Week 4
Logistic Regression

Applied Statistical Analysis II

Jeffrey Ziegler, PhD

Assistant Professor in Political Science & Data Science
Trinity College Dublin

Spring 2023

Roadmap through Stats Land

Where we've been:

- Over-arching goal: We're learning how to make inferences about a population from a sample
- What do we do when our outcome is not continuous?!
  - Need a new framework: GLMs/MLE

Today we will learn:

- How to use this new framework for binary outcome data

- We approach our modeling a little differently when response variable is a binary variable, such as
  - 0 or 1
  - Vote or not
  - Success or failure
- What if we apply previous models (specifically OLS) we've learned?

Ex: Applying OLS to binary outcomes

Do partisans select positive or negative political news stories?[1]

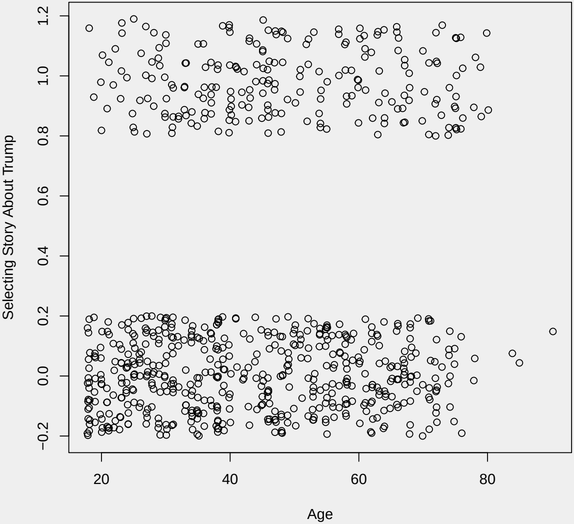
- Age: Age in years
- `SelectTrump`: Whether participant selected news story about Trump ($0 =$did not, $1 =$did)
- N=743 subjects
- Goal: explore relationship between age & likelihood of selecting story about Trump

[1] Kane, J. (2020). "Fight Clubs: Media Coverage of Party (Dis)unity and Citizens' Selective Exposure to It". *Political Research Quarterly*, 73(2), 276-292.

Example: Trump data

```r
# import data
trump_data <- read.
    csv("../datasets/
    Trump_select.csv"
    , header = T,
    stringsAsFactors
    = F)
# inspect
head(trump_data)
```

ID	Age	SelectTrump
1	20	0
2	23	0
3	24	0
4	25	0
5	25	1
6	26	0



Ex: Applying OLS to binary outcomes

Consider usual regression model:

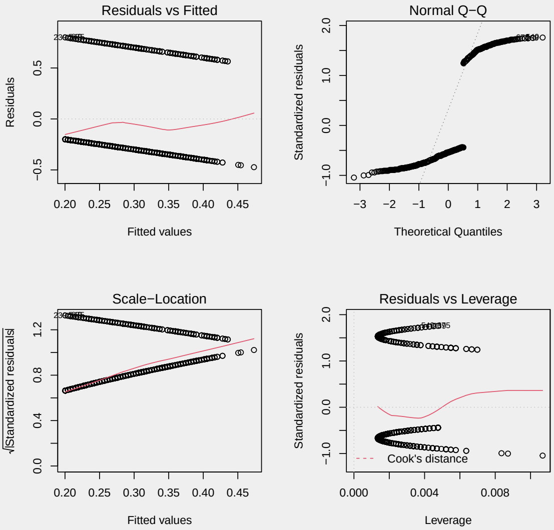
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

```r
# run OLS model with SelectTrump regressed on Age
ols_base <- lm(SelectTrump ~ age, data=trump_data)
```

Fitted model: $\hat{Y}_i = 0.132 + 0.004 X_{age}$

Check linear regression assumptions

```r
plot(ols_base)
```



Issues with using OLS now

(1) Predictions for observations could be outside [0,1]

- For $x = 1$, $\hat{Y} = 0.004$, and this is not possibly an average of y values (0s and 1s) at $x = 1$ (like a conditional mean given x)
- Sometimes for large x, say $x = 1000$, $\hat{Y} \approx 4$, which is not in [0,1]

Issues with using OLS now

(2) Our usual assumptions $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ is not reasonable

- Distributions of Y don't have same variability across all x values
- Means of y at different values of x don't have a linear relationship with x

