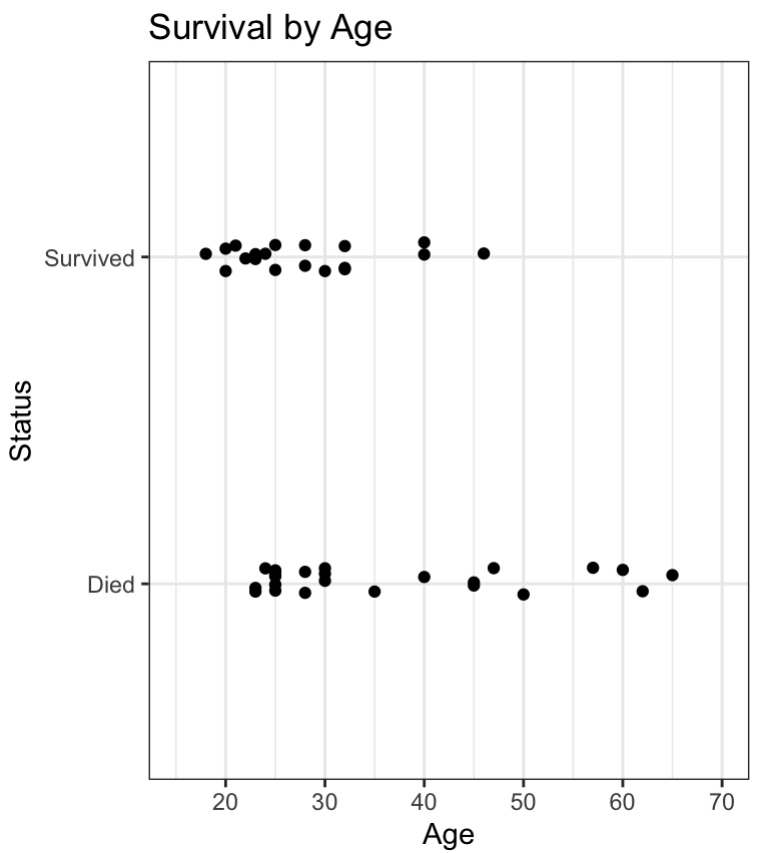
The Donner Party

▶ Code



The Donner Party

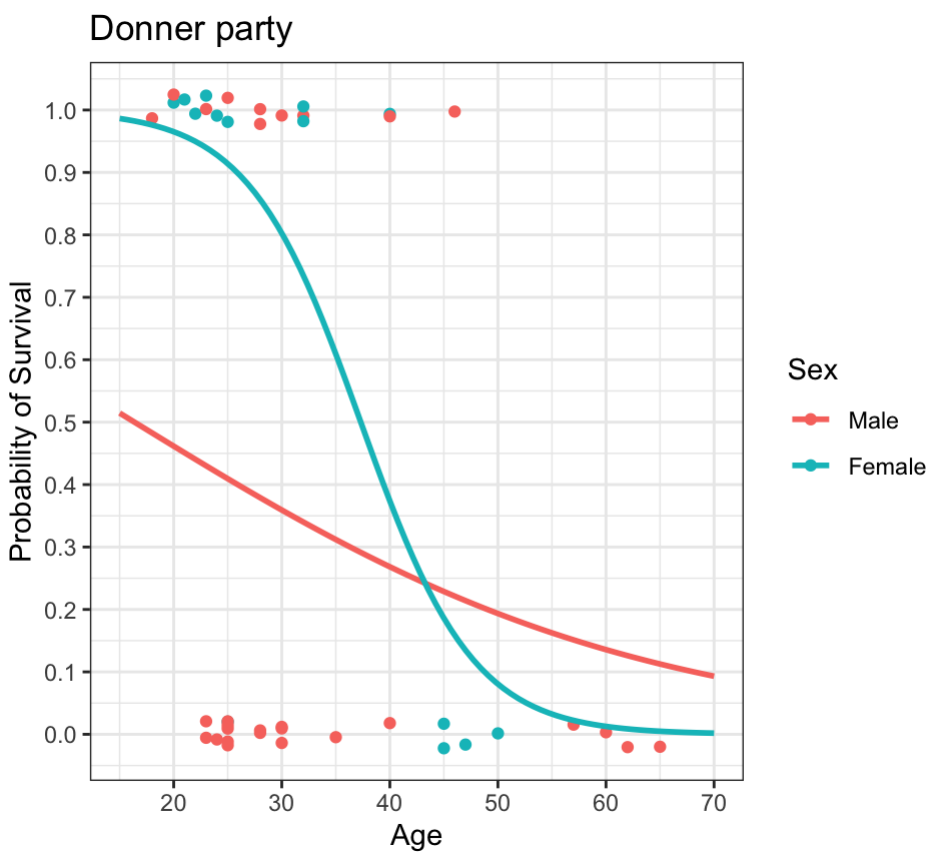
▶ Code



- what's the probability of survival given age?
- logistic regression models the **probability of a binary outcome** as a function of one or more predictors
- e.g. probability of `Status` (binary outcome: `Survived` vs `Died`) as a function of `Age` (continuous variable)

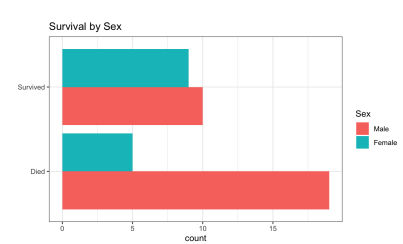
The Donner Party

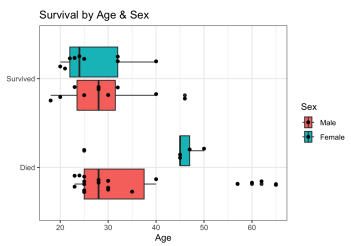
▶ Code

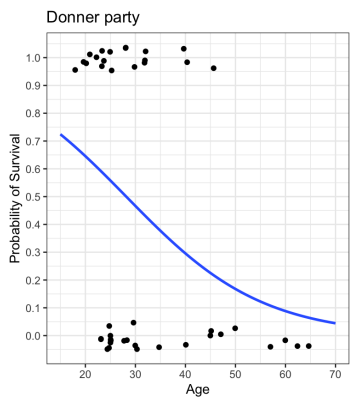


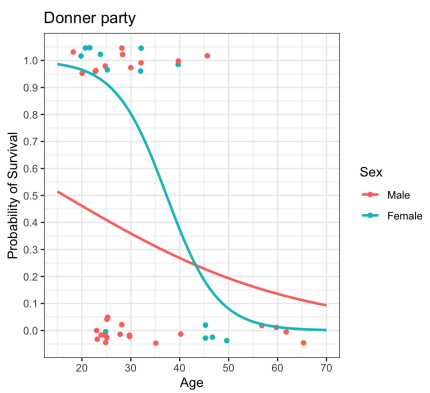
- what's the probability of survival **as a function of age, by sex**?
- we can plot a logistic regression fit in `ggplot` easily using `geom_smooth()`
- logistic regression lines are usually S-shaped (sigmoidal)

