POL 213 – Spring 2024

Lecture 6

Binary Choice Models II - Interpreting the Logit Model

Lauren Peritz

U.C. Davis

lperitz@ucdavis.edu

May 23, 2024

Table of Contents

Regression with Binary Dependent Variables

Last time...

- Discussed motivations for Logistic and Probit regression
- Demonstrated how linear probability model has several undesired features
- Derived the logistic regression model
- Showed how logistic regression results can be interpreted as (log) odds ratios

This time...

- Practice with logistic regression in R.
- Interpretation of logistic regression results and odds ratios.
- Working with interaction model specifications.

Practice with Logistic Regression

Please use the following files:

`Field.r` and `field_201202.txt`

The data and example code are from Scott MacKenzie, who kindly shared the teaching materials. The data reflect a 2012 Field Poll on support for same sex marriage, conducted among 515 California residents.

Table of Contents

Interpreting the Logit Model

The logit model is nonlinear in probabilities but linear in the log-odds; how do we interpret α and β?

- Example: Support for same sex marriage is the binary response variable. The predictors are: age, party affiliation (-1,0,1) and born-again Christian.

```
> summary(mydata)
same_sex           age              party              born_christian
Min.   :0.0000   Min.   :1.000   Min.   :-1.00000   Min.   :0.0000
1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:-1.00000   1st Qu.:0.0000
Median :1.0000   Median :4.000   Median : 0.00000   Median :0.0000
Mean   :0.5586   Mean   :4.044   Mean   : 0.02092   Mean   :0.3033
3rd Qu.:1.0000   3rd Qu.:6.000   3rd Qu.: 1.00000   3rd Qu.:1.0000
Max.   :1.0000   Max.   :6.000   Max.   : 1.00000   Max.   :1.0000
```

Interpreting the Logit Model

- Plot Support as a function of Age

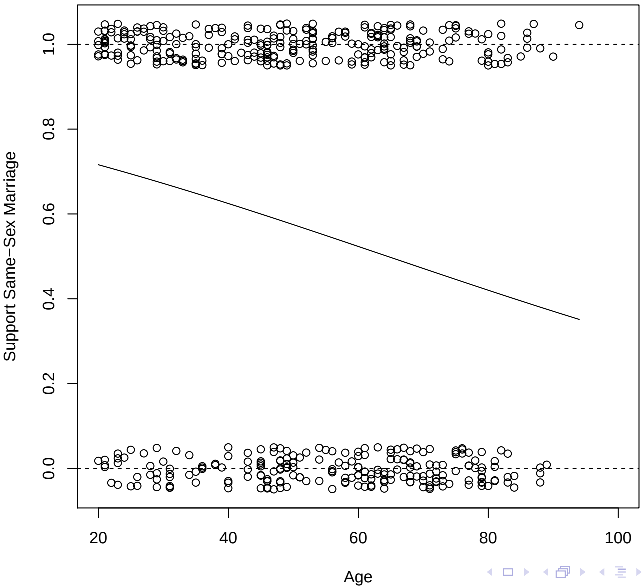
```
# Scatterplot of Support for Same-sex  Marriage against Age

plot(jitter(same_sex, .25) ~ age, mydata,  xlab="Age", xlim=c(20,100)
ylab="Support Same-Sex Marriage")
abline(h = 1, lty = 2)
abline(h = 0, lty = 2)
```

- We can estimate the logit model for π_i, i.e. $\Pr(Y_i = yes)$

$$\pi_i = \Lambda(\alpha + \beta X_i) = \frac{1}{1 + exp(-\alpha + \beta X_i)}$$

where support is a function of the respondent's age.