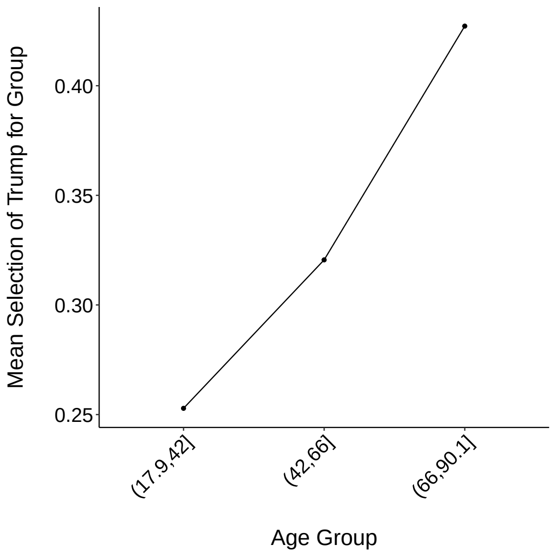
A better model: Logitistic regression

- What is a better way to model a $0 - 1$ response using regression?
- Frequency table of selecting Trump by age group helps

Age Group			**Select Trump**	
	Age	Present	Total	Mean
	(17.9,42]	89	352	0.25
	(42,66]	92	287	0.32
	(66,90.1]	44	103	0.43
Total		225	742	0.30

A better model

Plot of percentage of subjects who selected Trump story in each age



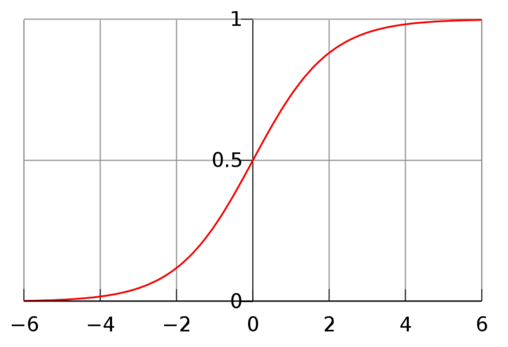
Plot between $E(Y|x)$ and Age

- Estimated values in table are close enough to true values of $E(Y|x)$ to provide a reasonable assessment of functional relationship between *SelectTrump* and *Age*
- With a dichotomous outcome variable, conditional mean must be greater than or equal to zero and less than or equal to 1 ($0 \leq E(Y|x) \leq 1$)
- Plot shows that this mean approaches 0 and 1 “gradually"
  - Δ in $E(Y|x)$ per unit change in x becomes progressively smaller as the conditional mean $\rightarrow$ zero or one
- Curve is said to be "S-shaped" and resembles a plot of cumulative distribution function (CDF) of a continuous random variable

- It shouldn't surprise you that some well-known CDFs have been used to provide a model for $E(Y|x)$ when Y is dichotomous

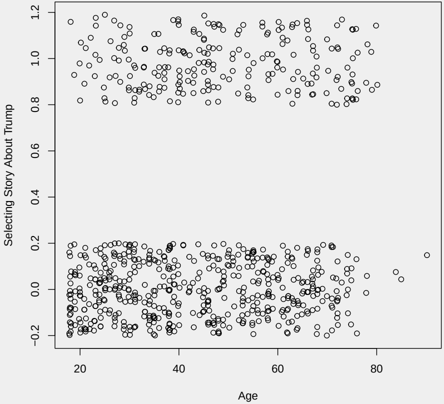
This is the **logistic function**:

$$logit(p) = \ln(\frac{p}{1-p})$$



The logistic model: Intuition

At each x value, there is a certain chance of a 0 and a certain chance of a 1



For example, in this data set, it looks more likely to get a 1 at higher values of x

The logistic model: Setup

- $P(Y = 1)$ changes with x values
- Conditioning on a given x, we have $P(Y = 1|x) \in (0, 1)$
- We will consider this probability of getting a 1 given x values in our modeling

$$\pi_i = P(Y_i = 1|X_i)$$

- Another way to think of the regression is that we are modeling a Bernoulli random variable occurring for each X_i, and the Bernoulli parameter π_i depends on covariate value x

$$Y_i|X_i \sim Bernoulli(\pi_i)$$

where $\pi_i = P(Y_i = 1|X_i)$

$$E(Y_i|X_i) = \pi_i \text{ and } Var(Y_i|X_i) = \pi_i(1 - \pi_i)$$

thinking in terms of the conditional distribution of $Y|X$

- Writing π_i as a function of X_i

$$\pi_i = P(Y_i = 1|X_i) = E(Y_i|X_i)$$

- Writing π_i as a logistic function of X_i

$$\pi_i = P(Y_i = 1|X_i) = E(Y_i|X_i) = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)}$$

- Response variable we will model is a transformation of $P(Y_i = 1)$ for a given X_i

$$logit[P(Y_i = 1|X_i)] = \ln\left(\frac{P(Y_i = 1|X_i)}{1 - P(Y_i = 1|X_i)}\right) = \beta_0 + \beta_1 X_i$$

