Logit analysis and logistic regression

Lecture 12

Multivariate statistics

Psychology 613 – Spring 2022

Outline

- Problems with binary DVs
- Odds and odds ratios
- Logistic response function
- Maximum likelihood
- Logistic regression in R
- Interpreting results
- Likelihood ratio tests

Problems with binary DVs

1. Non-normal errors: When Y is dichotomous, the residual, e, can only have two values:

   We know that: $e = Y - X*b$

   When $y = 1$: $e = 1 - X*b$

   When $y = 0$: $e = - X*b$

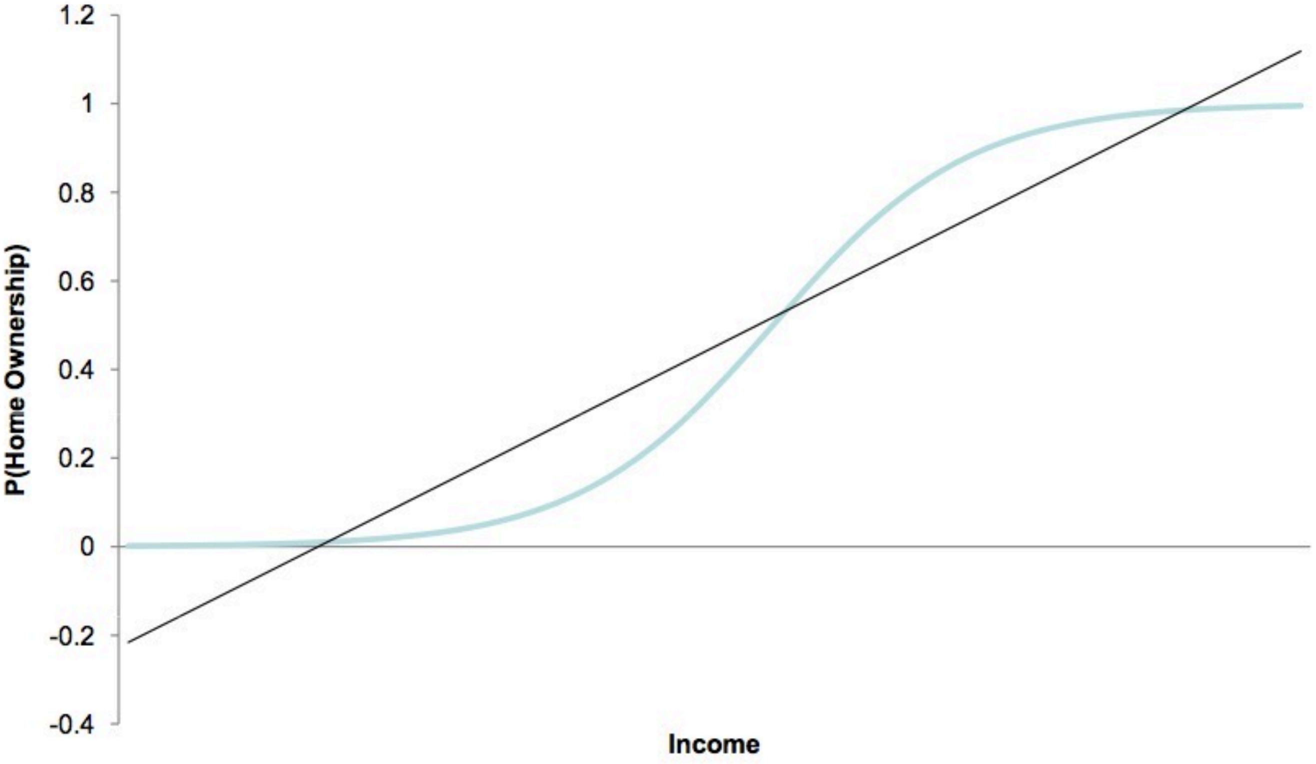
2. Non-constant error variance: When Y is dichotomous, it can be shown that the variance of the errors is:

   $\sigma^2 = Xb*(1-Xb)$,

Which involves X, so error is not homoscedastic

Problems with binary DVs

3. Phenomena are likely non-linear
4. No constraint on X*b (e.g., linear function)



Odds and odds ratios

E.g., subjects are told that a communiqué is EITHER from an economist or an administrator, then asked to judge whether the message is biased or not…

N	Economist	Administrator
Unbiased	38	20
Biased	12	30

Odds that a subject in the Economist condition thinks the message is unbiased: 38/12 = 3.1667

Note: Odds are not probabilities

Odds that a subject in the Administrator condition thinks the message is unbiased: 20/30=.66667

Asymmetrical odds

Odds for unbiased:

Economist: 3.1666 Administrator: 0.6667

Odds for biased:

Economist: .3158 Administrator: 1.5

Each category can be summarized by two different odds (one <1, the other one >1, e.g., 2/3 & 3/2)

Odds are *asymmetric about 1*:

Distance between 0 -> 1 and the distance between 1 -> +∞ express the same relation, though the two regions are vastly different.