Support Same−Sex Marriage
0.0
0.2
0.4
0.6
0.8
1.0
20
40
60
80
100
Age

Interpreting the Logit Model

- β is negative and indicates the effect of the log-odds of a 1-year increase in age.
  - When β is negative (positive), the curve descends (ascends)
  - The rate of change increases as $|\beta|$ increases
- Since the logit is S-shaped in probabilities, the rate of change in π_i depends on x. We can write slope = $\beta(\pi_i(1-\pi_i))$

```
-0.004 = -0.020776(0.35*0.65)
-0.005 = -0.020776(0.56*0.44)
-0.004 = -0.020776(0.72*0.28)
```

```
> logit.field <- glm(same_sex   age, mydata, family = binomial)

> summary(logit.field)


Call: glm(formula = same_sex   age, family = binomial, data = mydata)


Deviance Residuals:
Min 1Q Median 3Q Max
-1.587 -1.200 0.854 1.069 1.446


Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.340342 0.291319 4.601 4.21e-06 ***
age -0.020776 0.005166 -4.022 5.77e-05 ***


Null deviance: 654.43 on 476 degrees of freedom
Residual deviance: 637.65 on 475 degrees of freedom
AIC: 641.65


Number of Fisher Scoring iterations: 4
```

Interpreting the Logit Model

- Since β is negative, the estimated probability is smaller at greater ages

$$\frac{\exp(1.3403 - 0.0207 * 20)}{1 + \exp(1.3403 - 0.0207 * 20)} = 0.716$$

$$\frac{\exp(1.3403 - 0.0207 * 94)}{1 + \exp(1.3403 - 0.0207 * 94)} = 0.351$$

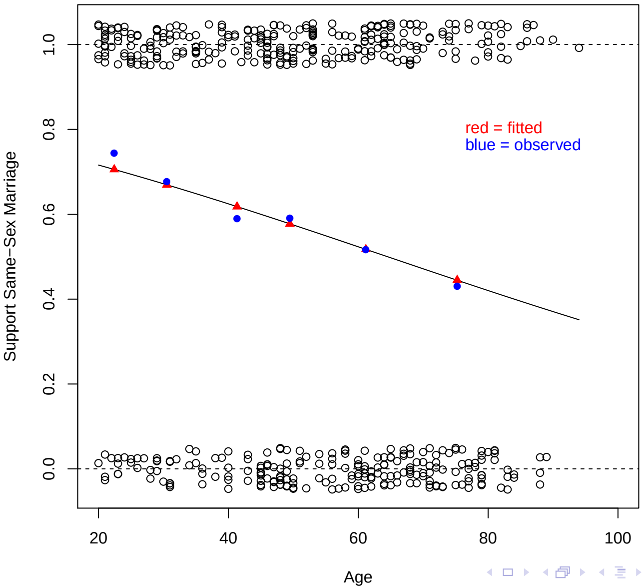
- The rate of change is greatest when $\pi_i = 0.50$; the age (i.e., median effective level) this occurs is $\dfrac{-\alpha}{\beta}$

$$-\frac{\alpha}{\beta} = -\frac{1.3403}{0.0207} = 64.514$$

$$\frac{\exp(1.3403 - 0.0207 * 64.5)}{1 + \exp(1.3403 - 0.0207 * 64.5)} = 0.50$$

Interpreting the Logit Model

- Plotting y against x when y takes only 0 and 1 values can provide only limited information about whether a logit model is reasonable
- By grouping age values into categories and plotting proportions, we can better assess the trend
- We can also plot the fitted values (grouped or not) and compare; the fit here looks adequate.


red = fitted
blue = observed
Support Same−Sex Marriage
Age
0.0
0.2
0.4
0.6
0.8
1.0
20
40
60
80
100

Interpreting the Logit Model

- One advantage of the logit model is that we can interpret the coefficients about as easily as those we get from OLS.
- Recall:

$$\log \frac{\pi_i}{1-\pi_i} = \alpha + \beta X_i$$

- This means the odds of a success versus failure (i.e. a respondent supports versus opposes same-sex marriage) is $\exp(\alpha + \beta X_i)$
  - For example, the odds ratio that a respondent at the median of age supports same-sex marriage is $\exp(1.3403 - 0.0207 * 54) = 1.2702$
- The odds multiply by $\exp(\beta)$ or $\exp(-0.0207) = 0.979$ for each 1-year decrease in age; that is, there is 2.1% decrease in the odds
  - For example, a 53-yr old, $\pi_i = 0.56$ with odds $\dfrac{(0.56)}{(0.44)} = 1.27$
  - For a 54-yr old, $\pi_i = 0.55$ with odds $\dfrac{(0.55)}{(0.45)} = 1.24$; $\dfrac{1.24}{1.27} = 0.979$