Logistic Regression









- these are all **descriptive** ways of looking at the relationship between characteristics (age, sex) and probability of survival
- logistic regression can give us a **quantitative** way of describing these relationships

Logistic Regression

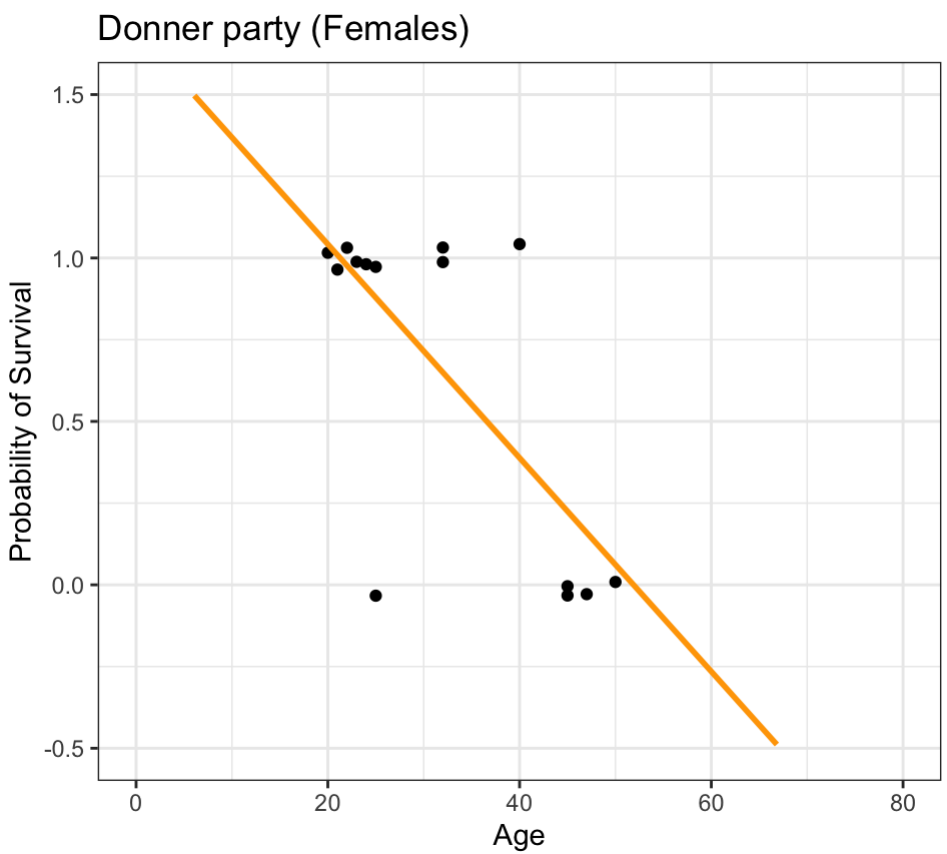
- we can also use logistic regression to **predict** outcomes for new observations
- e.g. predict 5-year survival of breast cancer patients based on tumour size, lymph node involvement, and tumour grade
- e.g. predict mortgage default based on income, credit score, and debt-to-income ratio
- **when the dependent variable is binary, we use logistic regression**
  - dependent variable is binary
  - independent variables can be continuous (e.g. Age) or categorical (e.g. Sex)

Why not linear regression?

- why do we need logistic regression?
- why not just use linear regression?
- let's try it and see what happens

Why not linear regression?

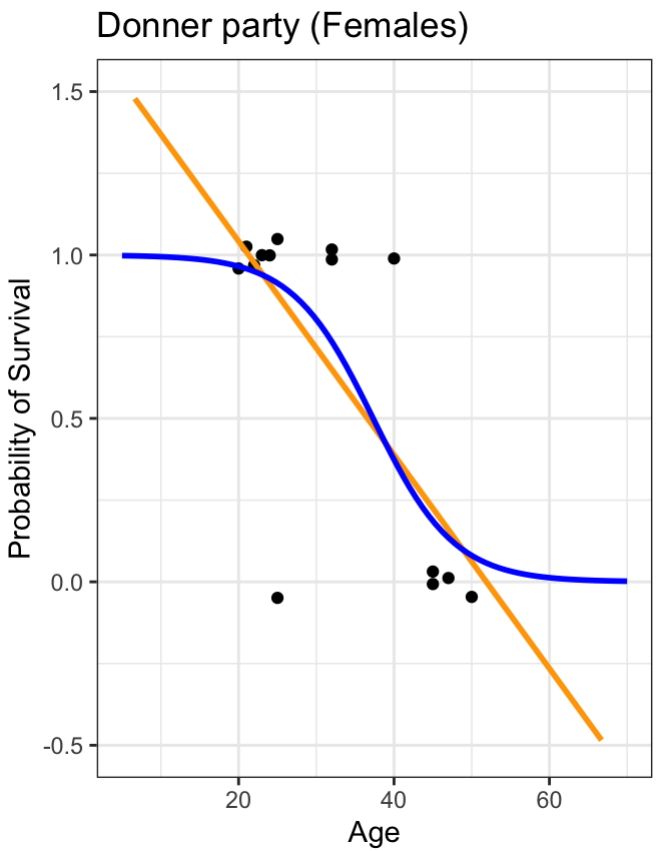
▶ Code



- linear regression model prediction extends above 1.0 and below 0.0
- doesn't make any sense given the thing we're trying to model is a probability
- probabilities can only live between 0.0 and 1.0

Why not linear regression?

▶ Code

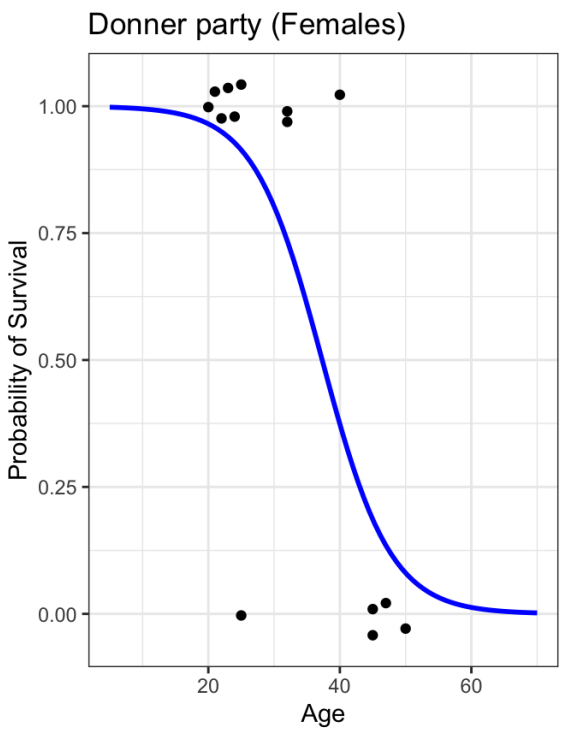


Also: we violate assumptions of linear regression

1. **linearity**: relationship between predictor and outcome is not linear
2. **normality**: residuals are not normally distributed
3. **homoscedasticity**: variance of residuals is not constant over the range of the predictor variable