Correcting for asymmetry

Take the *natural log* of the odds

Economist: $\ln(3.1667) = 1.1527$

$\ln(.31579) = -1.1527$

Administrator: $\ln(0.6667) = -.4055$

$\ln(1.5) = .4055$

Note on logs: if $y = \log_a(x)$, then $a^y=x$

ln = natural log, which is base *e*

The natural log of the odds is called the *logit*

Odds ratios

Odds ratio (unbiased) = 3.1667 / 0.667 = 4.75

Measures the degree of (nonlinear) association between binary X and Y variables

Null value (H_0) = 1

Interpretation:

> *The odds of perceiving the message as unbiased are 4.75 greater when the message is attributed to an economist rather than to an administrator.*

Odds ratios

Odds the Dodgers win when Kershaw pitches:

24/8 = 3

Odds the Dodgers win when ~Kershaw pitches:

59/71 = 0.83

Odds ratio = 3 / 0.83 = 3.61

The Dodgers are 3.61 times more likely to win when Kershaw pitches compared to when he does not.

(~$5.25 million per odds-ratio increase)

Asymmetric odds ratios

Just like *odds*, *odds ratios* are also asymmetric:

OR-unbiased = 4.75	OR-win = 3.61
OR-biased = .21	OR-lose = .28

And again, taking the natural log will fix this:

ln(4.75) = 1.56	ln(3.61) = 1.28
ln(.210) = -1.56	ln(.28) = -1.28

Odds ratios as probabilities

	Economist	Administrator
Unbiased	38	20
Biased	12	30
Marginal sum	50	50

P(unbiased | Economist) = 38/50 = .76

P(biased | Economist) = 1-P = 1-.76 = .24

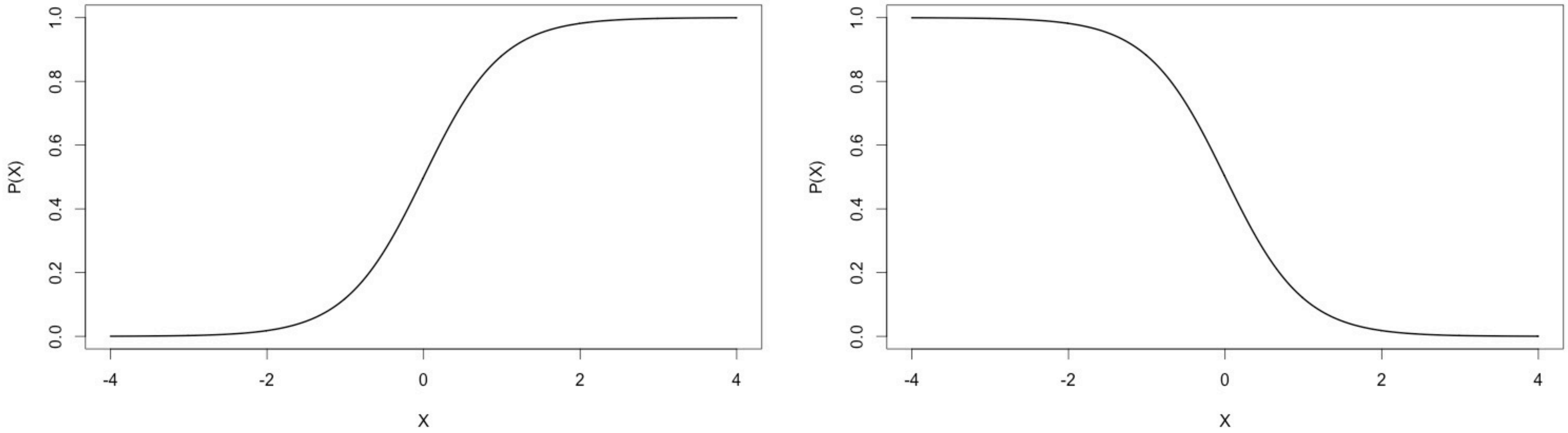
Odds are P / (1-P) = .76/.24 = 3.16667

P(unbiased | Administrator) = 20/50 = .4

P(biased | Administrator) = 1-P = 1-.4 = .6

Odds = .4/.6 = .6667

Logistic response functions



Often appropriate for binary DVs

Monotonic curves (=sigmoidal), asymptote at zero and one, approximately linear between P=.2 and .8

