# Lecture 22 - Model Selection

Sta102 / BME102

April 20, 2016

Colin Rundel

# Model diagnostics

# Modeling children's test scores

Predicting cognitive test scores of three- and four-year-old children using characteristics of their mothers. Data are a subsample from the National Longitudinal Survey of Youth.

| | kid_score | mom_hs | mom_iq | mom_work | mom_age |
|---|---|---|---|---|---|
| 1 | 65 | yes | 121.12 | yes | 27 |
| ⋮ | | | | | |
| 5 | 115 | yes | 92.75 | yes | 27 |
| 6 | 98 | no | 107.90 | no | 18 |
| ⋮ | | | | | |
| 434 | 70 | yes | 91.25 | yes | 25 |

Gelman, Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. (2007)

Cambridge University Press.

# Model output

```
summary(lm(kid_score ~ mom_hs + mom_iq + mom_work + mom_age, data = cognitive))
```

```
##
## Call:
## lm(formula = kid_score ~ mom_hs + mom_iq + mom_work + mom_age,
##     data = cognitive)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.134 -12.624   2.293  11.250  50.206
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.82261    9.18765   2.266   0.0239 *
## mom_hs       5.56118    2.31345   2.404   0.0166 *
## mom_iq       0.56208    0.06077   9.249   <2e-16 ***
## mom_work     0.13373    0.76763   0.174   0.8618
## mom_age      0.21986    0.33231   0.662   0.5086
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.17 on 429 degrees of freedom
## Multiple R-squared:  0.215,Adjusted R-squared:  0.2077
## F-statistic: 29.38 on 4 and 429 DF,  p-value: < 2.2e-16
```

# Conditions for MLR Inference

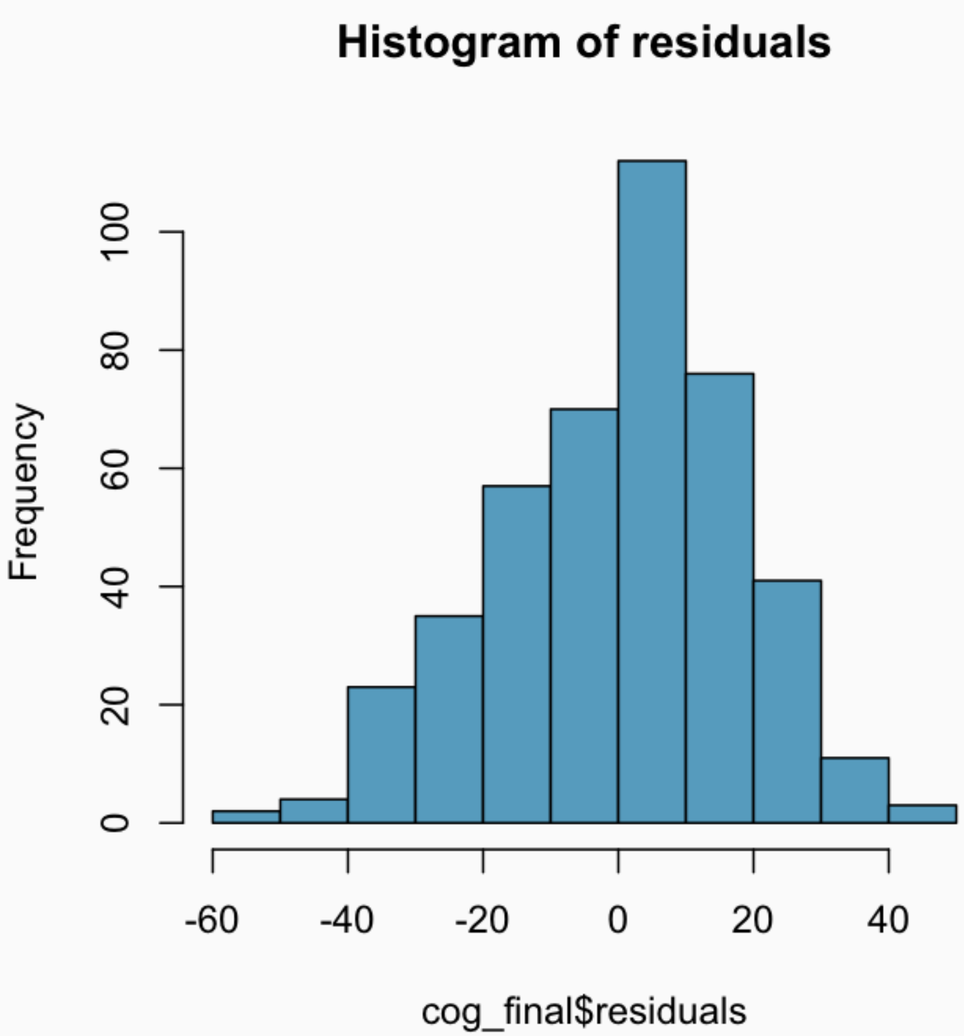
In order to conduct inference for multiple regression we require the following conditions:

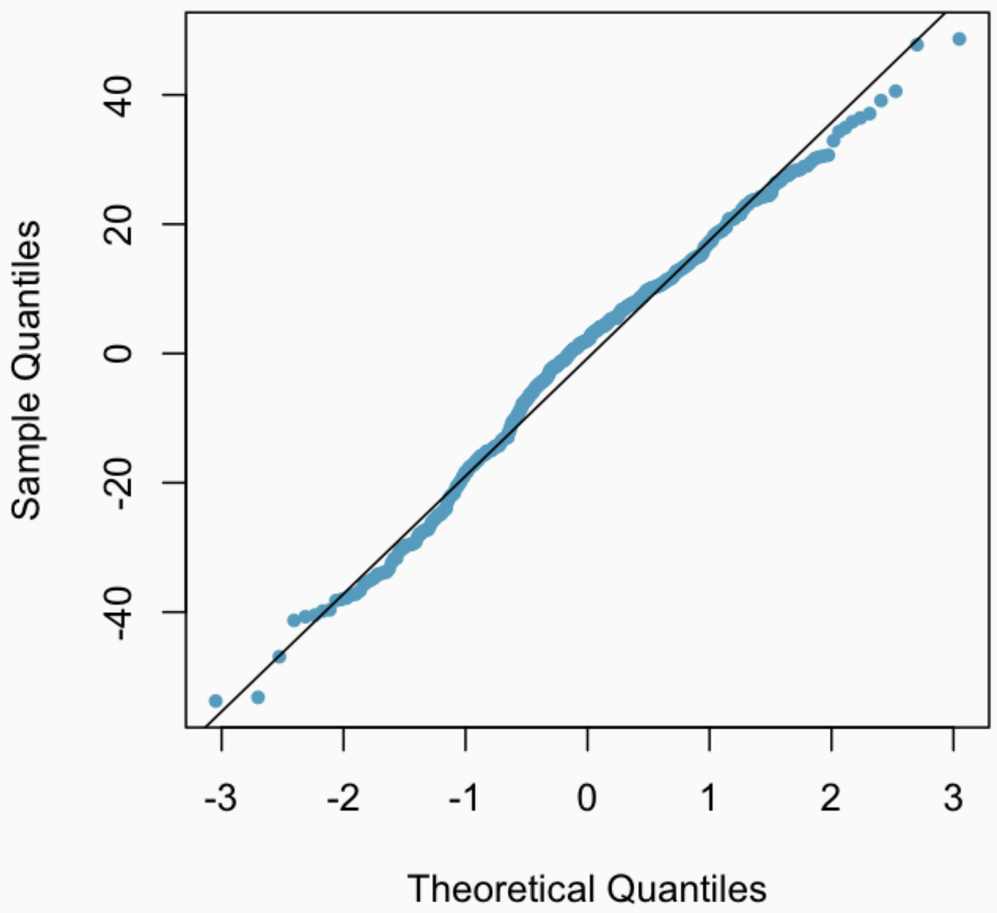
(1) Unstructured / nearly normal residuals