From simple linear regression to logit

- The **general logistic regression model**:

$$\ln\left(\frac{P(Y_i=1)}{1-P(Y_i=1)}\right) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$$

- Response on left is *logit transformation* of $P(Y_i = 1)$, right side is a *linear regression* model
- **The logistic regression model is a generalized linear model**

Logit Regression is a type of linear model

- Logistic regression model is a *generalized linear model* (glm) and applied when response variable is binary

- Logistic regression model describes expected value of Y ($E[Y]$) in terms of logistic formula:

$$E(Y|X_i) = P(Y_i = 1|X_i) = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})}$$

$$= \frac{1}{1 + \exp[-(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})]}$$

- It can be converted to $\ln$ of odds for $Y_i = 1$:

$$\ln\left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$$

Assumptions of logistic regression

- Unlike OLS regression, logistic regression does not assume
  - linearity between predictors and outcome variable
  - normally distributed errors
  - constant variance
- It does assume
  - we have independent observations
  - that covariates be linearly related to *logit* of outcome variable

Calculation of logistic regression

Remember: **Maximum likelihood estimation (MLE)** is used to calculate the regression coefficient estimates

- Ordinary Least Squares (OLS) minimizes sum of squared residuals
- MLE finds parameter estimates that maximize log-likelihood function

Why use logistic regression model?

This transformation *logit*(p) has many of the desirable properties of a linear regression model

$$P(Y_i = 1) \in (0, 1)$$

$$\text{Odds of } P(Y_i = 1) = \frac{P(Y_i = 1)}{1 - P(Y_i = 1)} \in (0, \infty)$$

$$-\infty < \ln\left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right) < \infty$$

- This response on left is not 'bounded' by $[0, 1]$ even though Y values themselves are bounded by $[0, 1]$
- Response on left can feasibly be any positive or negative quantity from negative infinity to positive infinity

Odds of an event

What are the odds of an event?

when two fair coins are flipped,

$P(\text{two heads}) = \frac{1}{4}$

$P(\text{not two heads}) = \frac{3}{4}$

Odds of an event

The odds in favor of getting two heads is:

$$\text{Odds of getting two heads} = \frac{P(\text{two heads})}{P(\text{not two heads})} = \frac{1/4}{3/4} = \frac{1}{3}$$

or sometimes referred to as 1 to 3 odds

Odds of not getting two heads is 3x larger than odds of getting two heads

- That means, we are 3x as likely to not get two heads as we are to get two heads

Odds of an event

For a binary variable Y (2 possible outcomes), odds in favor of $Y = 1$

$$= \frac{P(Y = 1)}{P(Y = 0)} = \frac{P(Y = 1)}{1 - P(Y = 1)}$$

- For example, if P(selecting Trump story) = 0.3, then
  - Odds of a selecting news story about Trump are $\frac{0.3}{1 - 0.3} = 0.428$

Odds of an event provides a more meaningful interpretation as a measure of association between Y and X

Odds ratio

- **Odds ratio** (OR) for two different groups is a quantity of interest when we interpret association between Y and X

$$\mathrm{OR} = \frac{\frac{\pi(1)}{1-\pi(1)}}{\frac{\pi(0)}{1-\pi(0)}}$$

- Odds of outcome being present among individuals with $x = 1$ is $\frac{\pi(1)}{1-\pi(1)}$, similarly, odds of outcome being present among individuals with $x = 0$ is $\frac{\pi(0)}{1-\pi(0)}$

Example of Odds ratio

- Consider heart attacks for “non-smoker" vs. “smoker"
- Suppose probability of heart attack is 0.0036 for a smoker, and 0.0018 for a non-smoker, then

$$\text{OR} = \frac{\text{odds of a heart attack for non-smoker}}{\text{odds of a heart attack for smoker}}$$

$$\text{OR} = \frac{\frac{P(\text{heart attack}|\text{non-smoker})}{1-P(\text{heart attack}|\text{non-smoker})}}{\frac{P(\text{heart attack}|\text{smoker})}{1-P(\text{heart attack}|\text{smoker})}}$$

$$\text{OR} = \frac{\frac{0.0018}{0.9982}}{\frac{0.0036}{0.9964}} = 0.4991$$

- Odds of having a heart attack among smokers is 2X greater than odds among non-smokers

- **Odds ratio** is widely used as a measure of association as it approximates how much more likely or unlikely (in terms of odds) it is for outcome to be present among those subjects with $X = x + 1$ as compared to those subjects with $X = x$
- So, how do we set up odds ratio with what we already know...?
- Remember, logistic function:

$$\frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)}$$

Constructing Odds ratios

P(heart attack | non-smoker)

$$= \frac{e^{\beta_0+\beta_1(x+1)}}{1+e^{\beta_0+\beta_1(x+1)}}$$

For a logistic regression model with one unit increase in predictor X, so relationship between odds ratio and regression coefficient is...