Formula: $P = \frac{e^{b_0 + b_1 X}}{1 + e^{b_0 + b_1 X}}$

Linearizing a logistic curve

If $$P = \frac{e^{b0+b1X}}{1+e^{b0+b1X}}$$

Linearized logistic function

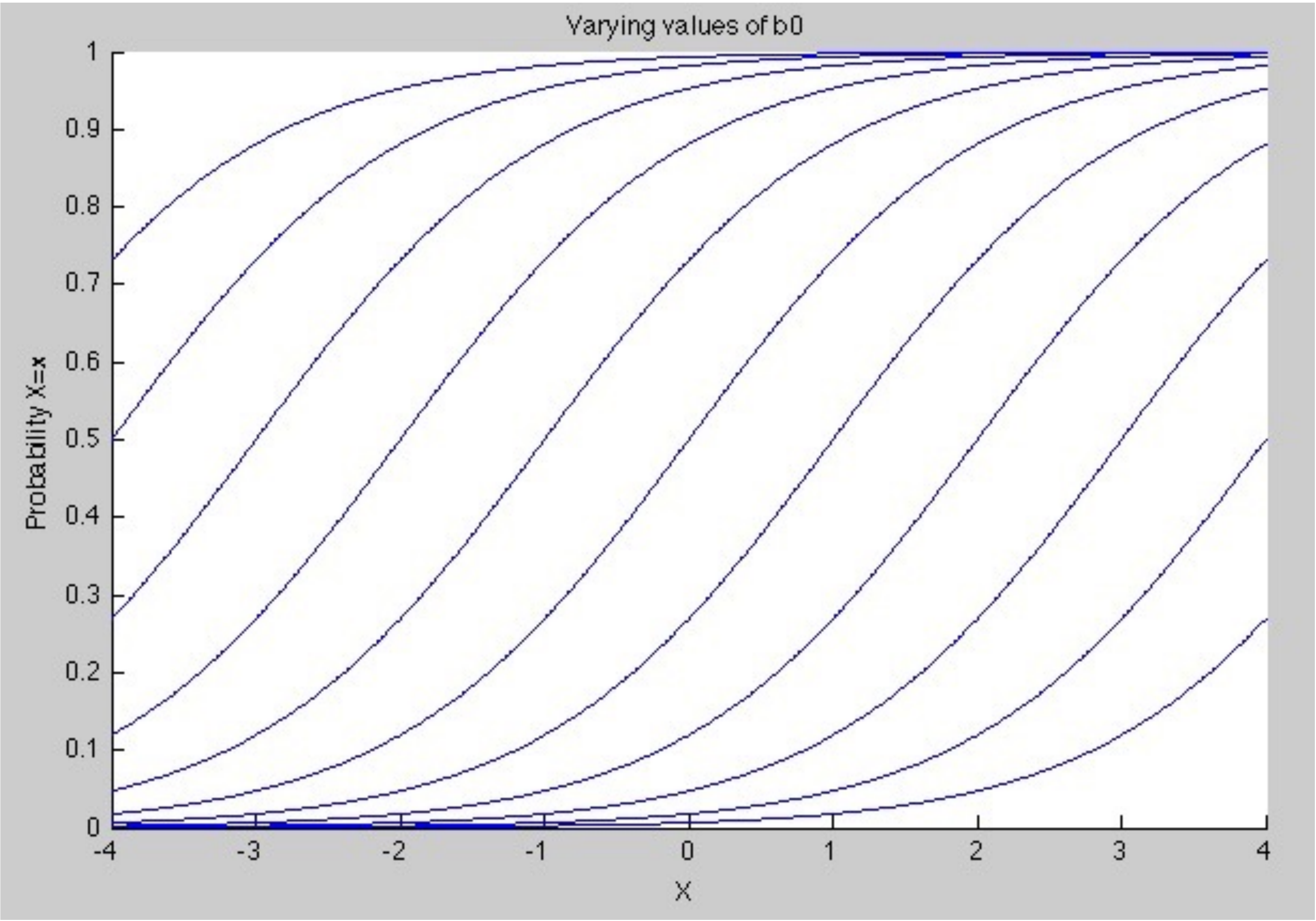
$$P = \frac{e^{b_0 + b_1 X}}{1 + e^{b_0 + b_1 X}} \Longrightarrow \ln\left(\frac{p}{1-p}\right) = b_0 + b_1 X$$

P	0.1	0.3	0.5	0.7	0.9	Range
Odds =P / (1-P)	0.1/0.9 =.11	0.3/0.7 =.43	0.5/0.5 =1	0.7/0.3 =2.33	0.9/0.1 =9	0 -> +∞
Logit =ln(odds)	-2.21	-.84	0	.84	2.21	-∞ -> +∞