(2) Constant variability of residuals

(3) Independent residuals

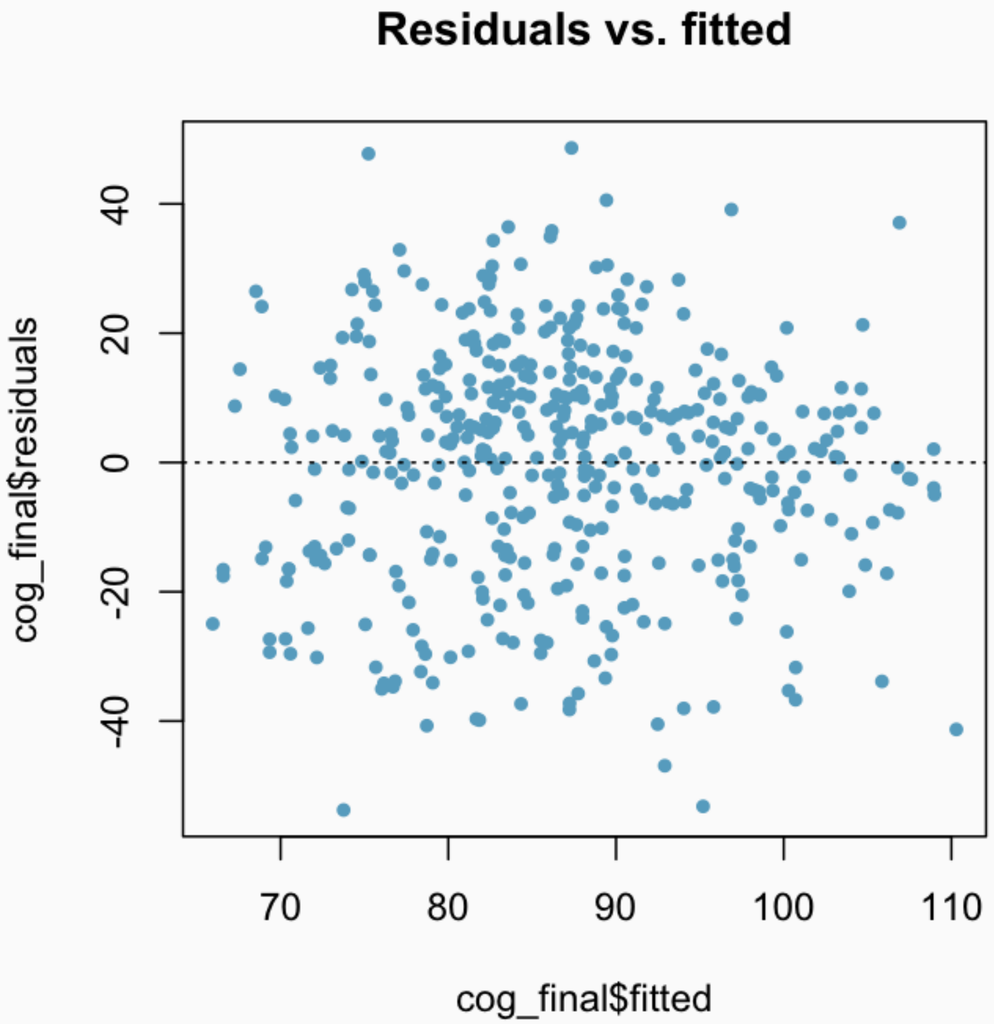
# Nearly normal residuals

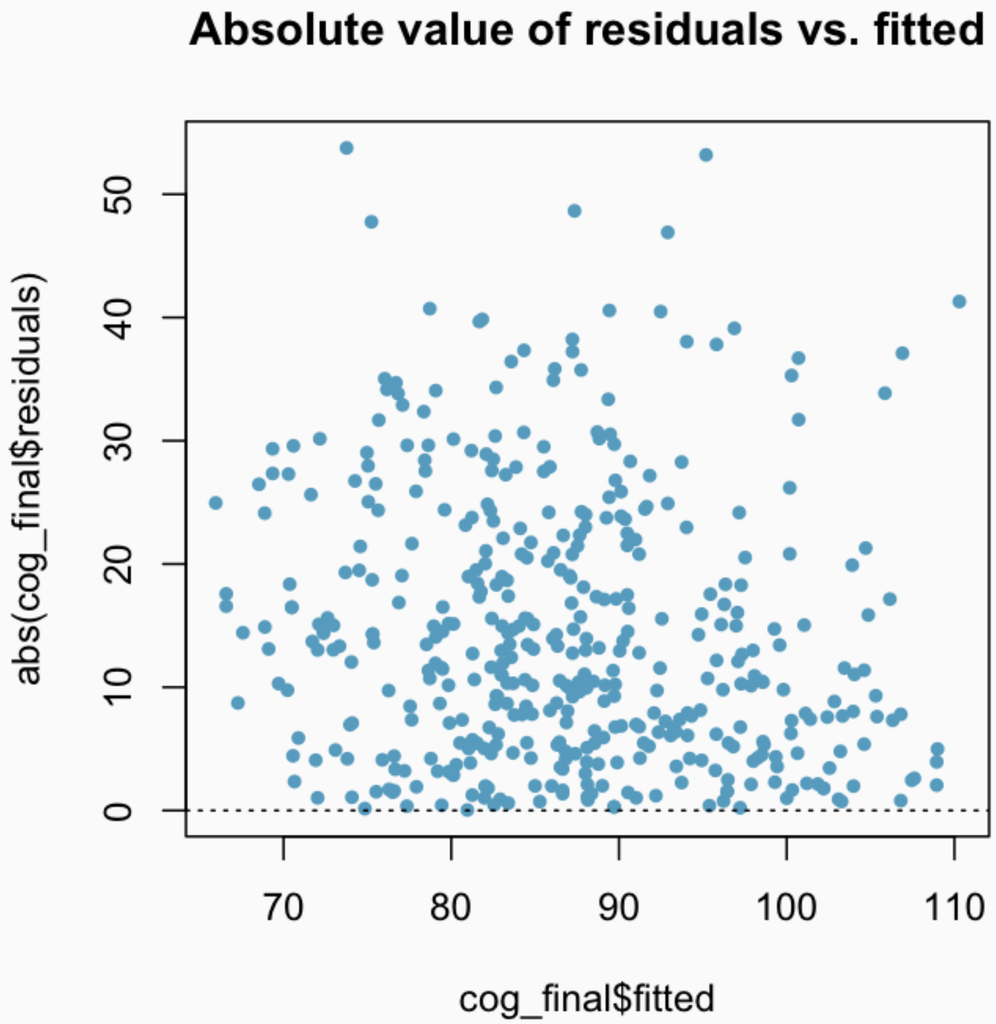


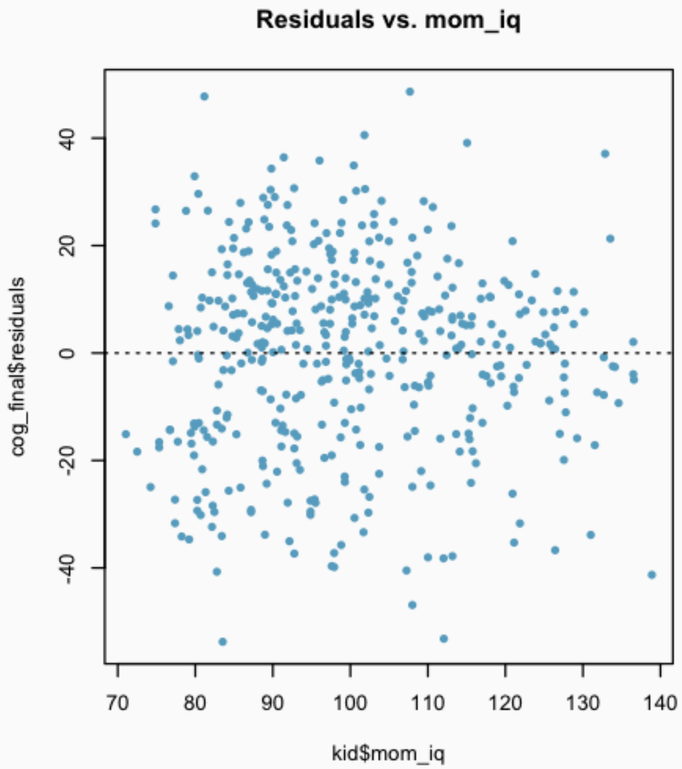
# Unstructured / Constant variability of residuals

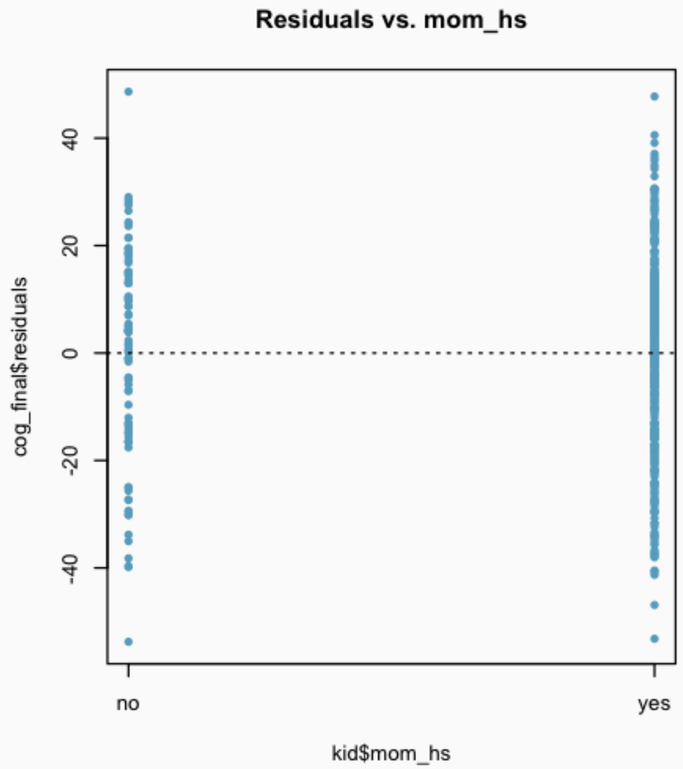
Why do we use the fitted (predicted) values in MLR?