Interpreting the Logit Model

- Using the output from the logit model, we can calculate Wald statistics and confidence intervals for α and β

```
> z_intercept <- (1.340342 - 0) / 0.291319; z_intercept
[1] 4.600943
> c(1.340342-1.96*0.291319, 1.340342+1.96*0.291319)
[1] 0.7693568 1.9113272
>
> z_beta <- (-0.020776 - 0) / 0.005166; z_beta
[1] -4.02168
> c(-0.020776-1.96*0.005166, -0.020776+1.96*0.005166)
[1] -0.03090136 -0.01065064
```

- The corresponding statistics from the likelihood-ratio test

```
> g_statusquo <- 2*(-318.8259 - -327.2174); g_statusquo
[1] 16.783

> confint(logit.field)
Waiting for profiling to be done...
2.5 %      97.5 %
(Intercept)  0.77703500  1.92077448
age         -0.03103374 -0.01075556
```

Interpreting the Logit Model

- We can also calculate confidence intervals for estimated probabilities of support at given values of age

```
> logit_53 <- 1.340342 - 0.020776*53; logit_53
[1] 0.239214

> vcov(logit.field)
(Intercept)              age
(Intercept)  0.084866 -0.0014245
age         -0.001424  0.0000266

> var_logit_53 <- 0.084866 + (53)^2*0.0000266 + 2*53*-0.001424; var_l
[1] 0.008817552
> c(logit_53 + 1.96*(var_logit_53)^.5, logit_53 - 1.96*(var_logit_53)
[1] 0.42326157 0.05516643
> logit_53_lo <- exp(0.05516643) / (1 + exp(0.05516643)); logit_53_lo
[1] 0.5137881
> logit_53_hi <- exp(0.42326157) / (1 + exp(0.42326157)); logit_53_hi
[1] 0.6042635
```

- The probability of supporting same-sex marriage for a 53-yr old is 0.559 with confidence interval (0.513, 0.604)

Interpreting the Logit Model

Expanding the logit model to multiple predictors, including categorical predictors is easy.

- For categorical predictors we need c-1 indicator variables where c is the number of categories

The logit model with age, party and born-again predictors is:

$$logit(\Pr(Y = 1)) = \alpha + \beta_1 IND + \beta_2 GOP + \beta_3 Born + \beta_4 age$$

Interpreting the Logit Model

```
> logit.field2 <- glm(same_sex ~ independent + republican + born_christian
mydata, family = binomial(link=logit))

glm(formula = same_sex ~ independent + republican + born_christian +
age, family = binomial(link = logit), data = mydata)

Deviance Residuals:
Min         1Q   Median       3Q      Max
-2.0052  -1.0155   0.5916   0.8237   2.0532

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)     2.225929   0.376157   5.918 3.27e-09 ***
independent    -0.404662   0.316698  -1.278    0.201
republican     -1.257992   0.230950  -5.447 5.12e-08 ***
born_christian -1.639462   0.232279  -7.058 1.69e-12 ***
age            -0.016335   0.005944  -2.748    0.006 **
---
(Dispersion parameter for binomial family taken to be 1)

Null deviance: 654.43  on 476  degrees of freedom
Residual deviance: 539.65  on 472  degrees of freedom
AIC: 549.65
```

Interpreting the Logit Model

The model assumes that the effects of age are constant across partisanship and born-again status. The implied logits are:

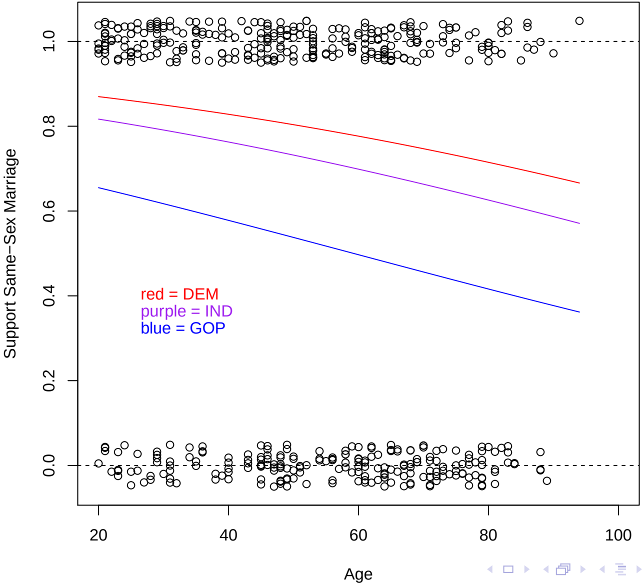
Party	Born Again	Logit
Dem	No	α
Dem	Yes	$\alpha + \beta_3$
Indept	No	$\alpha + \beta_1$
Indept	Yes	$\alpha + \beta_1 + \beta_3$
Rep	No	$\alpha + \beta_2$
Rep	Yes	$\alpha + \beta_2 + \beta_3$