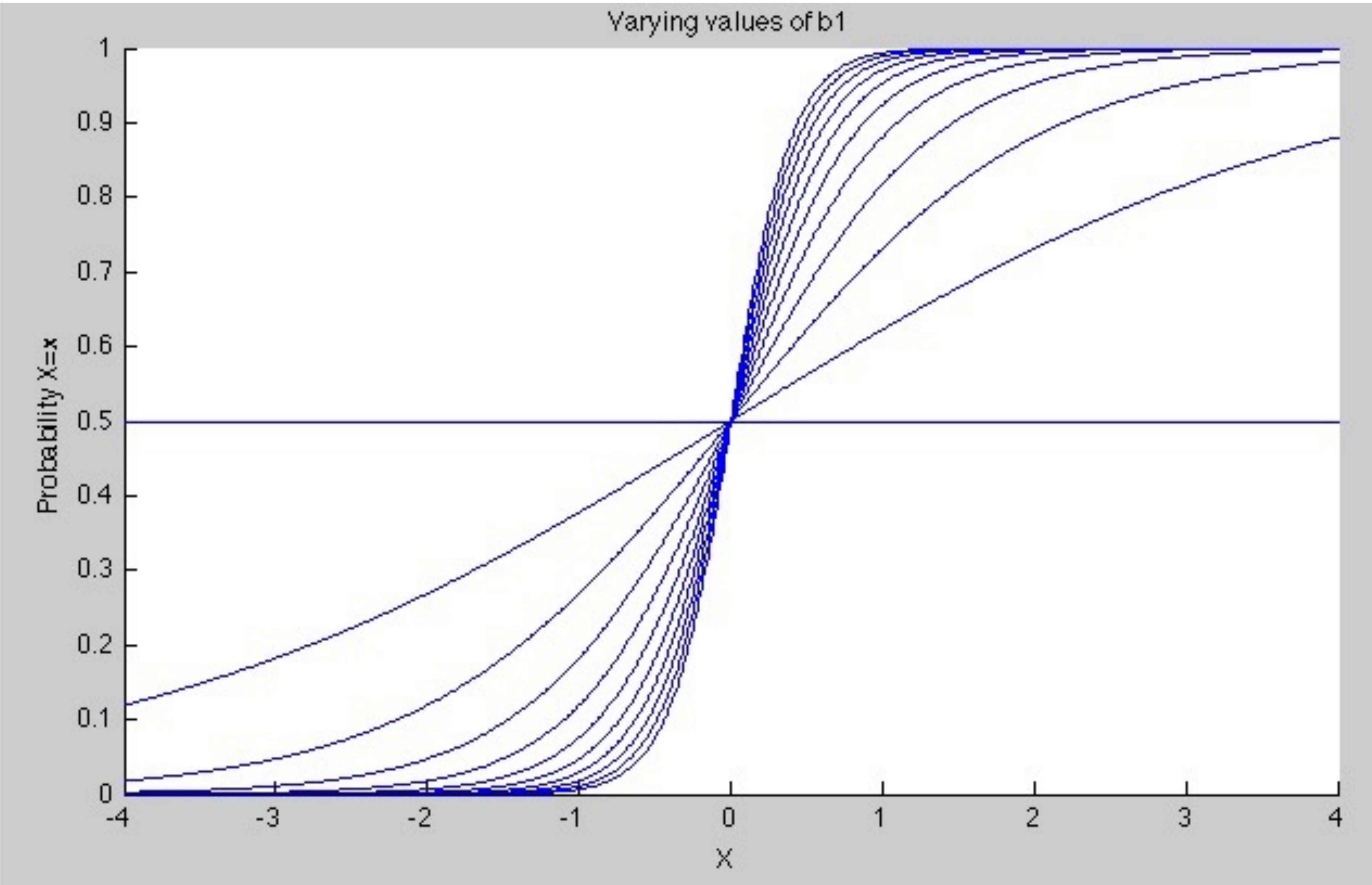
Addresses Problem 4 (constraint on X*b), but still heteroscedastic (i.e., unequal variances across range)

→ Estimate with *maximum likelihood*
(or weighted least squares)

Changes in intercept (b_0)