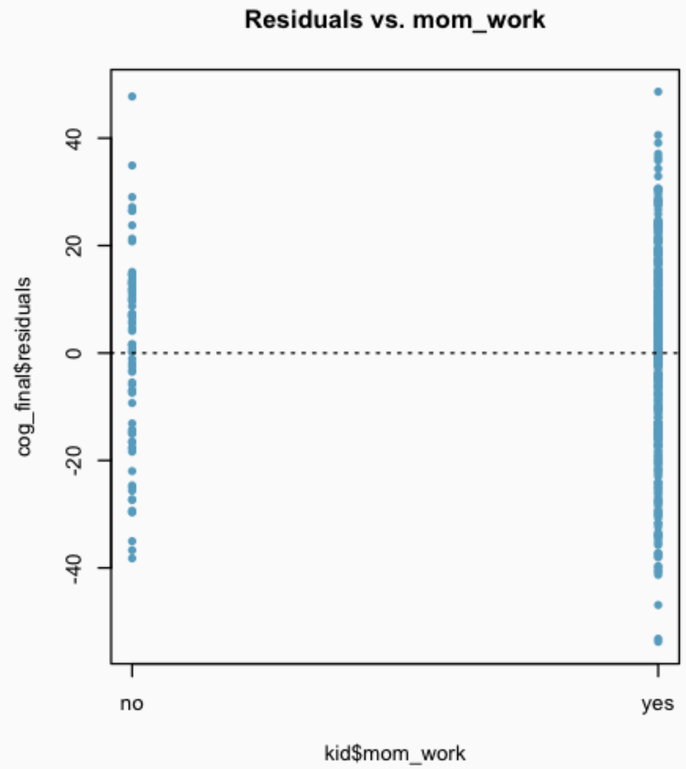


# Constant variability of residuals (cont.)

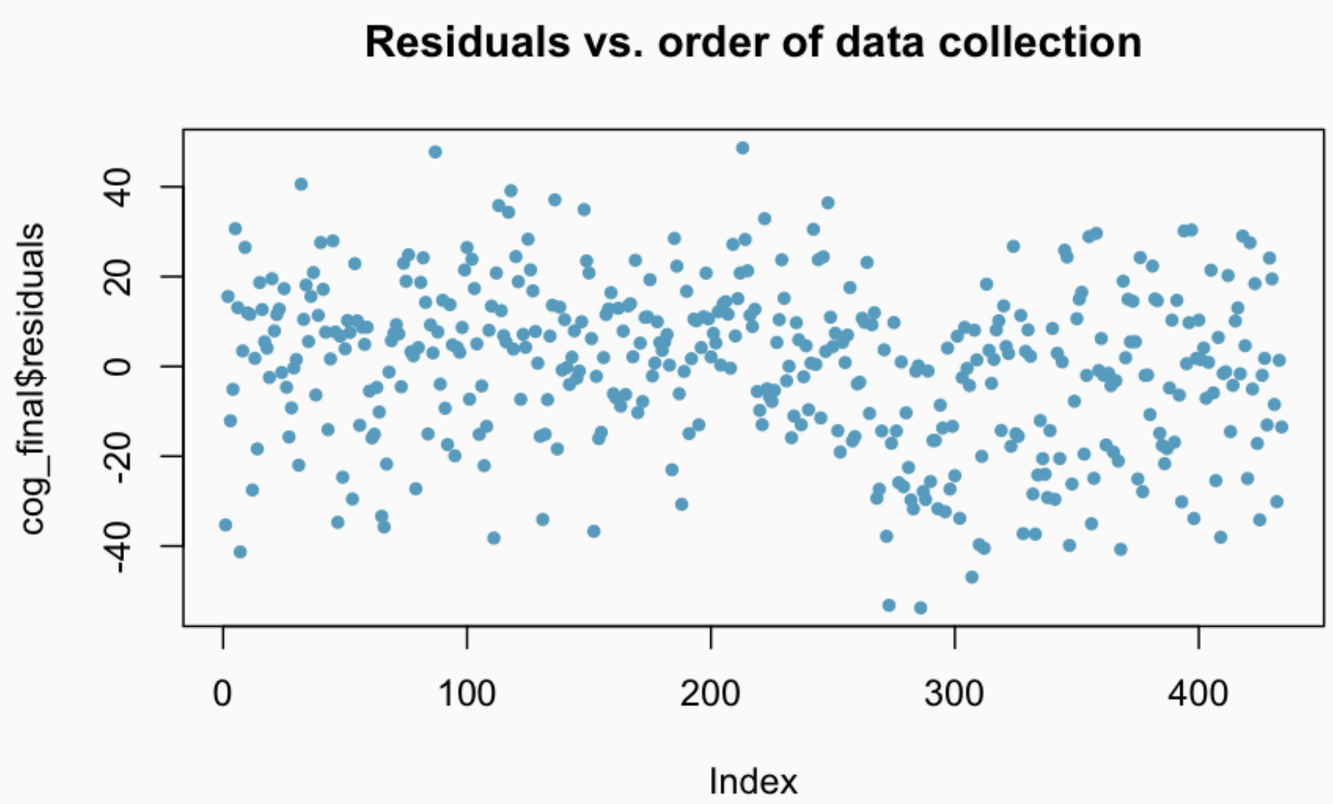


# Independent residuals

- If we suspect that order of data collection may influence the outcome (mostly in time series data):




- If not, think about how data are sampled.

# Inference for MLR

# Model output

```
summary(lm(kid_score ~ mom_hs + mom_iq + mom_work + mom_age, data = cognitive))
```

```
##
## Call:
## lm(formula = kid_score ~ mom_hs + mom_iq + mom_work + mom_age,
##     data = cognitive)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.134 -12.624   2.293  11.250  50.206
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.82261    9.18765   2.266   0.0239 *
## mom_hs       5.56118    2.31345   2.404   0.0166 *
## mom_iq       0.56208    0.06077   9.249   <2e-16 ***
## mom_work     0.13373    0.76763   0.174   0.8618
## mom_age      0.21986    0.33231   0.662   0.5086
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.17 on 429 degrees of freedom
## Multiple R-squared:  0.215,Adjusted R-squared:  0.2077
## F-statistic: 29.38 on 4 and 429 DF,  p-value: < 2.2e-16
```

df = ?

# Inference for the model as a whole

Is the model as a whole significant?

# Inference for the model as a whole

Is the model as a whole significant?

$$H_0 : \beta_1 = \cdots = \beta_k = 0$$

$$H_A : \text{At least one of the } \beta_i \neq 0$$

# Inference for the model as a whole

Is the model as a whole significant?

$$H_0 : \beta_1 = \cdots = \beta_k = 0$$

$$H_A : \text{At least one of the } \beta_i \neq 0$$

```
F-statistic: 29.38 on 4 and 429 DF,  p-value: < 2.2e-16
```

# Inference for the model as a whole

Is the model as a whole significant?

$$H_0 : \beta_1 = \cdots = \beta_k = 0$$

$$H_A : \text{At least one of the } \beta_i \neq 0$$

```
F-statistic: 29.38 on 4 and 429 DF,  p-value: < 2.2e-16
```

Since p-value $< 0.05$, the model as a whole is significant.

# Inference for the model as a whole

Is the model as a whole significant?

$$H_0 : \beta_1 = \cdots = \beta_k = 0$$

$$H_A : \text{At least one of the } \beta_i \neq 0$$

```
F-statistic: 29.38 on 4 and 429 DF,  p-value: < 2.2e-16
```

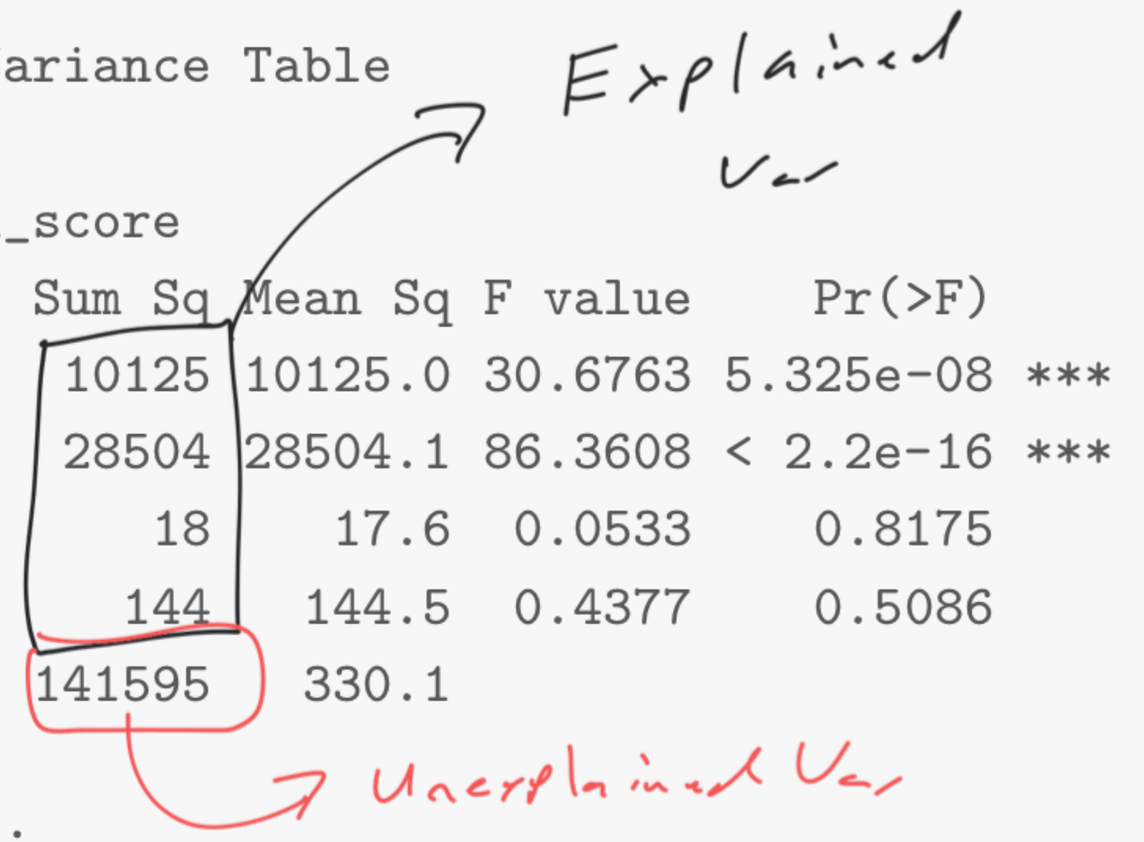
Since p-value $< 0.05$, the model as a whole is significant.

- The F test yielding a significant result doesn't mean the model fits the data well, it just means at least one of the $\beta$s is non-zero. i.e. the combination of these variables overall yields a model that is better than the intercept only model.

# ANOVA Table

```r
anova(lm(kid_score~.,data=cognitive))
```

```
## Analysis of Variance Table
##
## Response: kid_score
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## mom_hs      1  10125  10125.0 30.6763 5.325e-08 ***
## mom_iq      1  28504  28504.1 86.3608 < 2.2e-16 ***
## mom_work    1     18     17.6  0.0533    0.8175
## mom_age     1    144    144.5  0.4377    0.5086
## Residuals 429 141595    330.1
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



# ANOVA Table

```
anova(lm(kid_score~.,data=cognitive))
```

```
## Analysis of Variance Table
##
## Response: kid_score
##             Df Sum Sq Mean Sq F value    Pr(>F)
## mom_hs       1  10125 10125.0 30.6763 5.325e-08 ***
## mom_iq       1  28504 28504.1 86.3608 < 2.2e-16 ***
## mom_work     1     18    17.6  0.0533    0.8175
## mom_age      1    144   144.5  0.4377    0.5086
## Residuals  429 141595   330.1
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$MS_{Reg} = (18 + 144 + 10125 + 28504)/4 = 9697.75$$

$$F_{Reg} = 9697.75/330.1 = 29.38$$

# ANOVA Table

```
anova(lm(kid_score~.,data=cognitive))
```

```
## Analysis of Variance Table
##
## Response: kid_score
##            Df Sum Sq Mean Sq F value    Pr(>F)
## mom_hs      1  10125 10125.0 30.6763 5.325e-08 ***
## mom_iq      1  28504 28504.1 86.3608 < 2.2e-16 ***
## mom_work    1     18    17.6  0.0533    0.8175
## mom_age     1    144   144.5  0.4377    0.5086
## Residuals 429 141595   330.1
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$MS_{Reg} = (18 + 144 + 10125 + 28504)/4 = 9697.75$$

$$F_{Reg} = 9697.75/330.1 = 29.38$$

```
F-statistic: 29.38 on 4 and 429 DF,  p-value: < 2.2e-16
```

# Inference for the slope(s)

Is whether or not a mother graduated from high school a significant predictor of kid's cognitive test score, given all other variables in the model?

# Inference for the slope(s)

Is whether or not a mother graduated from high school a significant predictor of kid's cognitive test score, given all other variables in the model?

$H_0:$ $\beta_1 = 0$, when all other variables are included in the model

$H_A:$ $\beta_1 \neq 0$, when all other variables are included in the model

# Inference for the slope(s)

Is whether or not a mother graduated from high school a significant predictor of kid's cognitive test score, given all other variables in the model?

$H_0: \ \beta_1 = 0$, when all other variables are included in the model

$H_A: \ \beta_1 \neq 0$, when all other variables are included in the model

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.59241    9.21906   2.125   0.0341
mom_hsyes    5.09482    2.31450   2.201   0.0282
mom_iq       0.56147    0.06064   9.259   <2e-16
mom_workyes  2.53718    2.35067   1.079   0.2810
mom_age      0.21802    0.33074   0.659   0.5101

Residual standard error: 18.14 on 429 degrees of freedom
```

# Inference for the slope(s)

Is whether or not a mother graduated from high school a significant predictor of kid's cognitive test score, given all other variables in the model?