- For example, the logit for a 53-yr old born-again Republican is $logit(\Pr(y = 1|GOP = 1, born = 1)) = 2.225 - 1.258 - 1.639 - 0.016 * 53 = -1.537$
- The odds ratio for such a person is $\exp(-1.537) = 0.215$; $\pi_i = 0.177$

Interpreting the Logit Model

- We can plot the fitted lines for each party assuming Born=0.
  - Since β_1 and β_2 are negative, lines for INDs and GOPs are left of the line for DEMs
- At fixed values of age and born again status, the effect on the logit of changing from DEM (GOP = 0) to GOP (GOP=1) is

  $$[\alpha + \beta_2(1) + \beta_3(born) + \beta_4(age)] \\ - [\alpha + \beta_2(0) + \beta_3(born) + \beta_4(age)] = \beta_2$$


Support Same−Sex Marriage
0.0
0.2
0.4
0.6
0.8
1.0
20
40
60
80
100
Age
red = DEM
purple = IND
blue = GOP

Interpreting the Logit Model

- By itself, the estimate for β_2 (or any categorical predictor) is irrelevant; the estimate only makes sense when compared to the estimate for another category
- $\beta_2 = -1.258$ signifies that the conditional odds ratio between being GOP and support are $\exp(-1.258) = 0.284$; that is, the odds for GOPs are 0.284 times the odds for DEMS.
- Similarly, the odds of support among born again Christians are $\exp(-1.639) = 0.194$ times the odds of those not born again.

Interpreting the Logit Model

- Equivalently, we can take the inverse of 0.284 to get the odds for DEMs – the odds are 3.52 times those of GOPs
- Verify that if you re-estimated the logit model with DEM as a predictor and GOP as the excluded category, you would recover this value for β_2
- Similarly, those who are **not** born again are $\frac{1}{0.194} = 5.15$ times more likely that born again christians to support same-sex marriage

Interpreting the Logit Model

- Treating independents as a separate category does not seem to add much to the model. Can we simply by excluding it or treaty party as continuous?
- The continuous predictor has a small standard error, showing strong evidence of an effect.
- Compare the fit of this simple model against the more complex one with two terms:

```
g_statusquo <- 2*(-269.8239 - -270.1203)
g_statusquo
[1] 0.5928
```

- The simple model seems good.

Interpreting the Logit Model

```
> logit.field3 <- glm(same_sex   party + born_christian + age, mydata,
family = binomial(link=logit))

glm(formula = same_sex   party + born_christian + age,
family = binomial(link = logit), data = mydata)

Deviance Residuals:
Min 1Q Median 3Q Max
-2.0419 -1.0119 0.5732 0.8467 2.0571

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.708971 0.330556 5.170 2.34e-07 ***
party -0.630857 0.115964 -5.440 5.32e-08 ***
born_christian -1.655363 0.231573 -7.148 8.78e-13 ***
age -0.017628 0.005702 -3.091 0.00199 **

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 654.43 on 476 degrees of freedom
Residual deviance: 540.24 on 473 degrees of freedom
AIC: 548.24
```

Interpreting the Logit Model

- We can use similar procedures to test for interactions. Consider whether the effect of age differs for GOP and others.
- We can estimate a model with GOP*Age and compare against a simpler model with only GOP

$$logit(\Pr(Y = 1)) = \alpha + \beta_1(GOP) + \beta_2(born) + \beta_3(age) + \beta_4(GOP * age)$$

- The effect of the interaction is not statistically significant. For reasons we will discuss, this is not determinative.