Logistic Regression

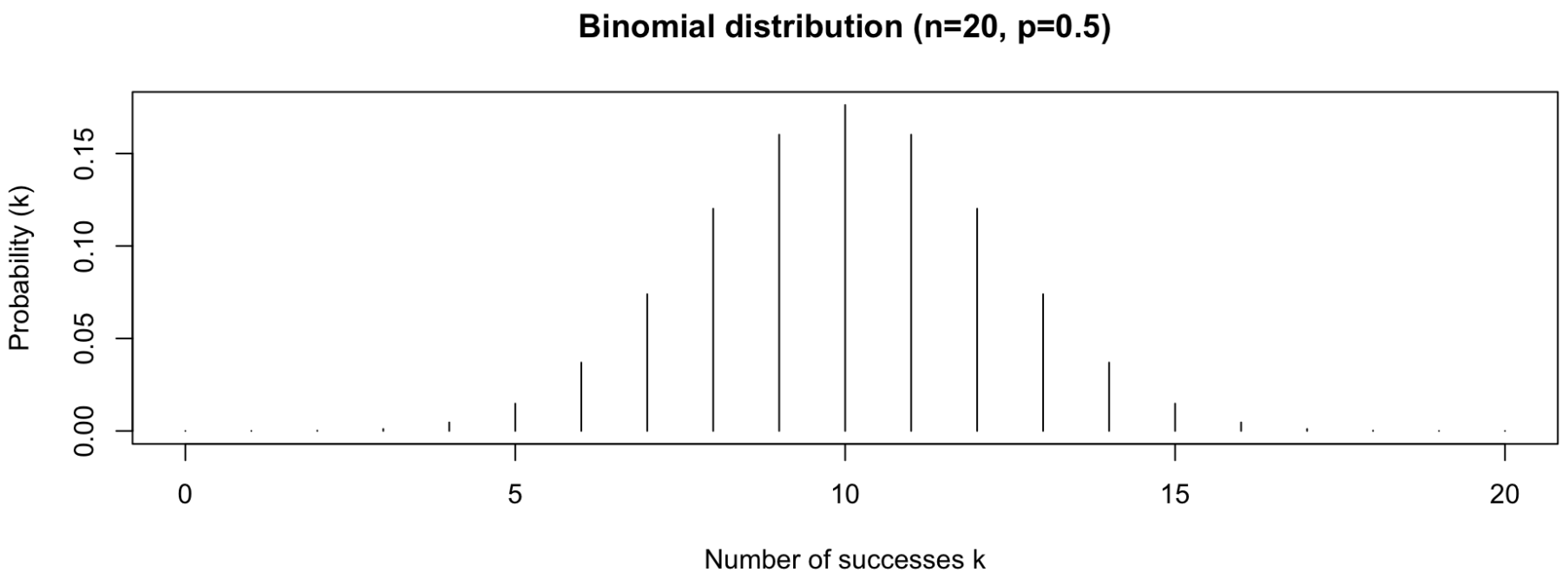
▶ Code



- we can treat Survived and Died as successes and failures arising from a binomial distribution where the probability of a success is given by *a transformation of a linear model of the predictor(s)*
- this is a general way of addressing this type of problem in regression, and the resulting models are called *generalized linear models* (**GLMs**)
- Logistic regression is just one example of a GLM
- but first: a reminder of the binomial distribution

Binomial Distribution

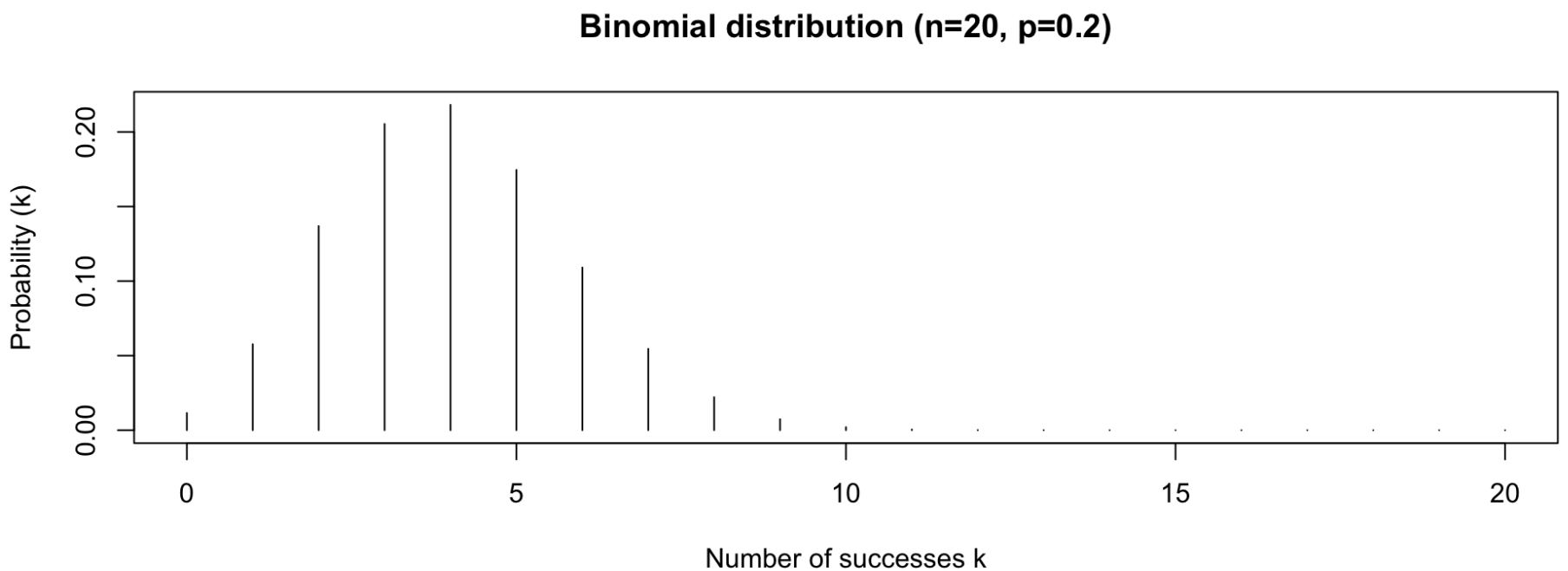
▶ Code



- the Binomial distribution has two parameters: *n* the number of trials, and *p* the probability of success on a given trial
- the Binomial distribution gives us the probability of observing *k* successes in *n* trials, given *p*
- e.g. probability of observing 3 heads in 20 coin flips, given that the probability of heads is 0.5

Binomial Distribution

▶ Code



- the Binomial distribution has two parameters: *n* the number of trials, and *p* the probability of success on a given trial
- the Binomial distribution gives us the probability of observing *k* successes in *n* trials, given *p*
- e.g. probability of observing 3 heads in 20 coin flips, given that the probability of heads is 0.2