$H_0:$ $\beta_1 = 0$, when all other variables are included in the model

$H_A:$ $\beta_1 \neq 0$, when all other variables are included in the model

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.59241    9.21906   2.125   0.0341
mom_hsyes    5.09482    2.31450   2.201   0.0282
mom_iq       0.56147    0.06064   9.259   <2e-16
mom_workyes  2.53718    2.35067   1.079   0.2810
mom_age      0.21802    0.33074   0.659   0.5101

Residual standard error: 18.14 on 429 degrees of freedom
```

$$T = 2.201, \ df = n - k - 1 = 434 - 4 - 1 = 429, \ \text{p-value} = 0.0282$$

# Inference for the slope(s)

Is whether or not a mother graduated from high school a significant predictor of kid's cognitive test score, given all other variables in the model?

$H_0:$ $\beta_1 = 0$, when all other variables are included in the model

$H_A:$ $\beta_1 \neq 0$, when all other variables are included in the model

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.59241    9.21906   2.125   0.0341
mom_hsyes    5.09482    2.31450   2.201   0.0282
mom_iq       0.56147    0.06064   9.259   <2e-16
mom_workyes  2.53718    2.35067   1.079   0.2810
mom_age      0.21802    0.33074   0.659   0.5101

Residual standard error: 18.14 on 429 degrees of freedom
```

$$T = 2.201, \ df = n - k - 1 = 434 - 4 - 1 = 429, \ \text{p-value} = 0.0282$$

Since p-value $< 0.05$, whether or not mom went to high school is a significant predictor of kid's test score, given all other variables in the model.

# Interpreting the slope

What is the correct interpretation of the slope for `mom_work`?

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 19.59 | 9.22 | 2.13 | 0.03 |
| mom_hs:yes | 5.09 | 2.31 | 2.20 | 0.03 |
| mom_iq | 0.56 | 0.06 | 9.26 | 0.00 |
| mom_work:yes | 2.54 | 2.35 | 1.08 | 0.28 |
| mom_age | 0.22 | 0.33 | 0.66 | 0.51 |

# Interpreting the slope

What is the correct interpretation of the slope for `mom_work`?

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 19.59 | 9.22 | 2.13 | 0.03 |
| mom_hs:yes | 5.09 | 2.31 | 2.20 | 0.03 |
| mom_iq | 0.56 | 0.06 | 9.26 | 0.00 |
| mom_work:yes | 2.54 | 2.35 | 1.08 | 0.28 |
| mom_age | 0.22 | 0.33 | 0.66 | 0.51 |

*All else being equal, children whose mothers worked during the first three years of the child's life are estimated to score 2.54 points higher than those whose mothers did not work.*

# CI Recap from last time

Inference for the slope for a SLR model (only one explanatory variable):

# CI Recap from last time

Inference for the slope for a SLR model (only one explanatory variable):

- Hypothesis test: $\beta_1$ from $H_0$

$$T = \frac{b_1 - \text{null value}}{SE_{b_1}} \qquad df = n - 2$$

$$\hookrightarrow \frac{1}{\sqrt{n-1}} \cdot \frac{s_e}{s_x}$$

# CI Recap from last time

Inference for the slope for a SLR model (only one explanatory variable):

- Hypothesis test:

$$T = \frac{b_1 - null\ value}{SE_{b_1}} \qquad df = n - 2$$

- Confidence interval:

$$b_1 \pm t^\star_{df} \times SE_{b_1}$$

# CI Recap from last time

Inference for the slope for a SLR model (only one explanatory variable):

- Hypothesis test:

$$T = \frac{b_1 - \textit{null value}}{SE_{b_1}} \qquad df = n - 2$$

- Confidence interval:

$$b_1 \pm t^\star_{df} \times SE_{b_1}$$

# of slope params in the model

The only difference for MLR is that we use $b_i$ instead of $b_1$, and use $df = n - k - 1$. Not that the formular for $SE_{b_i}$ also changes, but you will not be responsible for it in this class.

# CI for the slope

Construct a 95% confidence interval for the slope of `mom_work`.

$$b_k \quad \pm \quad t^\star SE_{b_k}$$

# CI for the slope

Construct a 95% confidence interval for the slope of `mom_work`.

$$
\begin{aligned}
b_k &\pm t^\star SE_{b_k} \\
df &= n - k - 1 = 434 - 4 - 1 = 429 \rightarrow 400
\end{aligned}
$$

# CI for the slope

Construct a 95% confidence interval for the slope of `mom_work`.

$$
\begin{aligned}
b_k &\pm t^\star SE_{b_k} \\
df &= n - k - 1 = 434 - 4 - 1 = 429 \rightarrow 400 \\
2.54 &\pm 1.97 \times 2.35
\end{aligned}
$$

# CI for the slope

Construct a 95% confidence interval for the slope of `mom_work`.

$$
\begin{aligned}
b_k &\pm t^\star SE_{b_k} \\
df &= n - k - 1 = 434 - 4 - 1 = 429 \rightarrow 400 \\
2.54 &\pm 1.97 \times 2.35 \\
2.54 &\pm 4.63
\end{aligned}
$$

# CI for the slope

Construct a 95% confidence interval for the slope of `mom_work`.

$$
\begin{aligned}
b_k &\pm t^\star SE_{b_k} \\
df &= n - k - 1 = 434 - 4 - 1 = 429 \rightarrow 400 \\
2.54 &\pm 1.97 \times 2.35 \\
2.54 &\pm 4.63 \\
(-2.0895 &\quad , \quad 7.1695)
\end{aligned}
$$

# CI for the slope

Construct a 95% confidence interval for the slope of `mom_work`.

$$
\begin{aligned}
b_k &\pm t^\star SE_{b_k} \\
df &= n - k - 1 = 434 - 4 - 1 = 429 \rightarrow 400 \\
2.54 &\pm 1.97 \times 2.35 \\
2.54 &\pm 4.63 \\
(-2.0895 &\quad , \quad 7.1695)
\end{aligned}
$$

Interpretation?

# CI for the slope

Construct a 95% confidence interval for the slope of `mom_work`.

$$
\begin{aligned}
b_k &\pm t^\star SE_{b_k} \\
df &= n - k - 1 = 434 - 4 - 1 = 429 \rightarrow 400 \\
2.54 &\pm 1.97 \times 2.35 \\
2.54 &\pm 4.63 \\
(-2.0895 &\quad , \quad 7.1695)
\end{aligned}
$$

Interpretation?

*We are 95% confident that, all else being equal, children whose mothers worked during the first three years of the child's life are estimated to score between -2.0895 and 7.1695 points higher than those whose mothers did not work.*

# Inference for the slope(s) (cont.)

Given all variables in the model, which variables are significant predictors of kid's cognitive test score?

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 19.59241 | 9.21906 | 2.125 | 0.0341 |
| mom_hsyes | 5.09482 | 2.31450 | 2.201 | 0.0282 |
| mom_iq | 0.56147 | 0.06064 | 9.259 | <2e-16 |
| mom_workyes | 2.53718 | 2.35067 | 1.079 | 0.2810 |
| mom_age | 0.21802 | 0.33074 | 0.659 | 0.5101 |

# Inference for the slope(s) (cont.)

Given all variables in the model, which variables are significant predictors of kid's cognitive test score?

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 19.59241 | 9.21906 | 2.125 | 0.0341 |
| mom_hsyes | 5.09482 | 2.31450 | 2.201 | 0.0282 |
| mom_iq | 0.56147 | 0.06064 | 9.259 | <2e-16 |
| mom_workyes | 2.53718 | 2.35067 | 1.079 | 0.2810 |
| mom_age | 0.21802 | 0.33074 | 0.659 | 0.5101 |

*mom_hs and mom_iq are significant*

*mom_work and mom_age are not.*

# Model selection

# Modeling kid's test scores (revisited)

Predicting cognitive test scores of three- and four-year-old children using characteristics of their mothers. Data are from a survey of adult American women and their children - a subsample from the National Longitudinal Survey of Youth.

| | kid_score | mom_hs | mom_iq | mom_work | mom_age |
|---|---|---|---|---|---|
| 1 | 65 | yes | 121.12 | yes | 27 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 5 | 115 | yes | 92.75 | yes | 27 |
| 6 | 98 | no | 107.90 | no | 18 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 434 | 70 | yes | 91.25 | yes | 25 |

Gelman, Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. (2007) Cambridge University Press.

# Model output

```
cog_full = lm(kid_score ~ mom_hs + mom_iq + mom_work + mom_age,
              data = cognitive)
summary(cog_full)
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19.59241    9.21906   2.125   0.0341
## mom_hsyes    5.09482    2.31450   2.201   0.0282
## mom_iq       0.56147    0.06064   9.259   <2e-16
## mom_workyes  2.53718    2.35067   1.079   0.2810
## mom_age      0.21802    0.33074   0.659   0.5101
##
## Residual standard error: 18.14 on 429 degrees of freedom
## Multiple R-squared: 0.2171, Adjusted R-squared: 0.2098
## F-statistic: 29.74 on 4 and 429 DF,  p-value: < 2.2e-16
```

# Backward-elimination

Adjusted $R^2$ approach:

- Start with the full model
- Drop one variable at a time and record $R^2_{adj}$ of each smaller model
- Pick the model with the largest increase in $R^2_{adj}$
- Repeat until none of the reduced models yield an increase in $R^2_{adj}$

# Backward-elimination

Adjusted $R^2$ approach:

- Start with the full model
- Drop one variable at a time and record $R^2_{adj}$ of each smaller model
- Pick the model with the largest increase in $R^2_{adj}$
- Repeat until none of the reduced models yield an increase in $R^2_{adj}$

When removing a categorical variable all levels should be included or removed *at the same time*

# Backward-selection: $R^2_{adj}$ approach

| Step | Variables included | $R^2_{adj}$ |
|---|---|---|
| Full | kid_score ~ mom_hs + mom_iq + mom_work + mom_age | *0.2098* |

# Backward-selection: $R^2_{adj}$ approach

| **Step** | **Variables included** | $\mathbf{R^2_{adj}}$ |
|---|---|---|
| Full | kid_score ~ mom_hs + mom_iq + mom_work + mom_age | *0.2098* |
| Step 1 | kid_score ~ mom_iq + mom_work + mom_age | 0.2027 |

# Backward-selection: $R^2_{adj}$ approach

| Step | Variables included | $R^2_{adj}$ |
|---|---|---|
| Full | kid_score ~ mom_hs + mom_iq + mom_work + mom_age | *0.2098* |
| Step 1 | kid_score ~ mom_iq + mom_work + mom_age | 0.2027 |
| | kid_score ~ mom_hs + mom_work + mom_age | 0.0541 |

# Backward-selection: $R^2_{adj}$ approach

| Step | Variables included | $R^2_{adj}$ |
|---|---|---|
| Full | kid_score ~ mom_hs + mom_iq + mom_work + mom_age | *0.2098* |
| Step 1 | kid_score ~ mom_iq + mom_work + mom_age | 0.2027 |
| | kid_score ~ mom_hs + mom_work + mom_age | 0.0541 |
| | kid_score ~ mom_hs + mom_iq + mom_age | 0.2095 |