Interpreting the Logit Model

```
glm(formula = same_sex   republican + born_christian + age + republican *
age,
family = binomial(link = logit), data = mydata)
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.771855 0.409882 4.323 1.54e-05 ***
republican -0.510810 0.668003 -0.765 0.444
born_christian -1.630770 0.231213 -7.053 1.75e-12 ***
age -0.009713 0.007554 -1.286 0.198
republican:age -0.011849 0.011756 -1.008 0.314

Null deviance: 654.43 on 476 degrees of freedom
Residual deviance: 540.24 on 472 degrees of freedom
AIC: 550.24
```

Interpreting the Logit Model

- We can get a better idea about this interaction by plotting the fitted lines for GOPs and others
- The slope of age for GOPs is noticeably steeper

```
> lrtest(logit.field5, logit.field4)

Likelihood ratio test

Model 1: same_sex ~ republican + born_christian + age + republican *

Model 2: same_sex ~ republican + born_christian + age

#Df  LogLik Df  Chisq Pr(>Chisq)

1   5 -270.12

2   4 -270.63 -1 1.0206     0.3124
```

- The LR test does not offer strong evidence of an interaction

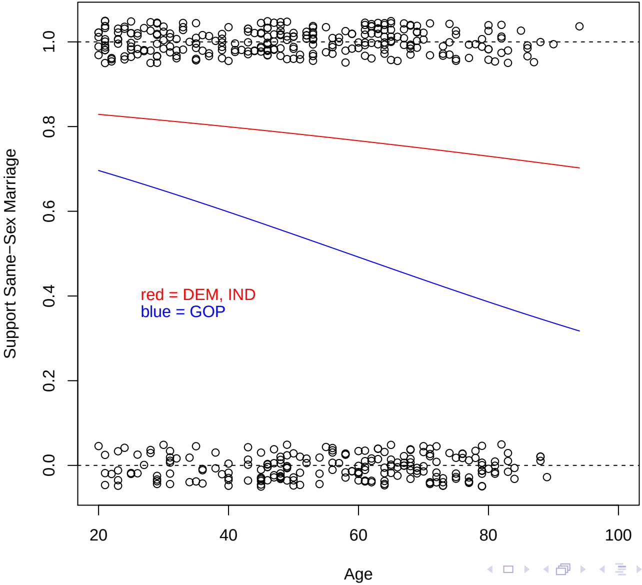

red = DEM, IND
blue = GOP
Support Same−Sex Marriage
0.0
0.2
0.4
0.6
0.8
1.0
20
40
60
80
100
Age

Table of Contents

Additional Quantities of Interest

- One method of assessment is to examine the signs of coefficients and determine their statistical significance.
- This is no more difficult in binary choice than for OLS models
  - A positive (negative) coefficient indicates that increases (decreases) in X lead to increases (decreases) in $\Pr(Y = 1)$
  - The ratio of β to its SE is a z-score that can be used for hypothesis testing.
- In the previous model, we observed that age was not a significant predictor (for Democrats); however, since we have an interaction, the effects of age are conditional.
  - We can use the `glht()` command in R to assess whether Age and GOP* Age are simultaneously 0

Additional Quantities of Interest

```
> summary(glht(logit.field5, linfct = c("age + republican:age = 0")))

Simultaneous Tests for General Linear Hypotheses

Fit: glm(formula = same_sex ~ republican + born_christian + age +

republican * age, family = binomial(link = logit), data = mydata)

Linear Hypotheses:

Estimate Std. Error z value Pr(>|z|)

age + republican:age == 0 -0.021563   0.009008  -2.394   0.0167 *
```

For GOPs, age is significant and the effect is twice as large as for DEMS.

Additional Quantities of Interest

- Simply focusing on the signs and significance of coefficients, however, is a poor method of interpreting and evaluating a binary choice (or any other) model.
- Typically, we are interested in whether and by how much a change in X changes the outcome of interest, i.e., $\Pr(Y_i = 1)$
- Assessing the effects of such changes is more involved than for OLS; the effects of our predictors are linear in the latent variable ξ_i but not in Y_i
- We also know that the net effect of a change in X depends critically on the values of other predictors and parameters in the model.
  - For example, the difference in the probability of support between GOPs and others is $0.61 - 0.80 = -0.19$ for a 37 year old.
  - For a 67-year old, the difference is $0.45 - 0.75 = -0.30$

Additional Quantities of Interest

- Using the formula for calculating probabilities, we can summarize the effect of an indicator variable by calculating estimated probability at its two values:

$$\Delta \Pr(Y_i = 1)_{XA \rightarrow XB} = \frac{exp(\alpha + \beta X_A + ... + \beta X_K)}{1 + exp(\alpha + \beta X_A + ... + \beta X_K)} - \frac{exp(\alpha + \beta X_B + ... + \beta X_K)}{1 + exp(\alpha + \beta X_B + ... + \beta X_K)}$$

- For continuous variables, it makes sense to calculate estimated probabilities at particular values, e.g. lower and upper quartiles, because the min and max reflect outliers.