Generalized Linear Models (GLMs)

All GLMs have four components:

1. an **outcome variable** that is being predicted and **predictor variables** to predict the outcome
2. a **probability distribution** describing the outcome variable
3. a **linear model** $\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$ where η is the linear predictor and $X_1, \ldots, X_n$ are the **predictor variable(s)**
4. a **link function** relates the linear model to the parameter of the outcome variable:
   $g(p) = \eta$
   or
   $p = g^{-1}(\eta)$

Logistic Regression

a GLM with a **binomial** distribution and a **logit** link function

1. the outcome variable is binary (e.g. Survived or Died) and the predictor variable is continuous (e.g. Age)
2. p (the probability of a success, e.g. Survival) is the parameter of the binomial distribution

Logistic Regression

a GLM with a **binomial** distribution and a **logit** link function

3. the linear model of our predictor variable is: $\eta = \beta_0 + \beta_1 \text{Age}$

4. the link function is: $logit(p) = log\left(\frac{p}{1-p}\right) = \eta$

putting it together:

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \text{Age}$$

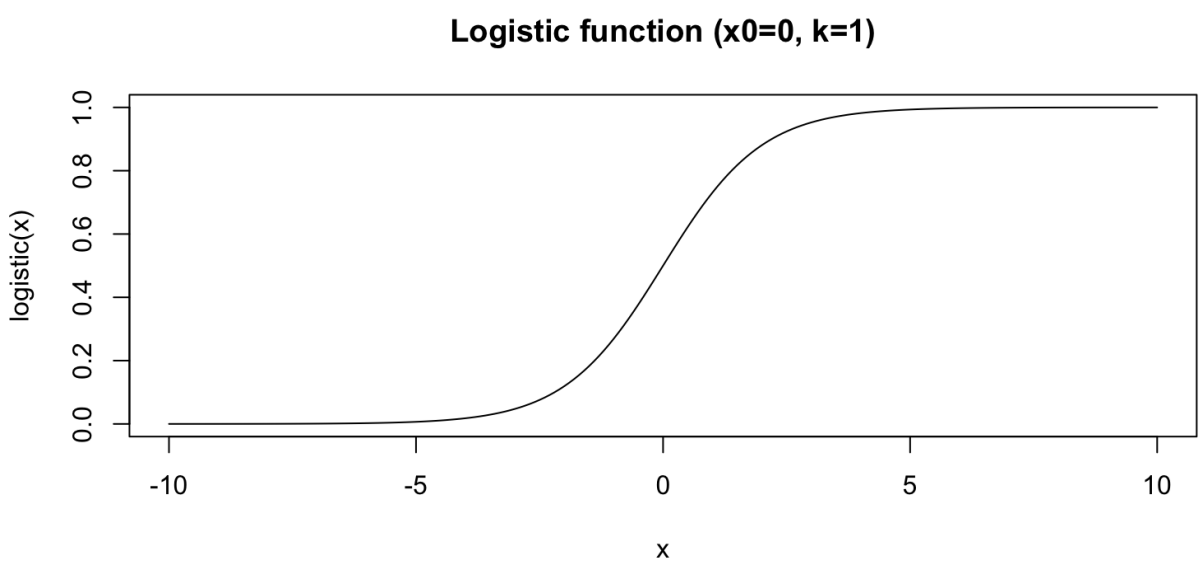
Properties of the Logit Function

- the Logit function is the inverse of the Logistic function
- the Logistic function is an S-shaped curve that maps continuous values ($-\infty$ to ∞) to the interval $(0, 1)$

▶ Code

$$f(x) = \frac{1}{1 + e^{-k(x-x_0)}}$$

- where x_0 is the sigmoid's midpoint and k is the steepness of the curve



Logistic Regression

- $log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \text{Age}$
- we can solve for p:

$$p = \frac{e^{(\beta_0 + \beta_1 \text{Age})}}{1 + e^{(\beta_0 + \beta_1 \text{Age})}}$$

- this is the **logistic regression model** for the probability of survival as a function of age

Logistic Regression

$$p = \frac{e^{(\beta_0 + \beta_1 \text{Age})}}{1 + e^{(\beta_0 + \beta_1 \text{Age})}}$$

- we can estimate the parameters β_0 and β_1 from the data
- then we can plug in any value of Age to get a predicted probability of survival
- we can use this model to predict the probability of survival for any age
- e.g. what's the probability of survival for a 30-year-old?

$$p = \frac{e^{\beta_0 + \beta_1 \text{Age}}}{1 + e^{\beta_0 + \beta_1 \text{Age}}} = \frac{e^{\beta_0 + \beta_1 (30)}}{1 + e^{\beta_0 + \beta_1 (30)}}$$

Logistic Regression in R

- we can fit a logistic regression model in R using the `glm()` function
- `glm()` is a general function for fitting generalized linear models
- for logistic regression, the family is `binomial` and the link function is `logit`

```r
mymod <- glm(Status ~ Age, data=donner, family=binomial(link="logit"))
summary(mymod)
```

```
Call:
glm(formula = Status ~ Age, family = binomial(link = "logit"),
    data = donner)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.06749    1.09914   1.881   0.0600 .
Age         -0.07339    0.03510  -2.091   0.0366 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 59.028  on 42  degrees of freedom
Residual deviance: 53.134  on 41  degrees of freedom
AIC: 57.134

Number of Fisher Scoring iterations: 4
```

Logistic Regression in R

- we can fit a logistic regression model in R using the `glm()` function
- `glm()` is a general function for fitting generalized linear models
- for logistic regression, the family is `binomial` and the link function is `logit`
- `logit` is the default so we can leave it out if we want:

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

```
Call:
glm(formula = Status ~ Age, family = binomial, data = donner)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.06749    1.09914   1.881   0.0600 .
Age         -0.07339    0.03510  -2.091   0.0366 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 59.028  on 42  degrees of freedom
Residual deviance: 53.134  on 41  degrees of freedom
AIC: 57.134

Number of Fisher Scoring iterations: 4
```