# Backward-selection: $R^2_{adj}$ approach

| Step | Variables included | $R^2_{adj}$ |
|---|---|---|
| Full | kid_score ~ mom_hs + mom_iq + mom_work + mom_age | *0.2098* |
| Step 1 | kid_score ~ mom_iq + mom_work + mom_age | 0.2027 |
| | kid_score ~ mom_hs + mom_work + mom_age | 0.0541 |
| | kid_score ~ mom_hs + mom_iq + mom_age | 0.2095 |
| → | kid_score ~ mom_hs + mom_iq + mom_work | *0.2109* |

# Backward-selection: $R^2_{adj}$ approach

| Step | Variables included | $R^2_{adj}$ |
|---|---|---|
| Full | kid_score ~ mom_hs + mom_iq + mom_work + mom_age | *0.2098* |
| Step 1 | kid_score ~ mom_iq + mom_work + mom_age | 0.2027 |
| | kid_score ~ mom_hs + mom_work + mom_age | 0.0541 |
| | kid_score ~ mom_hs + mom_iq + mom_age | 0.2095 |
| | kid_score ~ mom_hs + mom_iq + mom_work | *0.2109* |
| Step 2 | kid_score ~ mom_iq + mom_work | 0.2024 |

# Backward-selection: $R^2_{adj}$ approach

| Step | Variables included | $R^2_{adj}$ |
|---|---|---|
| Full | kid_score ~ mom_hs + mom_iq + mom_work + mom_age | *0.2098* |
| Step 1 | kid_score ~ mom_iq + mom_work + mom_age | 0.2027 |
| | kid_score ~ mom_hs + mom_work + mom_age | 0.0541 |
| | kid_score ~ mom_hs + mom_iq + mom_age | 0.2095 |
| | kid_score ~ mom_hs + mom_iq + mom_work | *0.2109* |
| Step 2 | kid_score ~ mom_iq + mom_work | 0.2024 |
| | kid_score ~ mom_hs + mom_work | 0.0546 |

# Backward-selection: $R^2_{adj}$ approach

| Step | Variables included | $R^2_{adj}$ |
|---|---|---|
| Full | kid_score ~ mom_hs + mom_iq + mom_work + mom_age | *0.2098* |
| Step 1 | kid_score ~ mom_iq + mom_work + mom_age | 0.2027 |
| | kid_score ~ mom_hs + mom_work + mom_age | 0.0541 |
| | kid_score ~ mom_hs + mom_iq + mom_age | 0.2095 |
| | kid_score ~ mom_hs + mom_iq + mom_work | *0.2109* |
| Step 2 | kid_score ~ mom_iq + mom_work | 0.2024 |
| | kid_score ~ mom_hs + mom_work | 0.0546 |
| | kid_score ~ mom_hs + mom_iq | *0.2105* |

# Backward-selection: $R^2_{adj}$ approach

| Step | Variables included | $R^2_{adj}$ |
|---|---|---|
| Full | kid_score ~ mom_hs + mom_iq + mom_work + mom_age | *0.2098* |
| Step 1 | kid_score ~ mom_iq + mom_work + mom_age | 0.2027 |
| | kid_score ~ mom_hs + mom_work + mom_age | 0.0541 |
| | kid_score ~ mom_hs + mom_iq + mom_age | 0.2095 |
| | kid_score ~ mom_hs + mom_iq + mom_work | *0.2109* |
| Step 2 | kid_score ~ mom_iq + mom_work | 0.2024 |
| | kid_score ~ mom_hs + mom_work | 0.0546 |
| | kid_score ~ mom_hs + mom_iq | *0.2105* |
| Step 3* | kid_score ~ mom_hs | 0.2024 |

# Backward-selection: $R^2_{adj}$ approach

| Step | Variables included | $R^2_{adj}$ |
|---|---|---|
| Full | kid_score ~ mom_hs + mom_iq + mom_work + mom_age | *0.2098* |
| Step 1 | kid_score ~ mom_iq + mom_work + mom_age | 0.2027 |
| | kid_score ~ mom_hs + mom_work + mom_age | 0.0541 |
| | kid_score ~ mom_hs + mom_iq + mom_age | 0.2095 |
| | kid_score ~ mom_hs + mom_iq + mom_work | *0.2109* |
| Step 2 | kid_score ~ mom_iq + mom_work | 0.2024 |
| | kid_score ~ mom_hs + mom_work | 0.0546 |
| | kid_score ~ mom_hs + mom_iq | *0.2105* |
| Step 3* | kid_score ~ mom_hs | 0.2024 |
| | kid_score ~ mom_iq | 0.0546 |

# Forward-selection

Adjusted $R^2$ approach:

- Start with regression of response vs. each explanatory variable
- Pick the model with the highest $R^2_{adj}$
- Add the remaining variables one at a time to the existing model, and once again pick the model with the highest $R^2_{adj}$
- Repeat until the addition of any of the remaining variables does not result in a higher $R^2_{adj}$

# Forward-selection: $R^2_{adj}$ approach

| Step | Variables included | $R^2_{adj}$ |
|---|---|---|
| Step 1 | kid_score ~ mom_hs | 0.0539 |

# Forward-selection: $R^2_{adj}$ approach

| Step | Variables included | $R^2_{adj}$ |
|---|---|---|
| Step 1 | kid_score ~ mom_hs | 0.0539 |
| | kid_score ~ mom_work | 0.0097 |

# Forward-selection: $R^2_{adj}$ approach

| Step | Variables included | $R^2_{adj}$ |
|---|---|---|
| Step 1 | kid_score ~ mom_hs | 0.0539 |
| | kid_score ~ mom_work | 0.0097 |
| | kid_score ~ mom_age | 0.0062 |

# Forward-selection: $R^2_{adj}$ approach

| Step | Variables included | $R^2_{adj}$ |
|---|---|---|
| Step 1 | kid_score ~ mom_hs | 0.0539 |
| | kid_score ~ mom_work | 0.0097 |
| | kid_score ~ mom_age | 0.0062 |
| | kid_score ~ mom_iq | *0.1991* |

# Forward-selection: $R^2_{adj}$ approach

| Step | Variables included | $R^2_{adj}$ |
|---|---|---|
| Step 1 | kid_score ~ mom_hs | 0.0539 |
| | kid_score ~ mom_work | 0.0097 |
| | kid_score ~ mom_age | 0.0062 |
| | kid_score ~ mom_iq | *0.1991* |
| Step 2 | kid_score ~ mom_iq + mom_work | 0.2024 |

# Forward-selection: $R^2_{adj}$ approach

| Step | Variables included | $R^2_{adj}$ |
|---|---|---|
| Step 1 | kid_score ~ mom_hs | 0.0539 |
| | kid_score ~ mom_work | 0.0097 |
| | kid_score ~ mom_age | 0.0062 |
| | kid_score ~ mom_iq | *0.1991* |
| Step 2 | kid_score ~ mom_iq + mom_work | 0.2024 |
| | kid_score ~ mom_iq + mom_age | 0.1999 |

# Forward-selection: $R^2_{adj}$ approach

| Step | Variables included | $R^2_{adj}$ |
|---|---|---|
| Step 1 | kid_score ~ mom_hs | 0.0539 |
| | kid_score ~ mom_work | 0.0097 |
| | kid_score ~ mom_age | 0.0062 |
| | kid_score ~ mom_iq | *0.1991* |
| Step 2 | kid_score ~ mom_iq + mom_work | 0.2024 |
| | kid_score ~ mom_iq + mom_age | 0.1999 |
| | kid_score ~ mom_iq + mom_hs | *0.2105* |
| Step 3 | kid_score ~ mom_iq + mom_hs + mom_age | 0.2095 |

# Forward-selection: $R^2_{adj}$ approach

| Step | Variables included | $R^2_{adj}$ |
|---|---|---|
| Step 1 | kid_score ~ mom_hs | 0.0539 |
| | kid_score ~ mom_work | 0.0097 |
| | kid_score ~ mom_age | 0.0062 |
| | kid_score ~ mom_iq | *0.1991* |
| Step 2 | kid_score ~ mom_iq + mom_work | 0.2024 |
| | kid_score ~ mom_iq + mom_age | 0.1999 |
| | kid_score ~ mom_iq + mom_hs | *0.2105* |
| Step 3 | kid_score ~ mom_iq + mom_hs + mom_age | 0.2095 |
| | kid_score ~ mom_iq + mom_hs + mom_work | *0.2109* |

# Forward-selection: $R^2_{adj}$ approach

| Step | Variables included | $R^2_{adj}$ |
|---|---|---|
| Step 1 | kid_score ~ mom_hs | 0.0539 |
| | kid_score ~ mom_work | 0.0097 |
| | kid_score ~ mom_age | 0.0062 |
| | kid_score ~ mom_iq | *0.1991* |
| Step 2 | kid_score ~ mom_iq + mom_work | 0.2024 |
| | kid_score ~ mom_iq + mom_age | 0.1999 |
| | kid_score ~ mom_iq + mom_hs | *0.2105* |
| Step 3 | kid_score ~ mom_iq + mom_hs + mom_age | 0.2095 |
| | kid_score ~ mom_iq + mom_hs + mom_work | *0.2109* |

# Forward-selection: $R^2_{adj}$ approach

| Step | Variables included | $R^2_{adj}$ |
|---|---|---|
| Step 1 | kid_score ~ mom_hs | 0.0539 |
| | kid_score ~ mom_work | 0.0097 |
| | kid_score ~ mom_age | 0.0062 |
| | kid_score ~ mom_iq | *0.1991* |
| Step 2 | kid_score ~ mom_iq + mom_work | 0.2024 |
| | kid_score ~ mom_iq + mom_age | 0.1999 |
| | kid_score ~ mom_iq + mom_hs | *0.2105* |
| Step 3 | kid_score ~ mom_iq + mom_hs + mom_age | 0.2095 |
| | kid_score ~ mom_iq + mom_hs + mom_work | *0.2109* |
| Step 4* | kid_score ~ mom_iq + mom_hs + mom_age + mom_work | 0.2098 |

# Final model choice

```
cog_final = lm(kid_score ~ mom_hs + mom_iq, data = kid)
summary(cog_final)

## Call:
## lm(formula = kid_score ~ mom_hs + mom_iq, data = kid)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.73154    5.87521   4.380 1.49e-05 ***
## mom_hsyes    5.95012    2.21181   2.690  0.00742 **
## mom_iq       0.56391    0.06057   9.309  < 2e-16 ***
##
## Residual standard error: 18.14 on 431 degrees of freedom
## Multiple R-squared: 0.2141, Adjusted R-squared: 0.2105
## F-statistic: 58.72 on 2 and 431 DF,  p-value: < 2.2e-16
```

# GLMs

# Odds

Odds are another way of quantifying the probability of an event, commonly used in gambling (and logistic regression).