Variable	β	**se**	**Comparison**	$\Delta \Pr(Y_i = 1)$
Intercept	1.772	0.410		
GOP	-0.511	0.668	$(1, 0)$ *at* X_{med}	0.53 - 0.78 = -0.25
Born Again	-1.631	0.231	$(1, 0)$ *at* X_{med}	0.41 - 0.78 = -037
Age	-0.010	0.008	(UQ, LQ) *at* $GOP = 0$	0.80 - 0.75 = 0.05
Age * GOP	-0.012	0.012	(UQ, LQ) *at* $GOP = 1$	0.61 - 0.45 = 0.15

Additional Quantities of Interest

- In addition to estimating point predictions for relevant profiles, we can use the `predict()` command to obtain predicted probabilities and standard errors.
- Process: we generate new dataset of hypothetical cases (i.e. profiles). Start with a baseline of non-GOP non-born again respondents whose ages range from 20 to 94. The hypothetical cases should be plausible (possible).

```
basedata = data.frame(age=x, republican=rep(0, length(x)), born_chris
length(x)))
```

- Next, calculate predicted probabilities and standard errors using the `glm` object `logit.field5`.

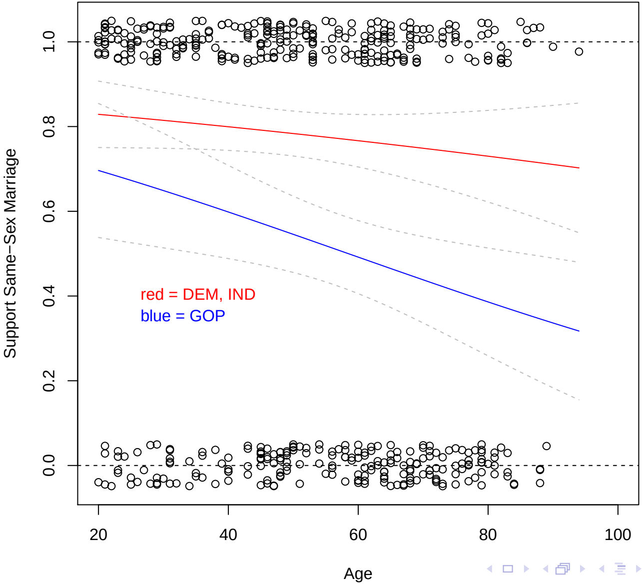
```
> base_field5 <- predict(logit.field5, newdata=basedata, type="respon
se.fit=TRUE)
```

- Overlay scatterplot with predictions and confidence intervals for GOP and others.

Additional Quantities of Interest

```
> lines(x, base_field5$fit, lty = 1, col = "red")
> lines(x, (base_field5$fit + 2*base_field5$se), lty = 2, col = "gray")
> lines(x, (base_field5$fit - 2*base_field5$se), lty = 2, col = "gray")
> lines(x, gop_field5$fit, lty = 1, col = "blue")
> lines(x, (gop_field5$fit + 2*gop_field5$se), lty = 2, col = "gray")
> lines(x, (gop_field5$fit - 2*gop_field5$se), lty = 2, col = "gray")
```