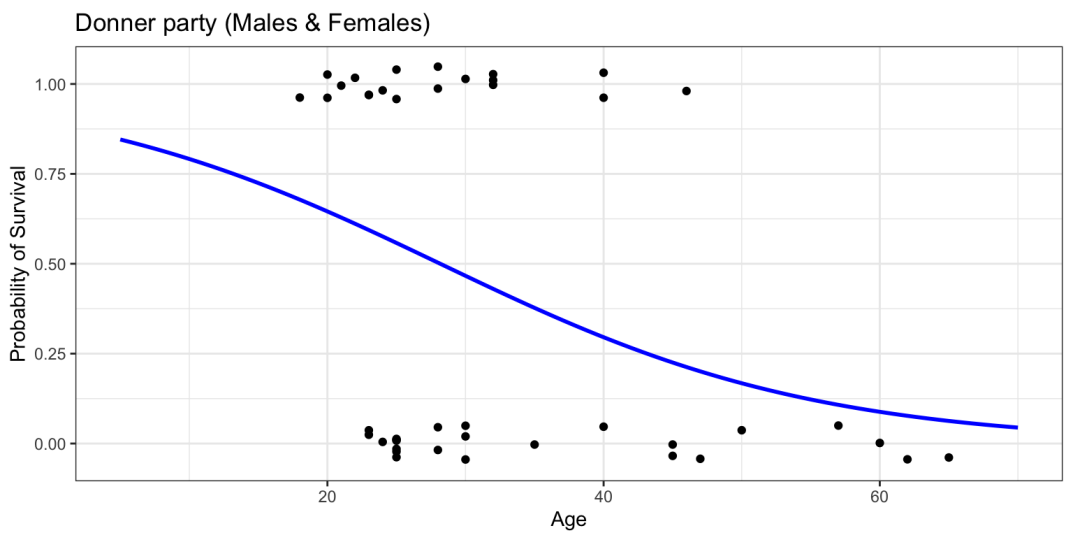
Logistic Regression in R

```r
mymod <- glm(Status ~ Age, data=donner,
summary(mymod)
```

```
Call:
glm(formula = Status ~ Age, family = binomial, data
= donner)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.06749    1.09914   1.881   0.0600 .
Age         -0.07339    0.03510  -2.091   0.0366 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1

(Dispersion parameter for binomial family taken to
be 1)

    Null deviance: 59.028  on 42  degrees of freedom
Residual deviance: 53.134  on 41  degrees of freedom
AIC: 57.134

Number of Fisher Scoring iterations: 4
```

- the model is: $log\left(\frac{p}{1-p}\right) = 2.07 - 0.07(\text{Age})$
- or, equivalently: $p = \frac{e^{2.07-0.07(\text{Age})}}{1+e^{2.07-0.07(\text{Age})}}$

▶ Code



Prediction with Logistic Regression

```
mymod <- glm(Status ~ Age, data=donner,
summary(mymod)
```

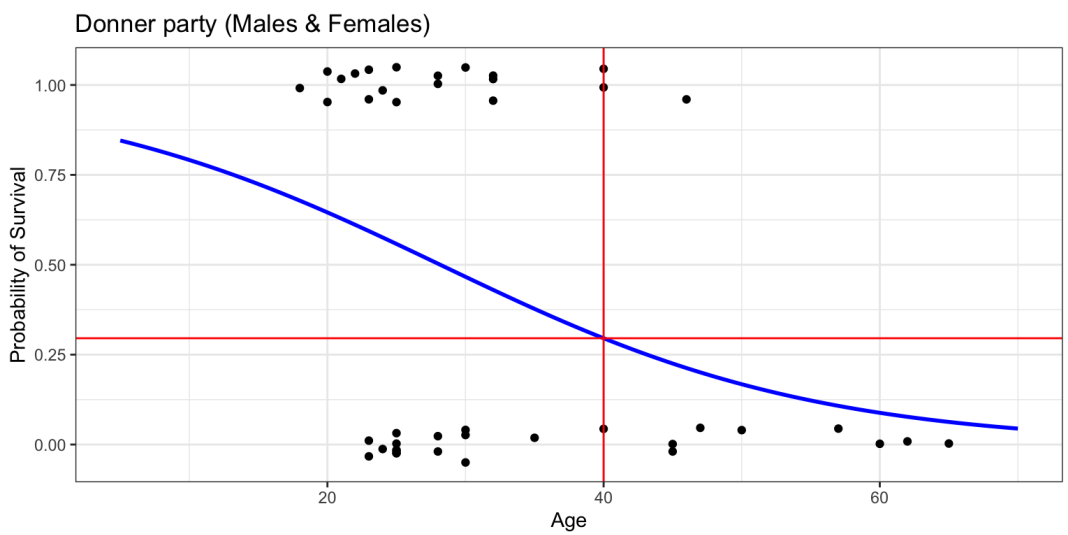
```
Call:
glm(formula = Status ~ Age, family = binomial, data
= donner)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.06749    1.09914   1.881   0.0600 .
Age         -0.07339    0.03510  -2.091   0.0366 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1

(Dispersion parameter for binomial family taken to
be 1)

    Null deviance: 59.028  on 42  degrees of freedom
Residual deviance: 53.134  on 41  degrees of freedom
AIC: 57.134

Number of Fisher Scoring iterations: 4
```

- what is the probability of survival for a 40-year-old?

```
pred_data <- tibble(Age=40)
prob_40 <- predict(mymod, newdata=pred_data, type="re
prob_40
```

```
        1
0.2956287
```

▶ Code



Interpretation of Logistic Regression Coefficients

- the regression coefficient of a variable corresponds to the change in log odds and its exponentiated form corresponds to the odds ratio
- the coefficients of a logistic regression model are in units of **log odds**
- the interpretation for a coefficient (a slope) is the *change in log odds* **per unit change** in the predictor variable

$$log\left(\frac{p}{1-p}\right) = 2.07 - 0.07(\text{Age})$$

$$p = \frac{e^{2.07-0.07(\text{Age})}}{1+e^{2.07-0.07(\text{Age})}}$$

```r
mymod <- glm(Status ~ Age, data=donner,
summary(mymod)
```

```
Call:
glm(formula = Status ~ Age, family = binomial, data
= donner)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.06749    1.09914   1.881   0.0600 .
Age         -0.07339    0.03510  -2.091   0.0366 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1

(Dispersion parameter for binomial family taken to
be 1)

    Null deviance: 59.028  on 42  degrees of freedom
Residual deviance: 53.134  on 41  degrees of freedom
AIC: 57.134

Number of Fisher Scoring iterations: 4
```

Interpretation of Logistic Regression Coefficients

- e.g. the coefficient for Age is -0.07
- this means that for each additional year of age, the log odds of survival decreases by 0.07
- or, equivalently, the **odds ratio** of survival for $(age + 1)$ vs (age) is 0.93
- +1 unit of Age = 0.93 odds ratio of survival
  - odds ratio > 1 : increase in odds of survival
  - odds ratio < 1 : decrease in odds of survival
- *not* the same as 0.93 decrease in probability of survival (next slide)

$$
\begin{aligned}
log\left(\frac{p}{1-p}\right) &= 2.07 - 0.07(\text{Age}) \\
p &= \frac{e^{2.07-0.07(\text{Age})}}{1+e^{2.07-0.07(\text{Age})}}
\end{aligned}
$$

```
mymod <- glm(Status ~ Age, data=donner,
summary(mymod)
```

```
Call:
glm(formula = Status ~ Age, family = binomial, data
= donner)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.06749    1.09914   1.881   0.0600 .
Age         -0.07339    0.03510  -2.091   0.0366 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1

(Dispersion parameter for binomial family taken to
be 1)

    Null deviance: 59.028  on 42  degrees of freedom
Residual deviance: 53.134  on 41  degrees of freedom
AIC: 57.134

Number of Fisher Scoring iterations: 4
```