For some event $E$,

$$\text{odds}(E) = \frac{P(E)}{P(E^c)} = \frac{P(E)}{1 - P(E)}$$

Similarly, if we are told the odds of E are $x$ to $y$ then

$$\text{odds}(E) = \frac{x}{y} = \frac{x/(x+y)}{y/(x+y)}$$

which implies

$$P(E) = x/(x+y), \quad P(E^c) = y/(x+y)$$

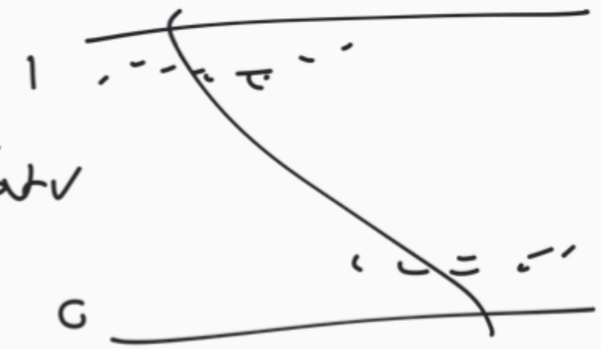
# Example - Donner Party

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 route to the Sacramento Valley. Having reached its full size of 87 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, 40 of the 87 members had died from famine and exposure to extreme cold.

From *Ramsey, Schafer (2002). The Statistical Sleuth*

# Example - Donner Party - Data

| | Age | Sex | Status |
|---|---|---|---|
| 1 | 23.00 | Male | Died |
| 2 | 40.00 | Female | Survived |
| ~~3~~ | ~~40.00~~ | ~~Male~~ | ~~Survived~~ |
| 4 | 30.00 | Male | Died |
| 5 | 28.00 | Male | Died |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 43 | 23.00 | Male | Survived |
| 44 | 24.00 | Male | Died |
| 45 | 25.00 | Female | Survived |




Age

Y

# Example - Donner Party - EDA

Status vs. Gender:

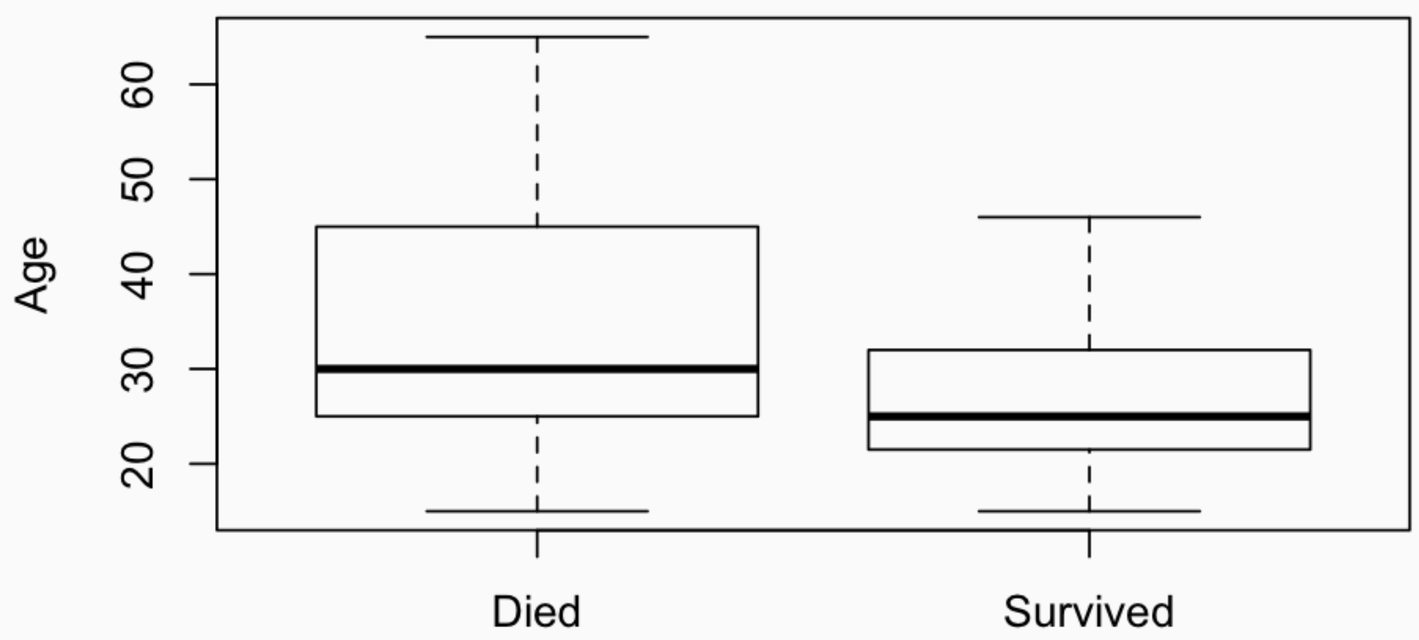
| | Male | Female |
|---|---|---|
| Died | 20 | 5 |
| Survived | 10 | 10 |

# Example - Donner Party - EDA

Status vs. Gender:

| | Male | Female |
|---|---|---|
| Died | 20 | 5 |
| Survived | 10 | 10 |

Status vs. Age:



# Example - Donner Party - ???

It seems clear that both age and gender have an effect on someone's survival, how do we come up with a model that will let us explore this relationship?

# Example - Donner Party - ???

It seems clear that both age and gender have an effect on someone's survival, how do we come up with a model that will let us explore this relationship?

Even if we set Died to 0 and Survived to 1, this isn't something we can reasonably fit a linear model to - we need something more.

# Example - Donner Party - ???

It seems clear that both age and gender have an effect on someone's survival, how do we come up with a model that will let us explore this relationship?

Even if we set Died to 0 and Survived to 1, this isn't something we can reasonably fit a linear model to - we need something more.

One way to think about the problem - we can treat Survived and Died as successes and failures arising from a Bernoulli trial where the probability of a success (survival) is given by a transformation of a linear model of the predictors.

# Generalized linear models

It turns out that this is a very 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 this type of model.

# Generalized linear models

It turns out that this is a very 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 this type of model.

All generalized linear models have the following three characteristics:

# Generalized linear models

It turns out that this is a very 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 this type of model.

All generalized linear models have the following three characteristics:

1. A probability distribution describing the outcome variable

# Generalized linear models

It turns out that this is a very 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 this type of model.

All generalized linear models have the following three characteristics:

1. A probability distribution describing the outcome variable
2. A linear model

$$\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

# Generalized linear models

It turns out that this is a very 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 this type of model.

All generalized linear models have the following three characteristics:

1. A probability distribution describing the outcome variable
2. A linear model

$$\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

3. A link function that relates the linear model to the parameter of the outcome distribution

$$g(p) = \eta \text{ or } p = g^{-1}(\eta)$$

# Logistic Regression

# Logistic Regression

Logistic regression is a GLM used to model a binary categorical variable using numerical and categorical predictors.

We assume a binomial distribution produced the outcome variable and we therefore want to model $p$ the probability of success for a given set of predictors.

# Logistic Regression

Logistic regression is a GLM used to model a binary categorical variable using numerical and categorical predictors.

We assume a binomial distribution produced the outcome variable and we therefore want to model $p$ the probability of success for a given set of predictors.

To finish specifying the Logistic model we just need to establish a reasonable link function that connects $\eta$ to $p$. There are a variety of options but the most commonly used is the logit function.

Logit function:

$$logit(p) = \log\left(\frac{p}{1-p}\right), \text{ for } 0 \leq p \leq 1$$

# Properties of the Logit

The logit function takes a value between 0 and 1 and maps it to a value between $-\infty$ and $\infty$.

Inverse logit (logistic) function:

$$g^{-1}(x) = \frac{\exp(x)}{1 + \exp(x)} = \frac{1}{1 + \exp(-x)}$$

The inverse logit function takes a value between $-\infty$ and $\infty$ and maps it to a value between 0 and 1.

This formulation is also useful for interpreting the model, since the logit can be interpreted as the log odds of a success - more on this later.

# The logistic regression model

The three GLM criteria give us:

$$y_i \sim \text{Bern}(p_i)$$

$$\eta_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_n x_{n,i}$$

$$\text{logit}(p_i) = \eta_i$$

From which we get,

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_n x_{n,i})}{1 + \exp(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_n x_{n,i})}$$

# Example - Donner Party - Model

In R we fit a GLM in the same was as a linear model except we use `glm` instead of `lm`. (We specify the type of GLM to fit using the `family` argument)

```
summary(glm(Status ~ Age, data=donner, family=binomial))
```

```
## Call:
## glm(formula = Status ~ Age, family = binomial, data = donner)
##
## 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 *
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 56.291  on 43  degrees of freedom
## AIC: 60.291
##
```

# Example - Donner Party - Prediction

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 1.8185 | 0.9994 | 1.82 | 0.0688 |
| Age | -0.0665 | 0.0322 | -2.06 | 0.0391 |

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

# Example - Donner Party - Prediction

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 1.8185 | 0.9994 | 1.82 | 0.0688 |
| Age | -0.0665 | 0.0322 | -2.06 | 0.0391 |

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a newborn (Age=0):

# Example - Donner Party - Prediction

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 1.8185 | 0.9994 | 1.82 | 0.0688 |
| Age | -0.0665 | 0.0322 | -2.06 | 0.0391 |

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a newborn (Age=0):

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 0$$

$$\frac{p}{1-p} = \exp(1.8185) = 6.16$$

$$p = 6.16/7.16 = 0.86$$

# Example - Donner Party - Prediction (cont.)

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a 25 year old:

# Example - Donner Party - Prediction (cont.)

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a 25 year old:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 25$$

$$\frac{p}{1-p} = \exp(0.156) = 1.17$$

$$p = 1.17/2.17 = 0.539$$

# Example - Donner Party - Prediction (cont.)

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a 25 year old:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 25$$

$$\frac{p}{1-p} = \exp(0.156) = 1.17$$

$$p = 1.17/2.17 = 0.539$$

Odds / Probability of survival for a 50 year old:

# Example - Donner Party - Prediction (cont.)

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a 25 year old:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 25$$

$$\frac{p}{1-p} = \exp(0.156) = 1.17$$

$$p = 1.17/2.17 = 0.539$$

Odds / Probability of survival for a 50 year old:

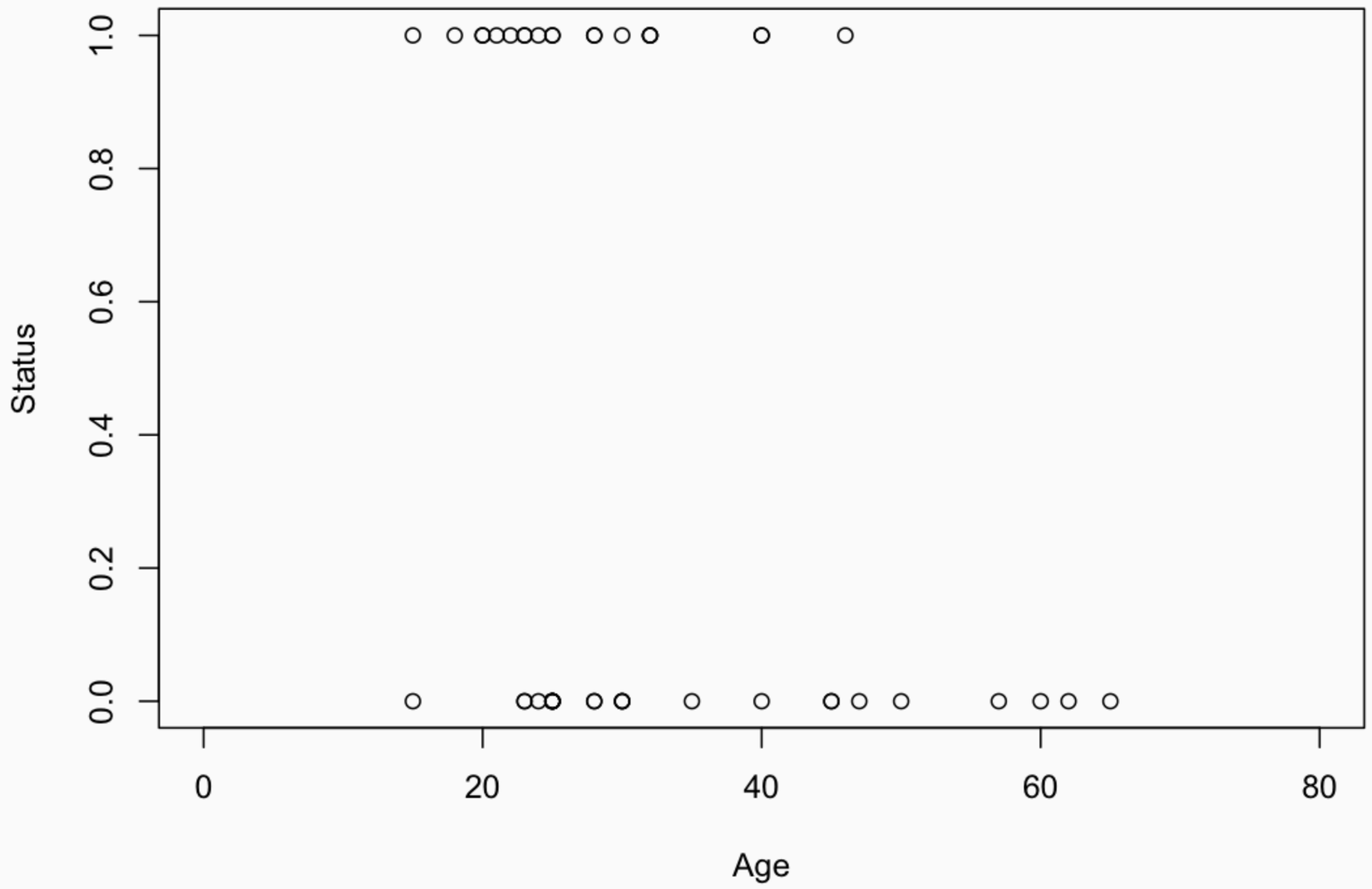
$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 0$$

$$\frac{p}{1-p} = \exp(-1.5065) = 0.222$$

$$p = 0.222/1.222 = 0.181$$

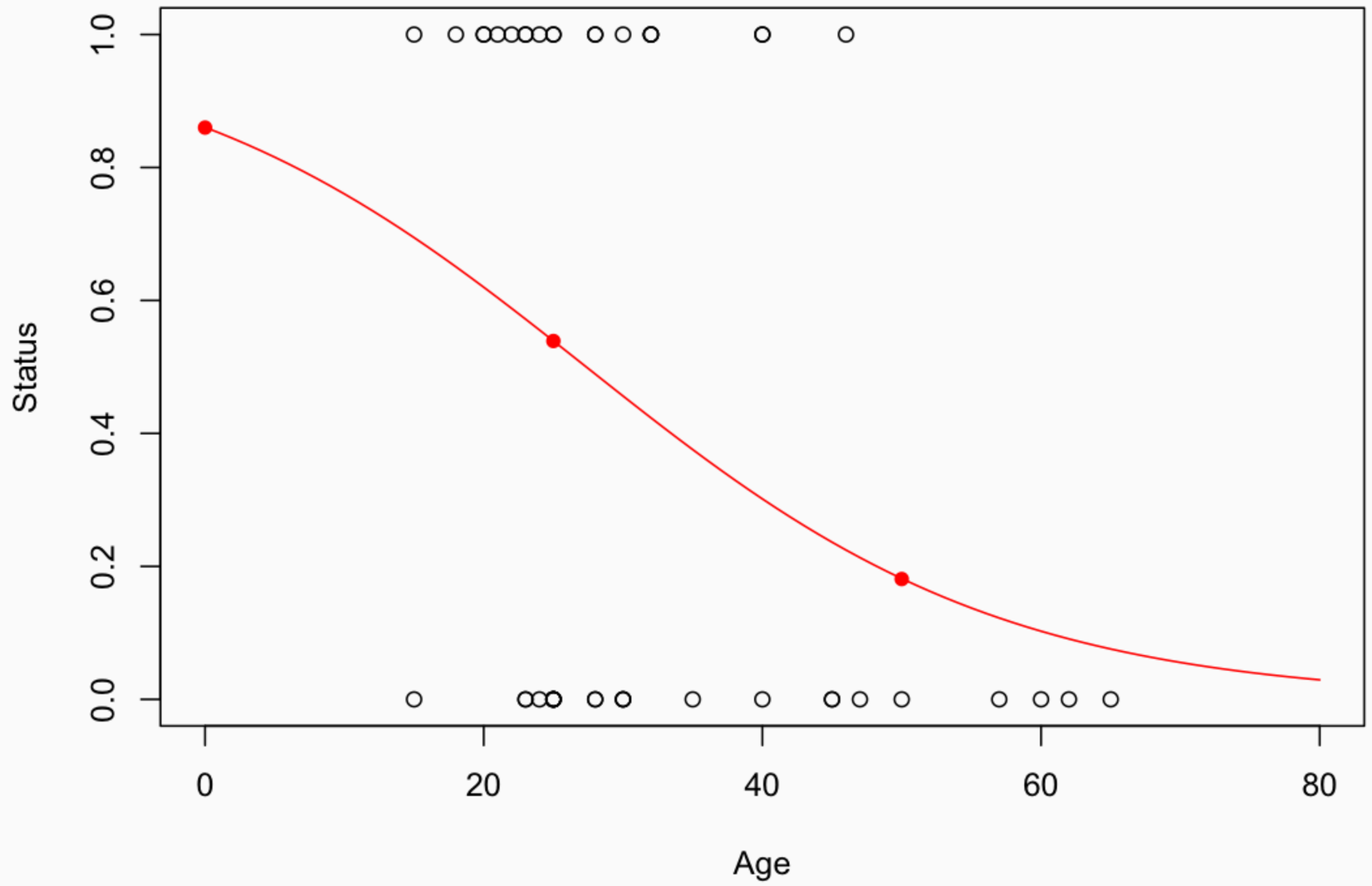
# Example - Donner Party - Prediction (cont.)

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$



# Example - Donner Party - Prediction (cont.)

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$



# Example - Donner Party - Interpretation

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 1.8185 | 0.9994 | 1.82 | 0.0688 |
| Age | -0.0665 | 0.0322 | -2.06 | 0.0391 |

Simple interpretation is only possible in terms of *log odds* and *log odds ratios* for intercept and slope terms.

*Intercept*: The *log odds* of survival for a party member with an age of 0. From this we can calculate the odds or probability, but additional calculations are necessary.

*Slope*: For a unit increase in age (being 1 year older) how much will the *log odds ratio* change, not particularly intuitive. More often than not we care only about sign and relative magnitude.

# Example - Donner Party - Interpretation - Slope

$$\log\left(\frac{p_1}{1-p_1}\right) = 1.8185 - 0.0665(x+1)$$

$$= 1.8185 - 0.0665x - 0.0665$$

$$\log\left(\frac{p_2}{1-p_2}\right) = 1.8185 - 0.0665x$$

$$\log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_2}{1-p_2}\right) = -0.0665$$