$$\mathrm{OR} = \frac{\frac{\frac{e^{\beta_0+\beta_1(x+1)}}{1+e^{\beta_0+\beta_1(x+1)}}}{\frac{1}{1+e^{\beta_0+\beta_1(x+1)}}}}{\frac{\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}}{\frac{1}{1+e^{\beta_0+\beta_1 x}}}}$$

$$= \frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1 x}}$$

$$= e^{\beta_1}$$

Interpretation of odds ratio

- $\text{OR} \geq 0$
- If $\text{OR} = 1$, then $P(Y = 1)$ is same in both samples
- If $\text{OR} < 1$, then $P(Y = 1)$ is less in numerator group than in denominator group
- If $\text{OR} = 0$, if and only if $P(Y = 1) = 0$ in numerator sample

Logit Example: Trump vs Age

Let's fit a logit model for selecting news story about Trump (`SelectTrump`) with `Age`

```
# run logit model with SelectTrump regressed on Age
logit_base <- glm(SelectTrump ~ age, data=trump_data, family=
    binomial(link="logit"))
```

Age	0.018***
	(0.005)
Constant	−1.663***
	(0.237)
N	742
Log Likelihood	−447.944
AIC.	899.888

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

So now what, how do we interpret results?

Fitted logistic regression model:

$$\text{Estimated} \quad \ln\left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right) = \hat{\beta}_0 + \hat{\beta}_1 X_i = -1.663 + 0.018 X_{age}$$

and when we solve for $P(Y_i = 1)$ we get...

$$\hat{E}(Y_i|x_i) = \hat{P}(Y_i = 1|x_i) = \frac{\exp(-1.663 + 0.018x_i)}{1 + \exp(-1.663 + 0.018x_i)}$$

- Value on right is bounded to $[0, 1]$
- Because our β_1 is positive,

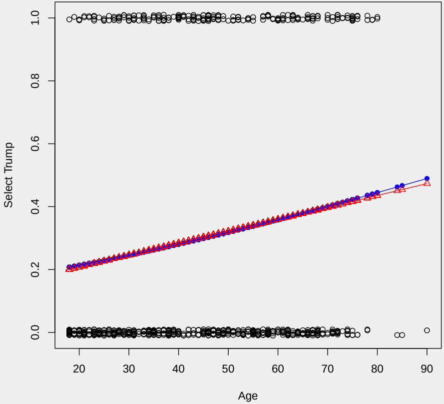
  as X gets larger, $P(Y_i = 1)$ goes to 1;

  as X gets smaller, $P(Y_i = 1)$ goes to 0

Fitted logistic model curve

Fitted curve, S shaped, represents $E(Y_i|x_i) = P(Y_i = 1|x_i)$

```r
plot(trump_data$age,jitter(trump_data$SelectTrump,0.05),xlab="Age",ylab="Select Trump")
points(trump_data$age, fitted(logit_base), col="blue", pch=16)
points(trump_data$age, fitted(ols_base), col="red", pch=2)
curve(predict(ols_base, data.frame(age=x), type="response"), col="red", add=T)
curve(predict(logit_base, data.frame(age=x), type="response"), col="blue", add=T)
```



Remember, estimation of logit model

In general case, we model one covariate like...

$$\ln\left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right) = \beta_0 + \beta_1 X_i$$

which means

$$E(Y_i|x_i) = P(Y_i = 1|x_i) = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)}$$

$$= \frac{1}{1 + \exp[-(\beta_0 + \beta_1 X_i)]}$$

- If β_1 is positive, we get S shape that goes to 1 as X_i goes to ∞ and goes to 0 as X_i goes to $-\infty$
- If β_1 is negative, we get opposite S shape that goes to 0 as X_i goes to ∞ and goes to 1 as X_i goes to $-\infty$

Interpretation of β_0

- When $X_i = 0$, subject with age of 0, then

$$\text{Estimated} \quad \ln\left(\frac{P(Y_i = 1|X_i = 0)}{1 - P(Y_i = 1|X_i = 0)}\right) = \hat{\beta}_0 = -1.663$$

  - $\hat{\beta}_0$ is estimated log-odds that a randomly selected baby with age=0 will select news story about Trump

- Interpret $e^{\hat{\beta}_0} = 0.1896$

  - $e^{\hat{\beta}_0}$ is estimated odds that a randomly selected patient Age=0 selects news story about Trump
  - So...

$$\text{Estimated Pr(Select=1 | Age=0)} = \hat{P}(Y_i = 1|X_i = 0) = \hat{\pi}_i = \frac{e^{\hat{\beta}_0}}{1 + e^{\hat{\beta}_0}}$$

Interpretation of β_1

- Consider log of odds ratio (OR) of selecting Trump for following two groups:
  - those with Age = x
  - those with Age = $x + 1$

$$\ln(OR) = \ln\left(\frac{\frac{\pi_2}{1-\pi_2}}{\frac{\pi_1}{1-\pi_1}}\right) = \ln\left(\frac{\exp(\beta_0 + \beta_1(x+1))}{\exp(\beta_0 + \beta_1 x)}\right) = \ln(e^{\beta_1}) = \beta_1$$

- β_1 is log-odds for a 1 unit increase in x
  - Compares the groups with x exposure and $x + 1$ exposure

Interpretation of β_1

In other words...