Odds vs Probability of outcome

- the **odds** of an outcome is the ratio of the probability of the outcome to the probability of the opposite outcome
- odds : $\frac{p}{1-p}$
  - e.g. if the probability of survival is 0.8, then the odds of survival are 0.8/0.2 = 4
- the **log odds** of an outcome is the natural logarithm of the odds
  - e.g. if the probability of survival is 0.8, then the log odds of survival are $log(4) = 1.39$
- the **odds ratio** is the ratio of odds between two situations
- the coefficients (slopes) of a logistic regression model are the change in the **log odds** per unit change in a continuous variable (e.g. Age) or a change in **log odds** across categories (e.g. Sex)

Interpretation of Logistic Regression Coefficients

- the coefficient for Age is -0.07
- this means that for each additional year of age, the log odds of survival decreases by 0.07
- or, equivalently, the odds ratio of survival is $e^{-0.07} = 0.93$ for an additional year of age
- +1 unit of Age = 0.93 **odds ratio** of survival
  - An odds ratio > 1 implies a positive association between the predictor and the outcome
  - An odds ratio < 1 implies a negative association between the predictor and the outcome

$$
\begin{aligned}
log\left(\frac{p}{1-p}\right) &= 2.07 - 0.07(\text{Age}) \\
p &= \frac{e^{2.07-0.07(\text{Age})}}{1+e^{2.07-0.07(\text{Age})}}
\end{aligned}
$$

```
mymod <- glm(Status ~ Age, data=donner,
summary(mymod)
```

```
Call:
glm(formula = Status ~ Age, family = binomial, data
= donner)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.06749    1.09914   1.881   0.0600 .
Age         -0.07339    0.03510  -2.091   0.0366 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1

(Dispersion parameter for binomial family taken to
be 1)

    Null deviance: 59.028  on 42  degrees of freedom
Residual deviance: 53.134  on 41  degrees of freedom
AIC: 57.134

Number of Fisher Scoring iterations: 4
```

Interpretation of Logistic Regression Coefficients

for example:

- Prob(Alive) for 40-year-old: $p = 0.296$
- Prob(Alive) for a 41-year-old: $p = 0.281$
- odds of survival for 40-year-old: $\text{odds} = \frac{0.296}{1-0.296} = 0.42$
- odds of survival for 41-year-old: $\text{odds} = \frac{0.281}{1-0.281} = 0.39$
- odds ratio of 41-yr old vs 40-yr old:
- $\text{odds ratio} = \frac{0.39}{0.42} = 0.93$
- $\log(0.93) = -0.07$

$$
\begin{aligned}
log\left(\frac{p}{1-p}\right) &= 2.07 - 0.07(\text{Age}) \\
p &= \frac{e^{2.07-0.07(\text{Age})}}{1+e^{2.07-0.07(\text{Age})}}
\end{aligned}
$$

```r
mymod <- glm(Status ~ Age, data=donner,
summary(mymod)
```

```
Call:
glm(formula = Status ~ Age, family = binomial, data
= donner)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.06749    1.09914   1.881   0.0600 .
Age         -0.07339    0.03510  -2.091   0.0366 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1

(Dispersion parameter for binomial family taken to
be 1)

    Null deviance: 59.028  on 42  degrees of freedom
Residual deviance: 53.134  on 41  degrees of freedom
AIC: 57.134

Number of Fisher Scoring iterations: 4
```

Logistic Regression with 2 Predictors

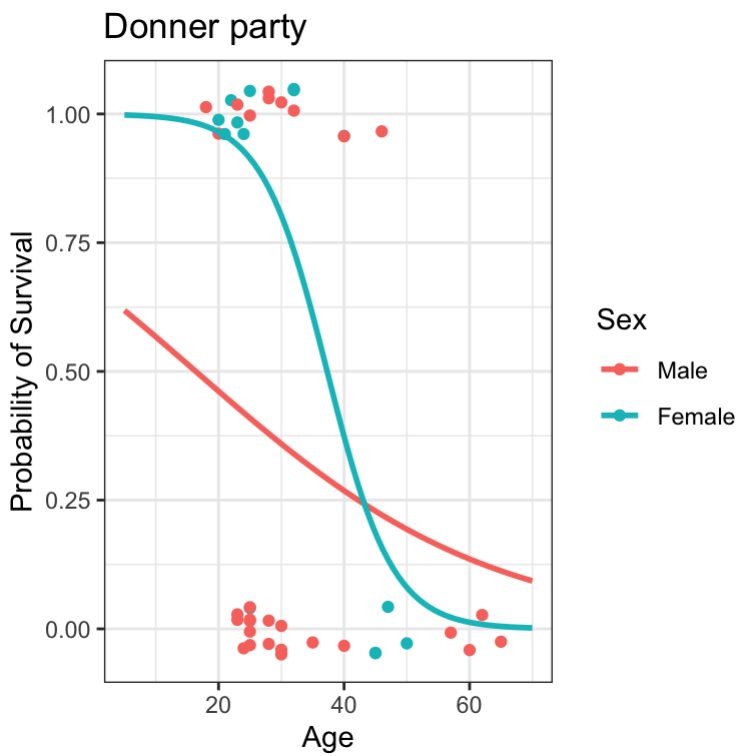
```
1 mymod <- glm(Status ~ Age + Sex, data=do
2 coef(mymod)
```

```
(Intercept)         Age   SexFemale
 2.00466996 -0.08762745  1.50667376
```

▶ Code

$$log\left(\frac{p}{1-p}\right) = 2.0 - 0.09(\text{Age}) + 1.5(\text{SexFen}$$

$$p = \frac{e^{2.0-0.09(\text{Age})+1.5(\text{SexFemale})}}{1+e^{2.0-0.09(\text{Age})+1.5(\text{SexFemale})}}$$



Interpretation of Logistic Regression Coefficients

$$
\begin{aligned}
log\left(\frac{p}{1-p}\right) &= 2.0 - 0.09(\text{Age}) + 1.5(\text{SexFemale}) \\
p &= \frac{e^{2.0-0.09(\text{Age})+1.5(\text{SexFemale})}}{1+e^{2.0-0.09(\text{Age})+1.5(\text{SexFemale})}}
\end{aligned}
$$

- the coefficient for **SexFemale** is 1.5
- this means R coded Males as 0 and Females as +1
- the model for Males is:

$$log\left(\frac{p}{1-p}\right) = 2.0 - 0.09(\text{Age}) + 1.5(0)$$

- the model for Females is:

$$log\left(\frac{p}{1-p}\right) = 2.0 - 0.09(\text{Age}) + 1.5(1)$$

```
1 mymod <- glm(Status ~ Age + Sex, data=do
2 summary(mymod)
```

```
Call:
glm(formula = Status ~ Age + Sex, family = binomial,
data = donner)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.00467    1.21857   1.645   0.1000 .
Age         -0.08763    0.04084  -2.146   0.0319 *
SexFemale    1.50667    0.78497   1.919   0.0549 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1

(Dispersion parameter for binomial family taken to
be 1)

    Null deviance: 59.028  on 42  degrees of freedom
Residual deviance: 48.992  on 40  degrees of freedom
AIC: 54.992

Number of Fisher Scoring iterations: 5
```