$$\log\left(\frac{p_1}{1-p_1} \Big/ \frac{p_2}{1-p_2}\right) = -0.0665$$

$$\frac{p_1}{1-p_1} \Big/ \frac{p_2}{1-p_2} = \exp(-0.0665) = 0.94$$

# Example - Donner Party - Age and Gender

```
summary(glm(Status ~ Age + Sex, data=donner, family=binomial))

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

*Gender slope*: When the other predictors are held constant this is the log odds ratio between the contrast (Female) and the reference level (Male).

# Example - Donner Party - Gender Models

Just like MLR we can plug in gender to arrive at two status vs age models for men and women respectively.

General model:

$$\log\left(\frac{p_1}{1-p_1}\right) = 1.63312 + -0.07820 \times \text{Age} + 1.59729 \times \text{Sex}$$

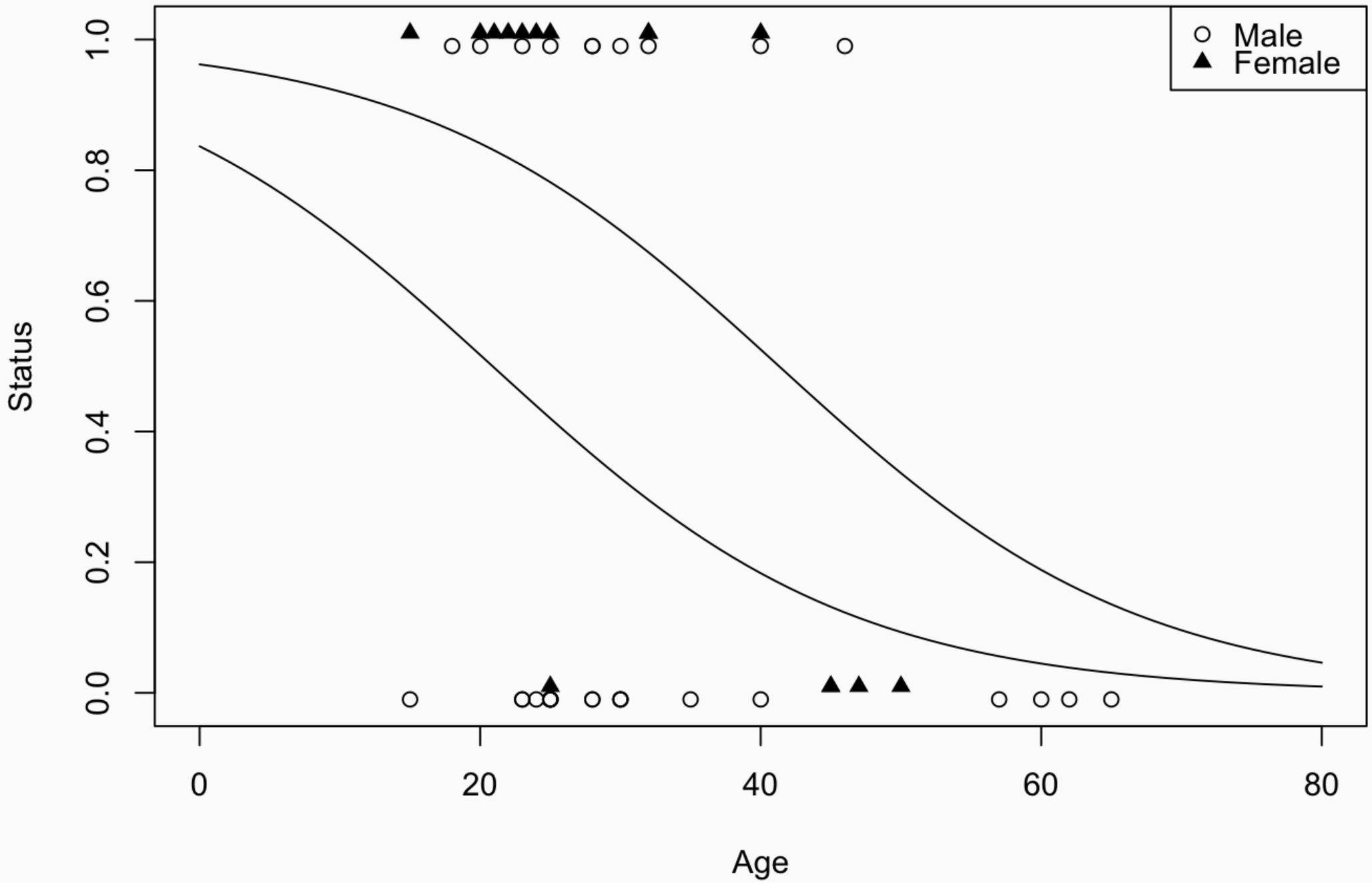
Male model:

$$\begin{aligned}\log\left(\frac{p_1}{1-p_1}\right) &= 1.63312 + -0.07820 \times \text{Age} + 1.59729 \times \textit{0} \\ &= 1.63312 + -0.07820 \times \text{Age}\end{aligned}$$

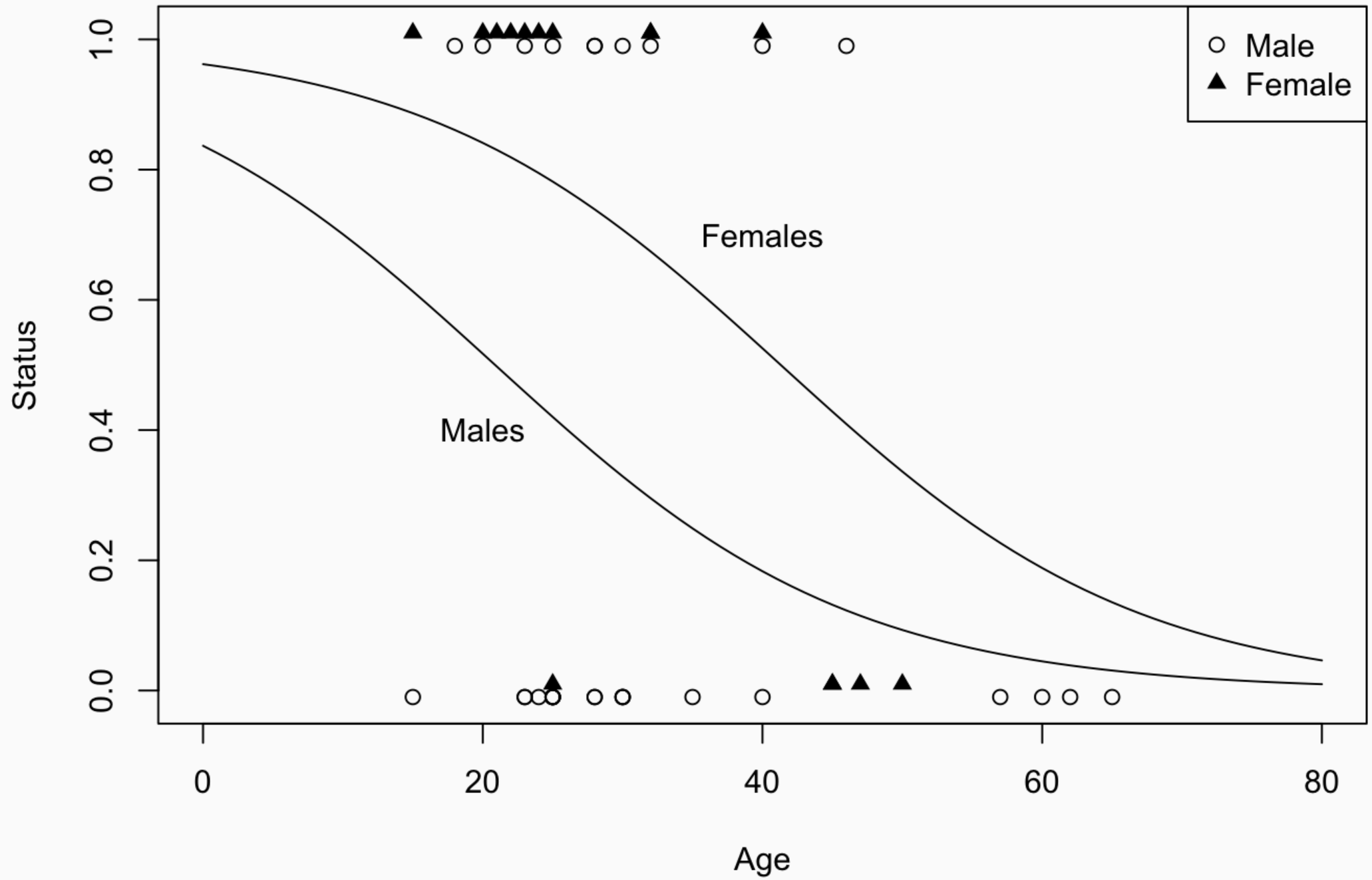
Female model:

$$\begin{aligned}\log\left(\frac{p_1}{1-p_1}\right) &= 1.63312 + -0.07820 \times \text{Age} + 1.59729 \times \textit{1} \\ &= 3.23041 + -0.07820 \times \text{Age}\end{aligned}$$

# Example - Donner Party - Gender Models (cont.)



# Example - Donner Party - Gender Models (cont.)



# Hypothesis test for the model

```
summary(glm(Status ~ Age + Sex, data=donner, family=binomial))

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

# Hypothesis test for the model

```
summary(glm(Status ~ Age + Sex, data=donner, family=binomial))

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

Note that the model output does not include any F-statistic, as a

# Hypothesis tests for a coefficient

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 1.6331 | 1.1102 | 1.47 | 0.1413 |
| Age | -0.0782 | 0.0373 | -2.10 | 0.0359 |
| SexFemale | 1.5973 | 0.7555 | 2.11 | 0.0345 |

We can still perform inference for individual coefficients, the basic framework is the same as SLR/MLR except we use a Z test instead of a t test.

Note the only tricky bit, which is beyond the scope of this course, is how the standard error is calculated.

# Testing for the slope of Age

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 1.6331 | 1.1102 | 1.47 | 0.1413 |
| Age | *-0.0782* | *0.0373* | *-2.10* | 0.0359 |
| SexFemale | 1.5973 | 0.7555 | 2.11 | 0.0345 |

# Testing for the slope of Age

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 1.6331 | 1.1102 | 1.47 | 0.1413 |
| Age | *-0.0782* | *0.0373* | *-2.10* | 0.0359 |
| SexFemale | 1.5973 | 0.7555 | 2.11 | 0.0345 |

$$H_0 : \beta_{age} = 0$$

$$H_A : \beta_{age} \neq 0$$

# Testing for the slope of Age

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 1.6331 | 1.1102 | 1.47 | 0.1413 |
| Age | *-0.0782* | *0.0373* | *-2.10* | 0.0359 |
| SexFemale | 1.5973 | 0.7555 | 2.11 | 0.0345 |

$$H_0 : \beta_{age} = 0$$

$$H_A : \beta_{age} \neq 0$$

$$Z = \frac{\hat{\beta_{age}} - \beta_{age}}{SE_{age}} = \frac{\textit{-0.0782} - 0}{\textit{0.0373}} = \textit{-2.10}$$

$$\text{p-value} = P(|Z| > \textit{2.10}) = P(Z > \textit{2.10}) + P(Z < \textit{-2.10})$$

$$= 2 \times 0.0178 = 0.0359$$

# Confidence interval for age slope coefficient

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 1.6331 | 1.1102 | 1.47 | 0.1413 |
| Age | -0.0782 | 0.0373 | -2.10 | 0.0359 |
| SexFemale | 1.5973 | 0.7555 | 2.11 | 0.0345 |

Remember, the interpretation for a slope is the change in log odds ratio per unit change in the predictor.

# Confidence interval for age slope coefficient

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 1.6331 | 1.1102 | 1.47 | 0.1413 |
| Age | -0.0782 | 0.0373 | -2.10 | 0.0359 |
| SexFemale | 1.5973 | 0.7555 | 2.11 | 0.0345 |

Remember, the interpretation for a slope is the change in log odds ratio per unit change in the predictor.

Log odds ratio:

$$CI = PE \pm CV \times SE = -0.0782 \pm 1.96 \times 0.0373 = (-0.1513, -0.0051)$$

# Confidence interval for age slope coefficient

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 1.6331 | 1.1102 | 1.47 | 0.1413 |
| Age | -0.0782 | 0.0373 | -2.10 | 0.0359 |
| SexFemale | 1.5973 | 0.7555 | 2.11 | 0.0345 |

Remember, the interpretation for a slope is the change in log odds ratio per unit change in the predictor.

Log odds ratio:

$$CI = PE \pm CV \times SE = -0.0782 \pm 1.96 \times 0.0373 = (-0.1513, -0.0051)$$

Odds ratio:

$$\exp(CI) = (\exp(-0.1513), \exp(-0.0051) = (0.8596, 0.9949)$$