- $e^{\hat{\beta}_1}$ is estimated odds ratio (OR) comparing two groups
- Interpret $e^{\hat{\beta}_1} = 1.018$:

  A one unit increase in X increases odds of selecting news story about Trump by a multiplicative factor of 1.018, it increases the odds by $\approx 2\%$

Testing individual coefficient $H_0 : \beta_j = 0$

- Estimates $\hat{\beta}_j$ are asymptotically normal, so R uses Z-tests (or Wald tests) for covariate significance in output
- General concepts of a linear regression model are easily extended to logistic regression with many covariates
- Hypothesis test:
  - $H_0 : \beta_j = 0$
  - $Z_{\text{ts}} = \frac{\hat{\beta}_j}{se_{\hat{\beta}_j}}$
  - If p-value $< \alpha$, then we can conclude that predictor is a significant predictor in logistic regression model
- A $(1 - \alpha) \times 100\%$ C.I. for β_j is: $(\hat{\beta}_j \pm Z_{\alpha/2} se_{\hat{\beta}_j})$

Significance Testing

Age	0.018^{***}
	(0.005)
Constant	-1.663^{***}
	(0.237)
N	742
Log Likelihood	-447.944
AIC.	899.888

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Testing several coefficients

- $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$
- H_a : at least one slope is equal to 0
- Compare full with reduced model
  - Reduced model is model without $X_1, \cdots, X_p$
- Likelihood ratio tests, which are chi-squared tests (χ^2 tests)
- We can use *anova()* function in R to do these likelihood ratio nested tests
  - `anova(glm.reduced, glm.full, test="LRT")`
- If p-value $< \alpha$, then we can conclude that at least one predictor is a significant predictor in logistic regression model

Prediction

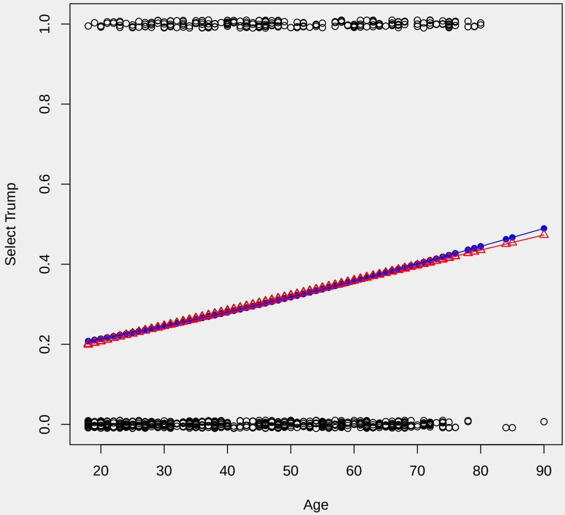
What is predicted probability of selecting Trump news story for a subject with $X = 69$?

$$P(Y_i = 1|X_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

$$= \frac{1}{1 + \exp[-(\beta_0 + \beta_1 x_i)]}$$

$$= \frac{1}{1 + \exp[-(-1.663 + 0.018 \times 69)]} = 0.396$$

Prediction



Another example: Characteristics of Judges

Does having daughters cause judges to rule for women's issues? (Glynn and Sen 2015)

- Outcome: Voted for decision that was more in favor of women in a gender-related case (1=yes, 0=no)
- Predictors:
  - # of daughters (>0 = 1, 0=0)
  - Republican/Democrat

Example: fitted multiplicative model

```r
judge_model <- glm(progressive.vote ~ girls_atLeast1 * republican,
                   family=binomial(link="logit"), data = women.subset)
```

(Intercept)	-0.349^{*}
	(0.160)
> 0 Girls	0.315
	(0.177)
Republican	-0.547^{**}
	(0.207)
> 0 Girls × Republican	-0.036
	(0.232)
AIC	2612.584
BIC	2634.935
N	1974