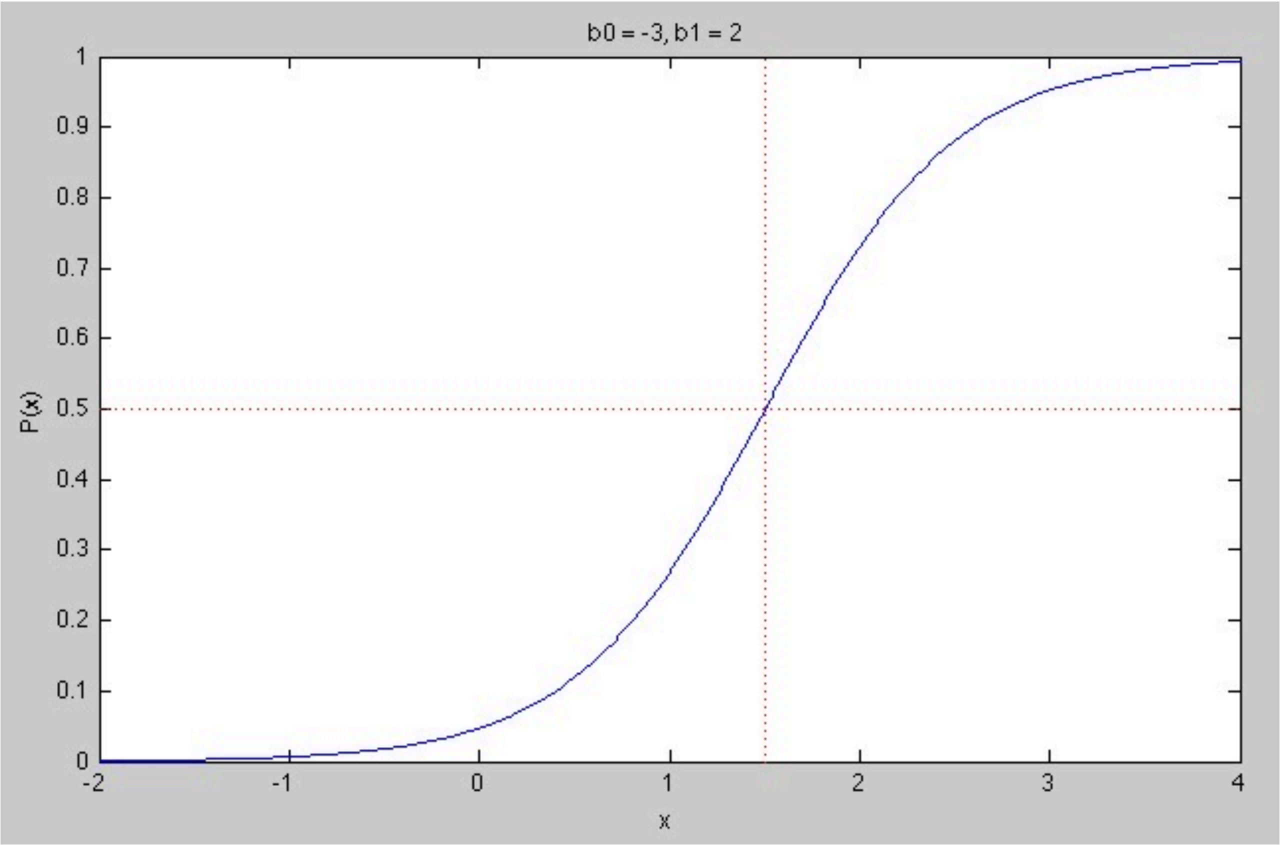


Changes in slope (b_1)



Threshold: P(0.5) when $-b_0/b_1$

The point on the curve when the probability is 0.5 is: $-b_0 / b_1$



Interpretation

Odds = P / (1-P) = e^{b0+b1X}

Log-odds = logit = $b_0 + b_1X$

b_0 = The expected value of the logit when X=0

b_1 = The "logit difference" = the amount the log-odds change for a one-unit change in X

e^{b0} = Odds of base rate (when X=0)

e^{b1} = Change in odds ratio with a 1-unit change in X

Administrator example

Odds of “biased” for Econon:	0.316
Odds of “biased” for Admin:	1.5

Odds ratio = 1.5 / .315 = 4.76

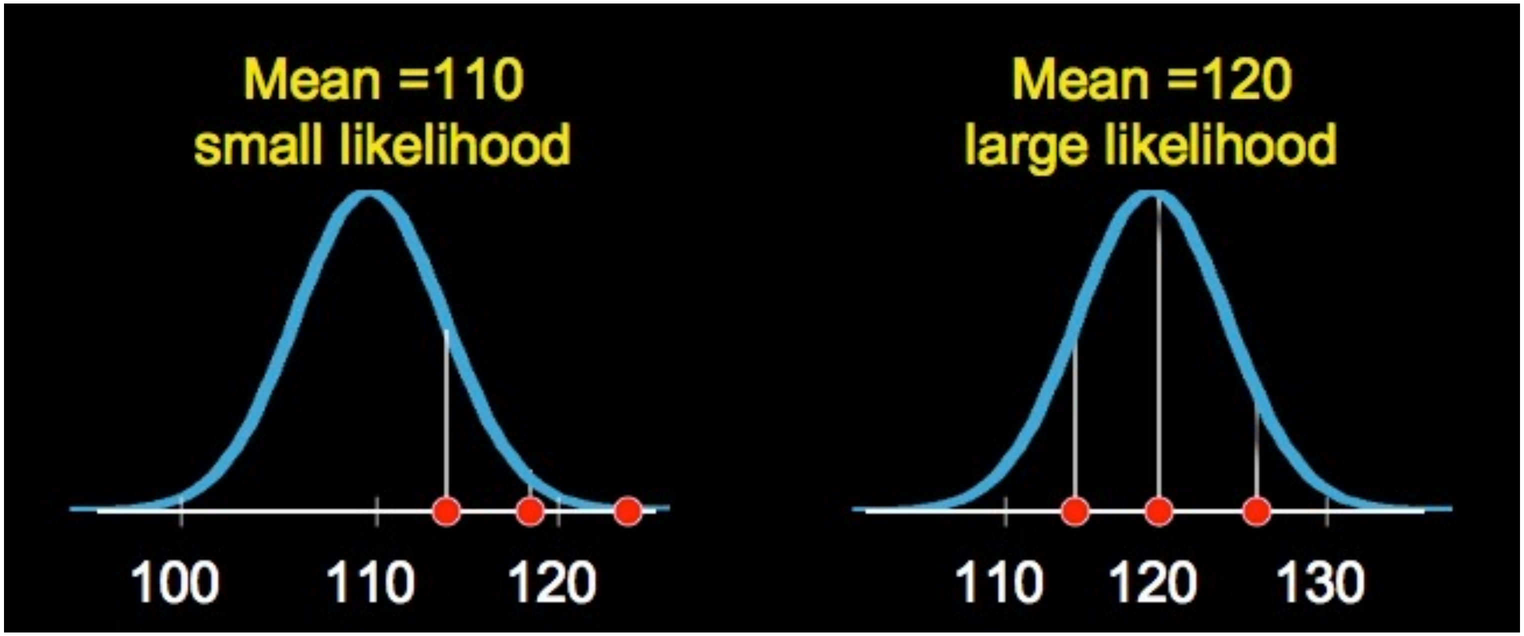
$$P = \frac{\exp(-1.15 + 1.56(cond))}{1 + \exp(-1.15 + 1.56(cond))}$$