Interpretation of Logistic Regression Coefficients

- the coefficient for **Age** is -0.09
- this means that for each additional year of age, the log odds of survival decreases by 0.09
- or, equivalently, the odds ratio of survival is $e^{-0.09} = 0.91$ for an extra year of age
- +1 unit of Age = 0.91 odds ratio of survival

```
1 mymod <- glm(Status ~ Age + Sex, data=do
2 summary(mymod)
```

```
Call:
glm(formula = Status ~ Age + Sex, family = binomial,
data = donner)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.00467    1.21857   1.645   0.1000 .
Age         -0.08763    0.04084  -2.146   0.0319 *
SexFemale    1.50667    0.78497   1.919   0.0549 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1

(Dispersion parameter for binomial family taken to
be 1)

    Null deviance: 59.028  on 42  degrees of freedom
Residual deviance: 48.992  on 40  degrees of freedom
AIC: 54.992

Number of Fisher Scoring iterations: 5
```

Interpretation of Logistic Regression Coefficients

- the coefficient for **SexFemale** is 1.5
- this means that for Females the log odds of survival increases by 1.5 over Males
- or, equivalently, the odds ratio of survival for females vs males is $e^{1.5} = 4.5$
- odds = $\frac{p}{1-p}$
- log odds = $\log(\frac{p}{1-p})$
- $\log(\frac{f}{1-f}) = \log(\frac{m}{1-m}) + 1.5$
- $\log(\frac{f}{1-f}) - \log(\frac{m}{1-m}) = 1.5$
- $\log(\frac{f}{1-f} / \frac{m}{1-m}) = 1.5$
- $\frac{f}{1-f} / \frac{m}{1-m} = e^{1.5} = 4.5$
- $\frac{\text{odds female}}{\text{odds male}} = e^{1.5} = 4.5$

```
1 mymod <- glm(Status ~ Age + Sex, data=do
2 summary(mymod)
```

```
Call:
glm(formula = Status ~ Age + Sex, family = binomial,
data = donner)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.00467    1.21857   1.645   0.1000 .
Age         -0.08763    0.04084  -2.146   0.0319 *
SexFemale    1.50667    0.78497   1.919   0.0549 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1

(Dispersion parameter for binomial family taken to
be 1)

    Null deviance: 59.028  on 42  degrees of freedom
Residual deviance: 48.992  on 40  degrees of freedom
AIC: 54.992

Number of Fisher Scoring iterations: 5
```

Model Fit and Model Selection

- does adding Sex to the model improve the fit? Use **AIC** to compare models

```r
mod_reduced <- glm(Status ~ Age, data=donner, 
summary(mod_reduced)
```

```
Call:
glm(formula = Status ~ Age, family = binomial, data =
donner)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.06749    1.09914   1.881   0.0600 .
Age         -0.07339    0.03510  -2.091   0.0366 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 59.028  on 42  degrees of freedom
Residual deviance: 53.134  on 41  degrees of freedom
AIC: 57.134

Number of Fisher Scoring iterations: 4
```

```r
mod_full <- glm(Status ~ Age + Sex, data=donne
summary(mod_full)
```

```
Call:
glm(formula = Status ~ Age + Sex, family = binomial, data
= donner)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.00467    1.21857   1.645   0.1000 .
Age         -0.08763    0.04084  -2.146   0.0319 *
SexFemale    1.50667    0.78497   1.919   0.0549 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 59.028  on 42  degrees of freedom
Residual deviance: 48.992  on 40  degrees of freedom
AIC: 54.992

Number of Fisher Scoring iterations: 5
```

- AIC goes down when Sex is added to the model, so keep it in the model
- as in multiple regression we can use the `step()` function to do stepwise model selection

Assumptions of Logistic Regression

- logistic regression does not require linearity (obviously)
  - it does require that the predictors are linearly related to the log odds of the outcome
  - **we will test for this**
- nor does it require normality of residuals
  - but it does require that the residuals are independent
  - there is no test of this, it's conceptual
- nor does it require homoscedasticity
- it does require (obviously) that the dependent variable is binary
  - we ought to know this by the nature of the variable
- like multiple regression, it does require that the predictors are independent of each other
  - collinearity is a problem for logistic regression as it is for multiple regression
  - **we can test for this**

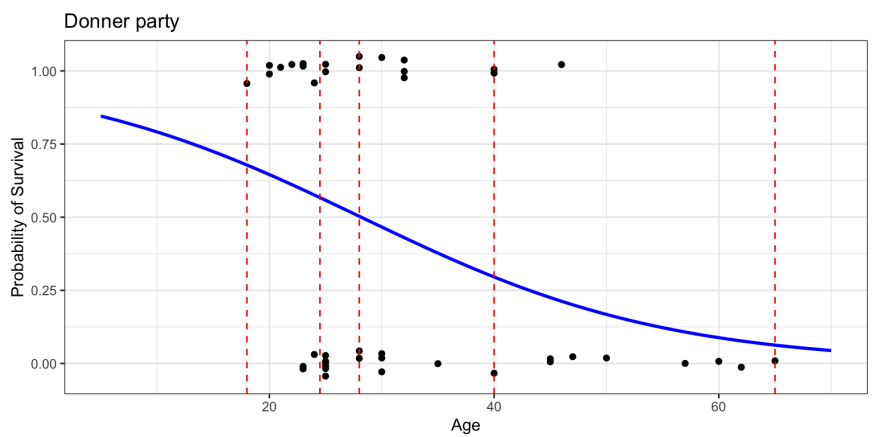
Assumptions of Logistic Regression

- logistic regression typically requires a relatively large sample size to provide a stable model fit
  - a rough minimum: 10 cases considering the least frequent outcome, for each predictor variable
  - e.g. if we have 3 predictor variables, and the probability of the least frequent outcome is 0.3, the minimum sample size should be around (10 x 3)/0.3 = 100
- as with multiple regression, logistic regression is sensitive to outliers
  - there are ways to test this

Check for log-odds linearity

- are the predictors linearly related to the log odds of the outcome?
- **`cut()`** the predictor `Age` into bins using `quantile()`
- within each bin, compute log odds of `Survived`

▶ Code




```r
linearity_data <- donner %>%
    mutate(prob = ifelse(Status == "Survived", 1, 0)
    mutate(Age_bin = cut(Age,
                         breaks=quantile(Age),
                         include.lowest=TRUE)) %>%
    group_by(Age_bin) %>%
    summarize(bin_n        = n(),
              bin_Age      = mean(Age),
              bin_prob     = mean(prob),
              bin_log_odds = log(bin_prob/(1-bin_prob
```