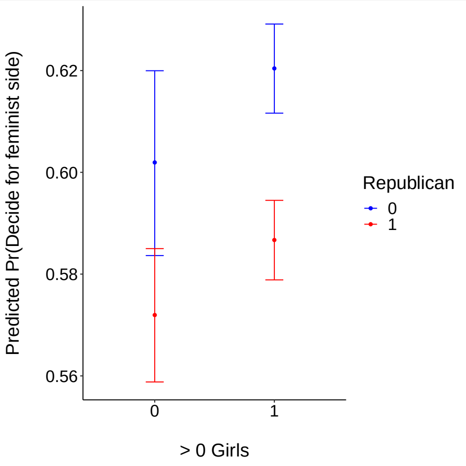
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Example: Interpretation

- Intercept: When a judge has no girls and is a Democrat, expected odds that judge sides with more feminist decision = $exp(-0.349) = 0.705$ (baseline odds ratio)
  - So, on average, judges do not favor feminist side
- > 0 Girls: For Democrats, having at least one girl increases the log odds of favoring feminist decision by 0.315 (relationship not differentiable from zero though...)
- Republican: When a judge has zero girls, being a Republican decreases the log odds of favoring feminist decision by 0.547 (& statistically reliable relationship!)
- Interaction: When a judge has at least one girl and is a Republican the log odds of favoring feminist decision decrease by 0.036 (not a statistically reliable relationship)

Example: Predicted probabilities

```
# create data with predicted values,
    including CIs
predicted_data <- with(women.subset,
    data.frame(girls_atLeast1 = as.
    factor(c(0, 1, 0, 1)), republican
    = as.factor(c(0, 1, 1, 0))))
predicted_data <- cbind(predicted_data,
    predict(judge_model, newdata =
    predicted_data, type = "response",
     se = T))
predicted_data <- within(predicted_data,
      {
  PredictedProb <- plogis(fit)
  LL <- plogis(fit - (1.96 * se.fit))
  UL <- plogis(fit + (1.96 * se.fit))
})
```



Fit of logit regression models

Deviance (or residual deviance)

- This is used to assess model fit
- In logistic regression, deviance has flavor of residual sum of squares (RSS) in ordinary regression
- Smaller the deviance, the better the fit

Null deviance

- It is similar to RSS in ordinary regression when only an overall mean is fit ($df = n - 1$)

Residual deviance

- It is similar to RSS in ordinary regression from full model fit ($df = n - k - 1$)

Likelihood ratio/χ^2 test

- A comparison of null deviance and residual deviance is used to test global null hypothesis

$$H_0 : \text{all slopes} = 0$$

$$H_a : \text{at least one } \beta_j \text{ not equal to } 0$$

- A likelihood ratio test is used for this nested test which follows a central χ^2 distribution under H_0 being true

Likelihood ratio/χ^2 test

- χ^2 is a chi-squared distribution with k degrees of freedom
  - k = # of coefficients being made in H_0 (# of covariates in full model for global null hypothesis test)

$$\chi^2_k = \text{Reduced model deviance} - \text{Full model deviance}$$
$$= \text{Null deviance} - \text{Residual deviance}$$

Likelihood ratio/χ^2 test

- This can be done using full vs. reduced likelihood ratio test

```r
judge_null <- glm(progressive.vote ~ 1, family=binomial(link="
    logit"), data = women.subset)
```

```
Coefficients:
(Intercept)
-0.4194

Null deviance: 2651.6  on 1973  degrees of freedom
Residual deviance: 2604.6  on 1970  degrees of freedom
```

Analysis of deviance table

```r
anova(judge_null, judge_interaction, test ="LRT")
```

```
Model 1: progressive.vote ~ 1
Model 2: progressive.vote ~ girls_atLeast1 * republican

Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      1973     2651.6
2      1970     2604.6  3   47.022 3.438e-10 ***
```

- p value < 0.01, we can conclude that at least one predictor is reliable in model

Individual tests for coefficients

- After rejecting global null hypothesis, we can consider individual Z-tests for predictors

(Intercept)	-0.349^{*}
	(0.160)
> 0 Girls	0.315
	(0.177)
Republican	-0.547^{**}
	(0.207)
> 0 Girls $\times$ Republican	-0.036
	(0.232)
AIC	2612.584
BIC	2634.935
N	1974

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

- *Republican* is a significant predictor, given we account for # of girls a judge has

Significance test for different slopes

- $H_0 : \beta_{\text{# of girls} \mid \text{Democrat}} = \beta_{\text{# of girls} \mid \text{Republican}}$
- H_a : Effect of having girls is different by party

```
Model 1: progressive.vote ~ girls_atLeast1 + republican
Model 2: progressive.vote ~ girls_atLeast1 * republican

Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       1971     2604.6
2       1970     2604.6  1 0.023961    0.877
```

- So, there is not evidence that including an interactive effect of party and # of girls a judge has is a significant predictor for odds of deciding in a feminist direction
    - We also could have gathered this from individual coefficient...