$e^{-1.15}$ = 0.316 = Odds of “bias” for X=0 (econ)

$e^{1.56}$ = 4.76 = Change in odds (multiplicative) for a one-unit change in X (i.e., from Econ to Admin)

Maximum likelihood (refresher)

Instead of minimizing error (which is what Ordinary Least Squares—OLS—does), maximize the likelihood that we would have gotten our data from some given parameters: P(Y | β)



Maximum likelihood (refresher)

Find estimates of parameter values that maximize the likelihood (the “joint probability” of all data points) of obtaining the data we have.

(Similar to how we minimize the errors in OLS.)

However, the system of equation is insoluble for logistic regression, so use an iterative guided trial-and-error process.

Maximum likelihood (refresher)

ML-based estimation yields a useful value:

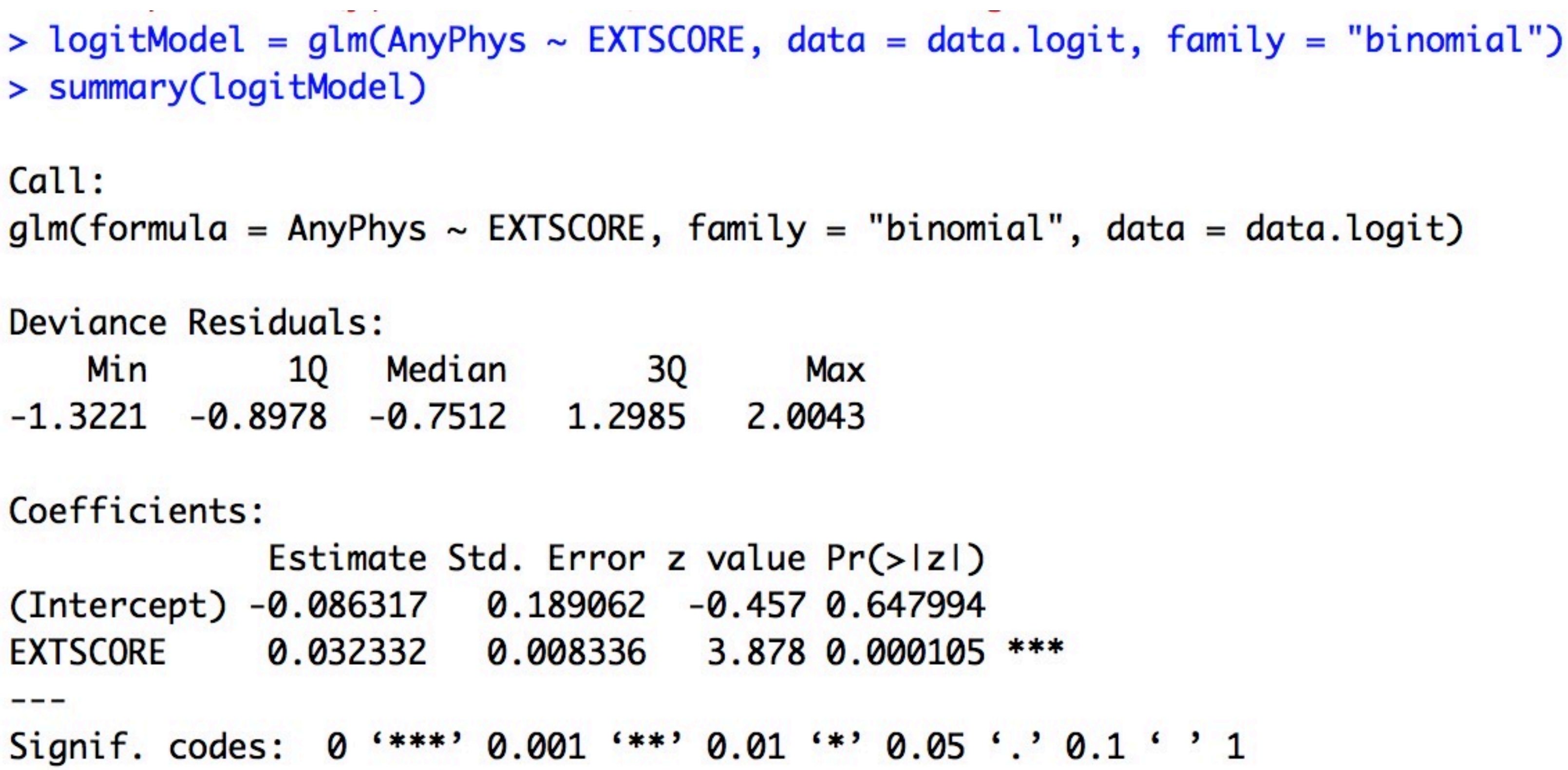
Deviance = -2 * log(likelihood) = -2LL

Positive number which indicates the *lack* of fit

Deviances closer to 0 are better.