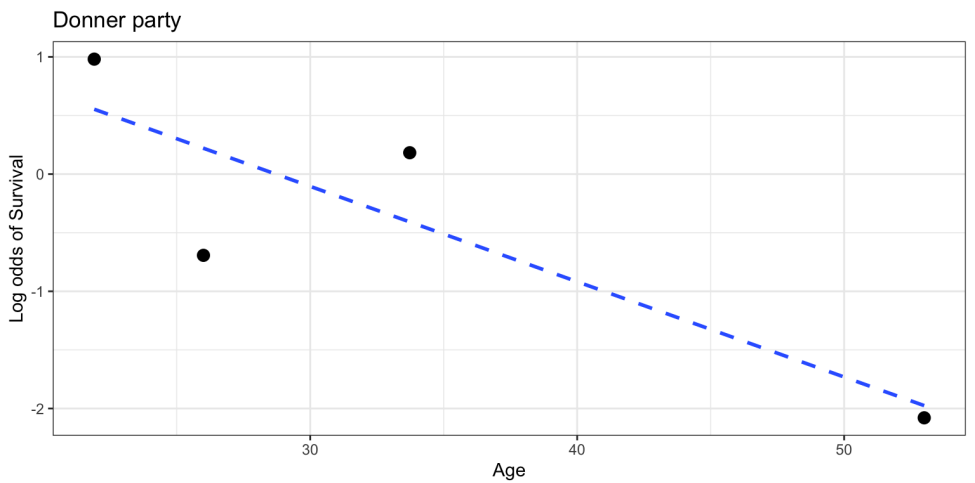
```r
linearity_data
```

```
# A tibble: 4 × 5
  Age_bin   bin_n bin_Age bin_prob bin_log_odds
  <fct>     <int>   <dbl>    <dbl>        <dbl>
1 [18,24.5]    11    21.9    0.727        0.981
2 (24.5,28]    12    26      0.333       -0.693
3 (28,40]      11    33.7    0.545        0.182
4 (40,65]       9    53      0.111       -2.08
```

Check for log-odds linearity

- then plot age vs log odds of survival within each quantile bin
- is it linear? sort of. maybe not.
  - (second bin should be higher)
- but we don't have much data in this toy dataset so 🤷
- you will see an example in the homework this week of a larger dataset

```
ggplot(data=linearity_data, aes(x=bin_Age, y=b
   geom_point(size=3) +
   geom_smooth(method="lm", se=FALSE, lty=2, s
   labs(title="Donner party",
        x="Age",
        y="Log odds of Survival") +
   theme_bw()
```



Box-Tidwell Test

- are the predictors linearly related to the log odds of the outcome?
- statistical test: Box-Tidwell test
- we test for the significance of the **interaction between the predictor and the log odds of the outcome**
- if there is an *interaction* between X and Y it means that
  - **the relationship between X and Y is different at different values of X**
  - i.e. the slope of the line relating X and Y is different at different values of X
  - i.e. it's not a constant slope, i.e. not a straight line, i.e. not linear, i.e. nonlinear
- if the Box-Tidwell test is significant, then the relationship between the predictor and the log odds of the outcome is nonlinear

Box-Tidwell Test

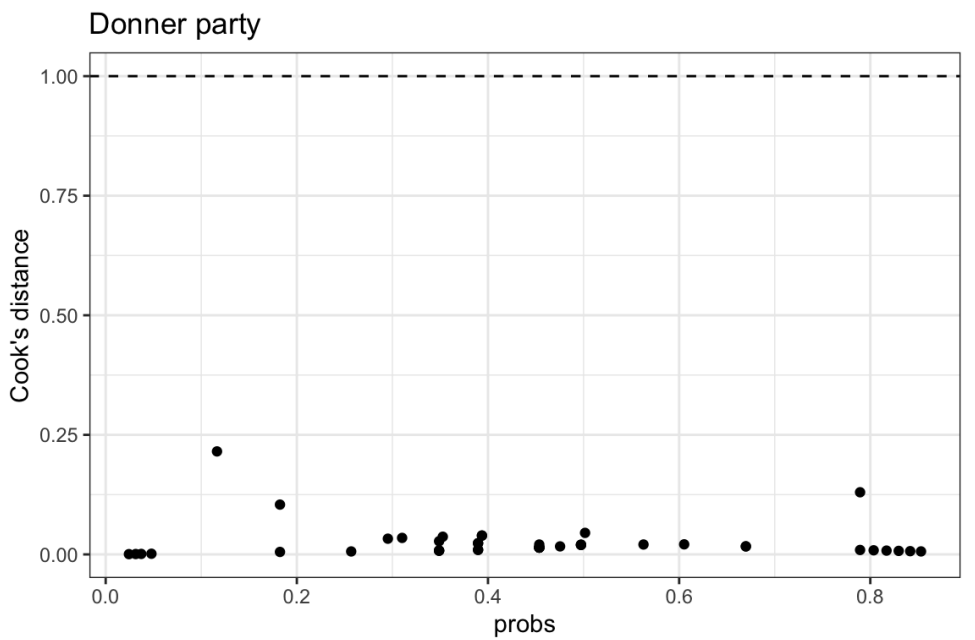
- we test for the significance of the **interaction between the predictor and the log odds of the outcome**
- in an R model formula, the notation `X:Y` means "the interaction between X and Y"
- so we simply add a term `Age:log(Age)` to the GLM model
- `Age:log(Age)` means "the interaction between Age and log(Age)"
- here: `p=0.7712` so we are `>.05` so we fail to reject the null hypothesis that there is no interaction between Age and log odds of survival
- the linearity assumption passes the Box-Tidwell test

```
Call:
glm(formula = Status ~ Age + Sex + (Age:log(Age)),
family = binomial,
    data = donner)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.60391    8.97625  -0.067   0.9464
Age           0.27300    1.23769   0.221   0.8254
SexFemale     1.52563    0.79610   1.916   0.0553 .
Age:log(Age) -0.07959    0.27374  -0.291   0.7712
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1

(Dispersion parameter for binomial family taken to
be 1)

    Null deviance: 59.028  on 42  degrees of freedom
Residual deviance: 48.902  on 39  degrees of freedom
AIC: 56.902

Number of Fisher Scoring iterations: 5
```

Outliers

- use Cook's distance
- Cook's distance is a measure of the influence of each observation on the model
- observations with high Cook's distance are outliers
- values > 1 are considered high
- all our values are < 1 so we're ok

```
donner <- donner %>%
    mutate(probs = predict(mod_full, type="resp
           cd = cooks.distance(mod_full))
ggplot(data=donner, aes(x=probs, y=cd)) +
   geom_point() +
   geom_hline(yintercept=1, linetype="dashed") +
   labs(title="Donner party", y="Cook's distanc
   theme_bw()
```



Collinearity

- as in multiple regression
- visualize using `ggpairs()` from the `GGally` package
- test using **variance inflation factor** (**VIF**)
- VIF ~ 4 or 5 is considered moderatly high and may be a problem
- VIF > 10 is considered high and is definitely a problem
- See Navarro 15.9.6 for more details
  - she discusses it in the context of multiple regression
  - the same principles apply to multiple continuous predictor variables in logistic regression