The "best" fitted model

	Model 1
(Intercept)	-0.33^{**}
	(0.12)
> 0 girl	0.29^{*}
	(0.11)
Republican	-0.58^{***}
	(0.09)
AIC	2610.61
BIC	2627.37
Log Likelihood	-1302.30
Deviance	2604.61
Num. obs.	1974

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Interpretation of estimates

$$\hat{\beta}_{\text{>=1 Girl}} = 0.29 \;\; \text{or} \;\; e^{\hat{\beta}_{\text{>=1 Girl}}} = 1.336$$

- Controlling value of party, having at least one girl is associated with an increase in the odds of survival by a factor of 1.336
- Holding value of party constant, having at least one girl increases odds of survival by a multiplicative factor of 1.336, it increases odds by 29%
- $e^{\hat{\beta}_{\text{>=1 Girl}}}$ represents multiplicative effect of having >= one girl

$$\text{Odds}_{\text{>=1 Girl}} = 1.336 \times \text{Odds}_{\text{No Girls}}$$

Outliers, Influential data

car library diagnostics can also be used on generalised linear models:

- Remember, *rstudent, hatvalues, cookd, vif, outlierTest, etc.*

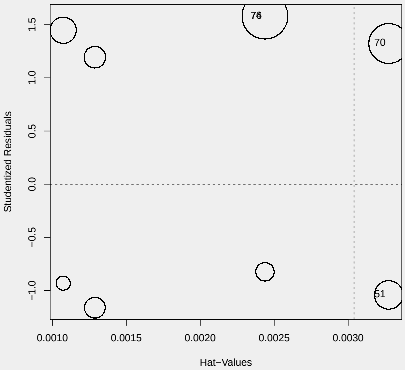
```
girls_atLeast1     republican
1.002729       1.002729
```

```
vif(judge_additive)
```

```
No Studentized residuals with Bonferroni p < 0.05

Largest |rstudent|:
rstudent unadjusted p-value Bonferroni p
74 1.580987           0.11388          NA
```

Outliers, Influential data

```
pdf("../../graphics/influencePlot_logit.pdf")
```




```
   StudRes         Hat       CookD
51 -1.041087 0.003274238 0.000788022
70  1.323106 0.003274238 0.001531570
74  1.580987 0.002437317 0.002023854
76  1.580987 0.002437317 0.002023854
```

Diagnostics: Goodness of fit

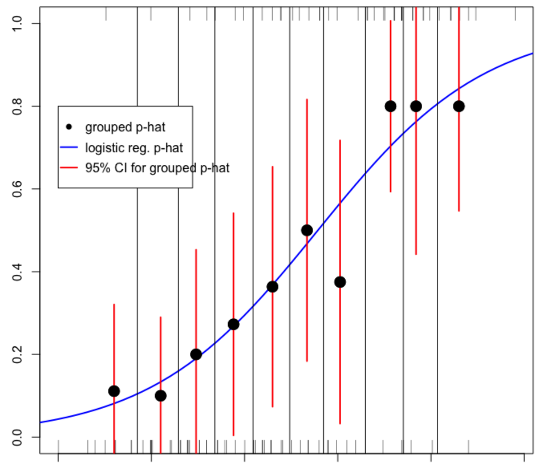
- Since responses are all 0's and 1's, it is more difficult to check how well model fits our data (compared to ordinary regression)
- If you have data points fairly evenly spread across your x values, we could try to check fit using *Hosmer-Lemeshow Goodness of fit test*

Diagnostics: Goodness of fit

- Fitted value is a probability or ($\hat{p}$)
- The logistic regression provides a $\hat{p}$ for every x value
- To check the fit, we will partition observations into 10 groups based on the x values
- For each group g, we will estimate a $\hat{p}_g$
- $\hat{p}_g$ does not consider the other fitted $\hat{p}$ values (unlike the logistic regression fitted values which all fall along a smooth curve)
- Now we will compare the ‘freely fit’ estimated probabilities, with logistic regression fitted probabilities

The Hosmer-Lemeshow test curve

Vertical lines represent grouping structure of observations



If the dots fall close to the fitted logistic curve, it is a reasonably good fit, the short red lines represent a 95% C.I. around $\hat{p}$

The Hosmer-Lemeshow Test: Judge Progressivism

Hosmer-Lemeshow Test takes these values and tests goodness of fit using a χ^2 test statistic

- H_0 : Fit is sufficient

  H_a : Fit is not sufficient (fitted curve doesn't explain data very well)

```
hoslem.test(women.subset$progressive.vote, fitted(judge_
    additive))
```