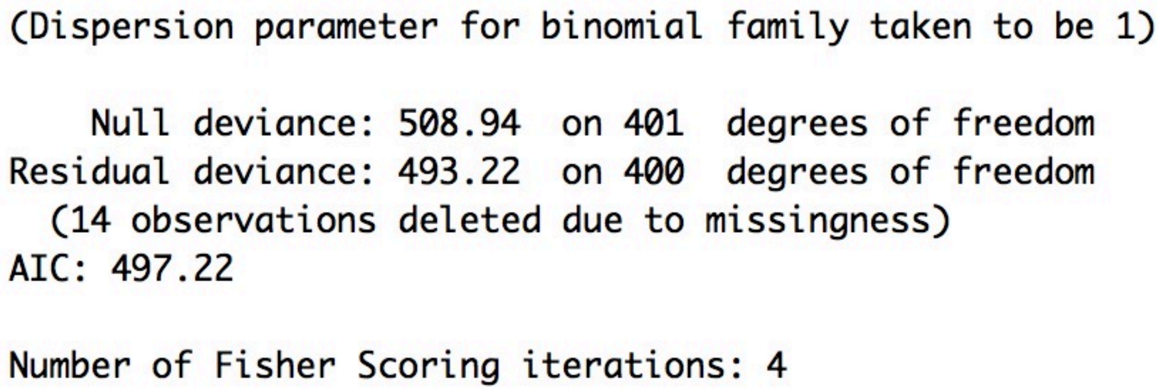
Not interpretable in isolation; useful in comparison between nested models

Logistic Regression in R



```
> logitModel = glm(AnyPhys ~ EXTSCORE, data = data.logit, family = "binomial")
> summary(logitModel)

Call:
glm(formula = AnyPhys ~ EXTSCORE, family = "binomial", data = data.logit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3221  -0.8978  -0.7512   1.2985   2.0043

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.086317   0.189062  -0.457 0.647994
EXTSCORE     0.032332   0.008336   3.878 0.000105 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Logistic Regression in R



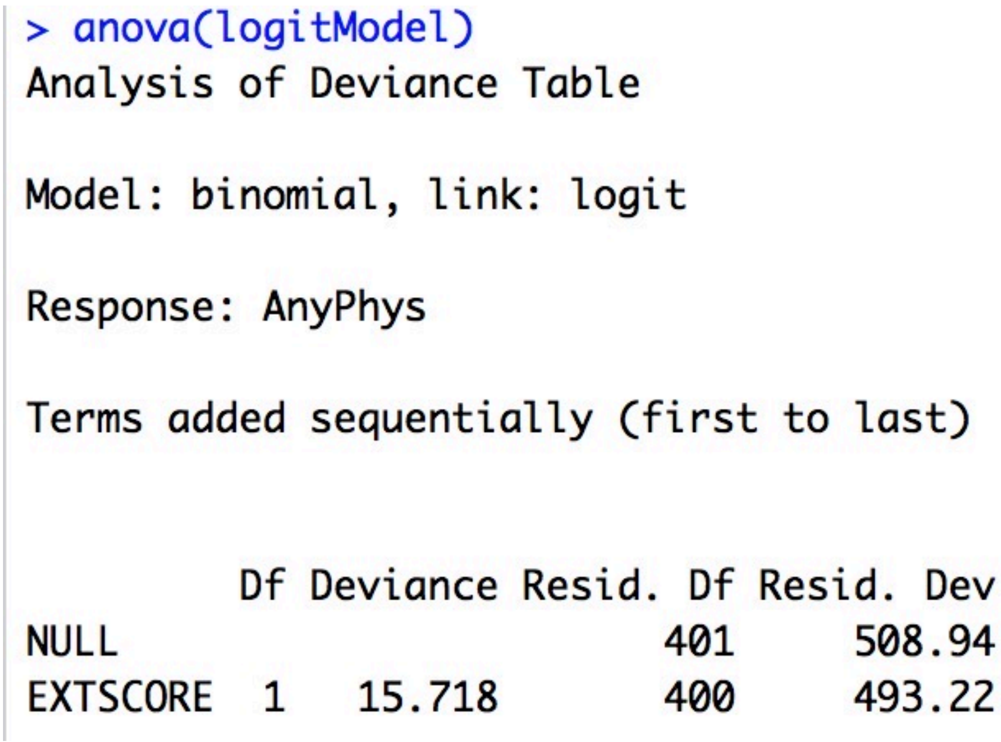
```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 508.94  on 401  degrees of freedom
Residual deviance: 493.22  on 400  degrees of freedom
  (14 observations deleted due to missingness)
AIC: 497.22