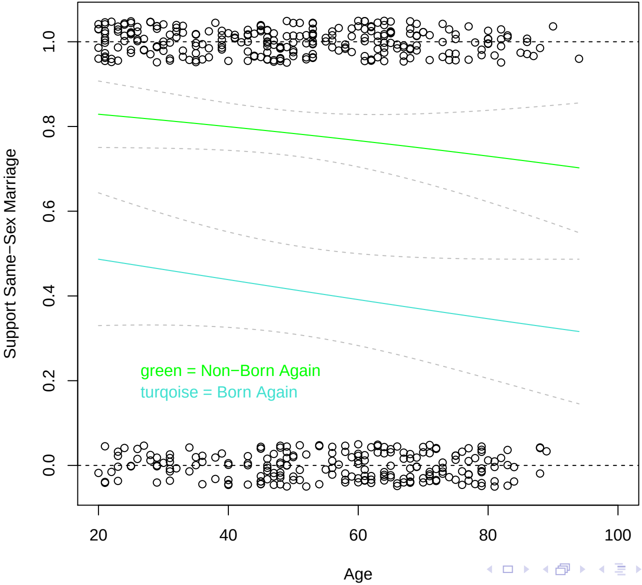
Quantities of Interest



Additional Quantities of Interest

- There is some overlap but clear separation for 40+ respondents.
- We can plot some other comparisons, e.g. born again Christians v. others; no overlap in these lines.

```
> plot(jitter(same_sex, .25)   age, mydata, xlab="Age", xlim=c(20,100),
+ ylab="Support Same-Sex Marriage")
> abline(h = 1, lty = 2); abline(h = 0, lty = 2)
> > lines(x, base_field5$fit, lty = 1, col = "green")
> lines(x, (base_field5$fit + 2*base_field5$se), lty = 2, col = "gray")
> lines(x, (base_field5$fit - 2*base_field5$se), lty = 2, col = "gray")
> > lines(x, born_field5$fit, lty = 1, col = "turquoise")
> lines(x, (born_field5$fit + 2*born_field5$se), lty = 2, col = "gray")
> lines(x, (born_field5$fit - 2*born_field5$se), lty = 2, col = "gray")
```



Support Same−Sex Marriage
Age
0.0
0.2
0.4
0.6
0.8
1.0
20
40
60
80
100
green = Non−Born Again
turqoise = Born Again

Additional Quantities of Interest

- One advantage of a logit model over a probit is the ability to substantively interpret the effects of predictors in terms of odds. Recall:

$$\log \frac{\pi_i}{1-\pi_i} = \frac{\dfrac{exp(\alpha+\beta X_i)}{1+exp(\alpha+\beta X_i)}}{1-\dfrac{exp(\alpha+\beta X_i)}{1+exp(\alpha+\beta X_i)}} = \alpha+\beta X_i$$

- The odds ratio is $exp(\alpha+\beta X_i)$. For a coefficient of an indicator variable, the odds ratio is $exp\beta$ holding other factors constant.
  - For example, the odds that a born again Christian supports same sex marriage is $exp(-1.631) = 0.196$ times that of others. Taking the inverse gives us the odds ratio for others vs. born again Christians, 5.109. The odd that someone who is <u>not</u> born again supports same sex marriage is <u>five times</u> that of a born again Christian.
- For an indicator variable, the interpretation is simple and intuitive.

Additional Quantities of Interest

- Since odds-ratios are proportional between adjacent categories, we can calculate the (approximate) percentage change in the odds given changes in X:

$$\%\Delta \frac{\pi_i}{1-\pi_i} = \frac{exp(\beta x) - exp(\beta x')}{exp \beta x'} * 100$$

- For born again Christians we have:

$$= \frac{exp(-1.631 * 1) - exp(-1.631 * 0)}{exp{-1.631 * 0}} * 100 = \left[\frac{0.961 - 1}{1}\right] * 100 = -80.42$$

- That means the odds are 80% lower.

Additional Quantities of Interest

- For DEMS, increasing age from 37 to 67 results in a 25.92% decrease in the odds.

$$\frac{exp(0.010 * 67) - exp(0.010 * 37)}{exp(0.010 * 37)} * 100 =$$
$$\left[\frac{0.512 - 0.691}{0.691}\right] * 100 =$$
$$-25.92$$

- For GOPs, increasing age from 37 to 67 results in a 48.31% decrease in the odds.

$$\frac{exp(0.010 * 67 - 0.012 * 67) - exp(0.010 * 37 - 0.012 * 37)}{exp(0.010 * 37 - 0.012 * 37)} * 100 =$$
$$\left[\frac{0.229 - 0.443}{0.443}\right] * 100 =$$
$$-48.31$$

Table of Contents

title

content...