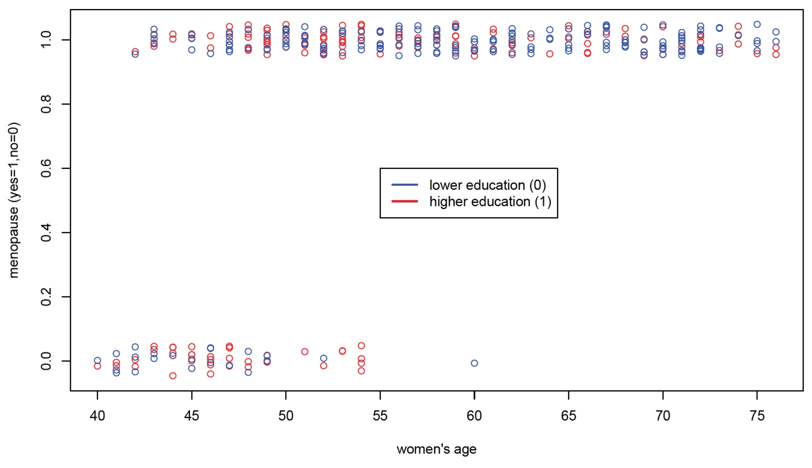
```
Hosmer and Lemeshow goodness of fit (GOF) test

data:  women.subset$progressive.vote, fitted(judge_additive)
X-squared = 0.013809, df = 8, p-value = 1
```

- Cannot reject null hypothesis that logit regression model describes data reasonably well

Example: Menopause presence among women

- Binary response variable: *Menopause* (present: 1; not present: 0)
- Predictor variables:
  - One continuous variable: age
  - One categorical variable: *highed* (lower education: 0; higher education: 1)

Example: Menopause and age



Example: Run logit regression

```
> glm.out=glm(menopause ~ age + highed,
                        family=binomial(logit))
> summary(glm.out)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.51368    1.97708  -6.329 2.46e-10 ***
age           0.28383    0.04117   6.894 5.42e-12 ***
highed       -0.70705    0.36576  -1.933   0.0532 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 313.01  on 337  degrees of freedom
Residual deviance: 193.16  on 335  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 199.16

Number of Fisher Scoring iterations: 7
```

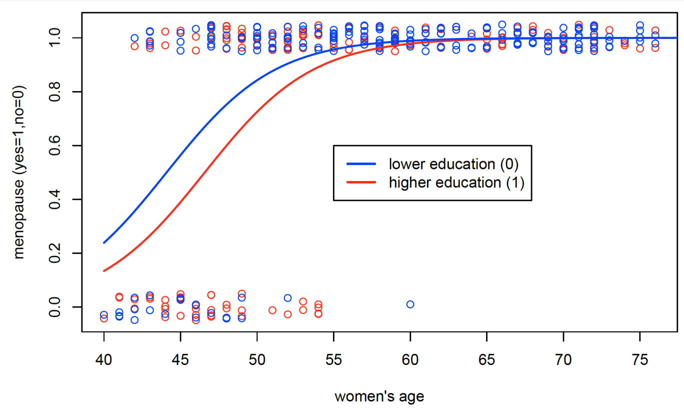
Menopause: Interpretation of results

- For women in a given `highed` level (i.e. holding `highed` constant), increasing age by 1 year is associated with an increase in the odds of having entered menopause by a factor of $e^{0.28383} = 1.3267$
- `highed` is an almost significant predictor $\hat{\beta}_{\text{highed}} = -0.70705$

Menopause: Interpretation of results

- age is a significant predictor of the probability of having entered menopause
- For a women in a given age group (i.e. holding age constant), an increase in highed (having more education, or changing highed from 0 to 1) is associated with a decrease in the probability of having entered menopause
- Specifically, in a given age group, odds of a woman with higher education having entered menopause is $e^{-0.70705} = 0.4931$ times less than odds of a woman with lower education having entered menopause

EXAMPLE: FITTED CURVE



INTERPRETATION OF THE ADDITIVE MODEL

- In terms of interpretation of parameters, a 1 unit change in *age* has same effect on both groups because there is no interaction in this model
- Fitted curves are not on top of each other because one is shifted to right of other, shift is present because of *highed* effect
- Predicted probability of a woman having entered menopause at a given age, is different between two groups

Possible interaction model

- If a model was fit that included interaction, these two curves could feasibly crossover each other
- But, as with classical regression models, there are a variety of forms of interaction, and shape of fitted curves will depend on data

Wrap Up

In this lesson, we went over how to...

- Intuition behind logitistic regression
- Assumptions of logitistic regression
- How to estimate logitistic regressions
- Interpret individual coefficients from additive logit models
  - And, logit models with interactions
- Diagnose assumptions and determine model fit of logit regressions