Number of Fisher Scoring iterations: 4
```

The bottom half of the output tells you *deviances* for the null model (i.e., with no predictors) and the model you fit (“residual”).

Remember that these are -2*log-likelihoods.

Their difference is chi-squared distributed…

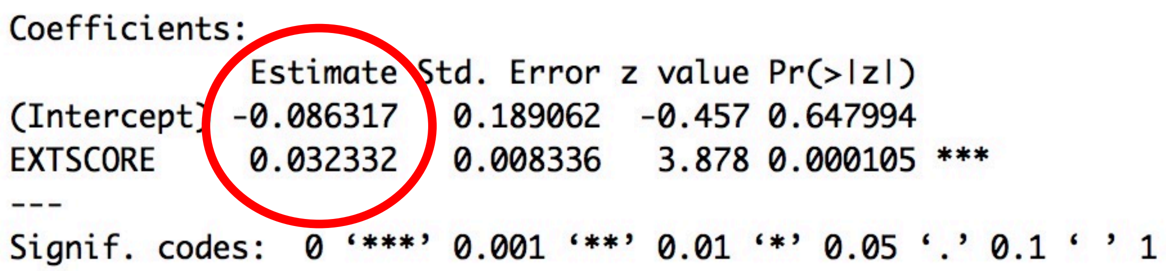
Logistic regression in R



```
> anova(logitModel)
Analysis of Deviance Table

Model: binomial, link: logit

Response: AnyPhys

Terms added sequentially (first to last)


         Df Deviance Resid. Df Resid. Dev
NULL                       401     508.94
EXTSCORE  1   15.718       400     493.22
```

Chi-square test of change in -2*Log-likelihood (vs. constant-only model)

Logistic regression in R



```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.086317   0.189062  -0.457 0.647994
EXTSCORE     0.032332   0.008336   3.878 0.000105 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Formula for the **log odds of physical aggression:**

$$\ln\left(\frac{P}{1-P}\right) = -.086 + .032 * EXTSCORE$$

Formula for the **odds of physical aggression:**

$$\frac{P}{1-P} = e^{-.086+.032*EXTSCORE} = 0.917 * e^{.032*EXTSCORE}$$

Interpretation

Interpret the *b*s as changes in logits – the expected change in logit units (log-odds) for a one-unit increase in predictor:

E.g., for each one-unit change in *extscore*, the log odds of physical aggression increase by .032.

Ratio of b-to-SE could arguably be a t or Z

Wald chi-square is a squared value of that ratio, has an associated p-value

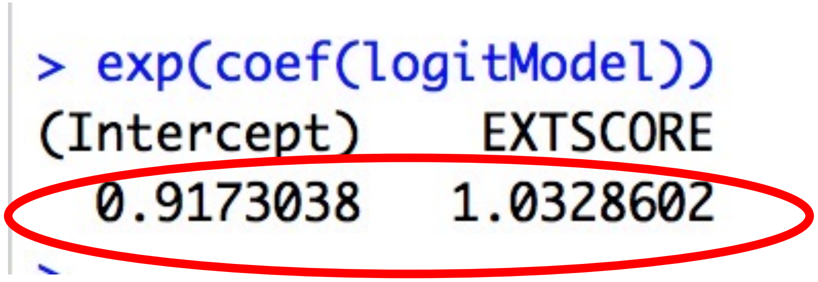
Interpretation

b is expected change in log-odds for a one-unit change in X…

…e^{b} is expected change in odds, P/(1-P), for a one-unit change in X

E.g., $e^{.032} = 1.033$ → For each one-unit increase in EXTSCORE, expected odds of violence are 1.033 times greater (i.e., the odds increase by 3.3%)

Interpretation



```
> exp(coef(logitModel))
(Intercept)    EXTSCORE
  0.9173038   1.0328602
```

Formula for the **odds of physical aggression:**

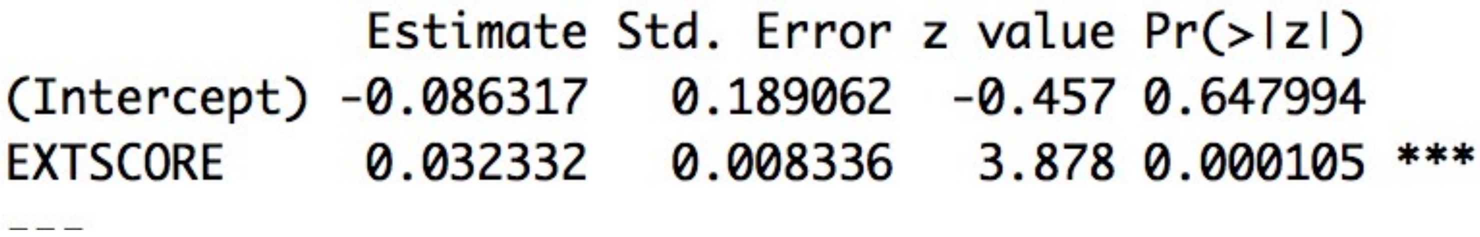
$$\frac{P}{1-P} = e^{-.086+.032*EXTSCORE} = 0.917 * e^{.032*EXTSCORE}$$

Interpretation of 1.033 (=e^b) is the odds multiply by a factor of 1.033 for each 1-unit increase in the IV.

Interpretation

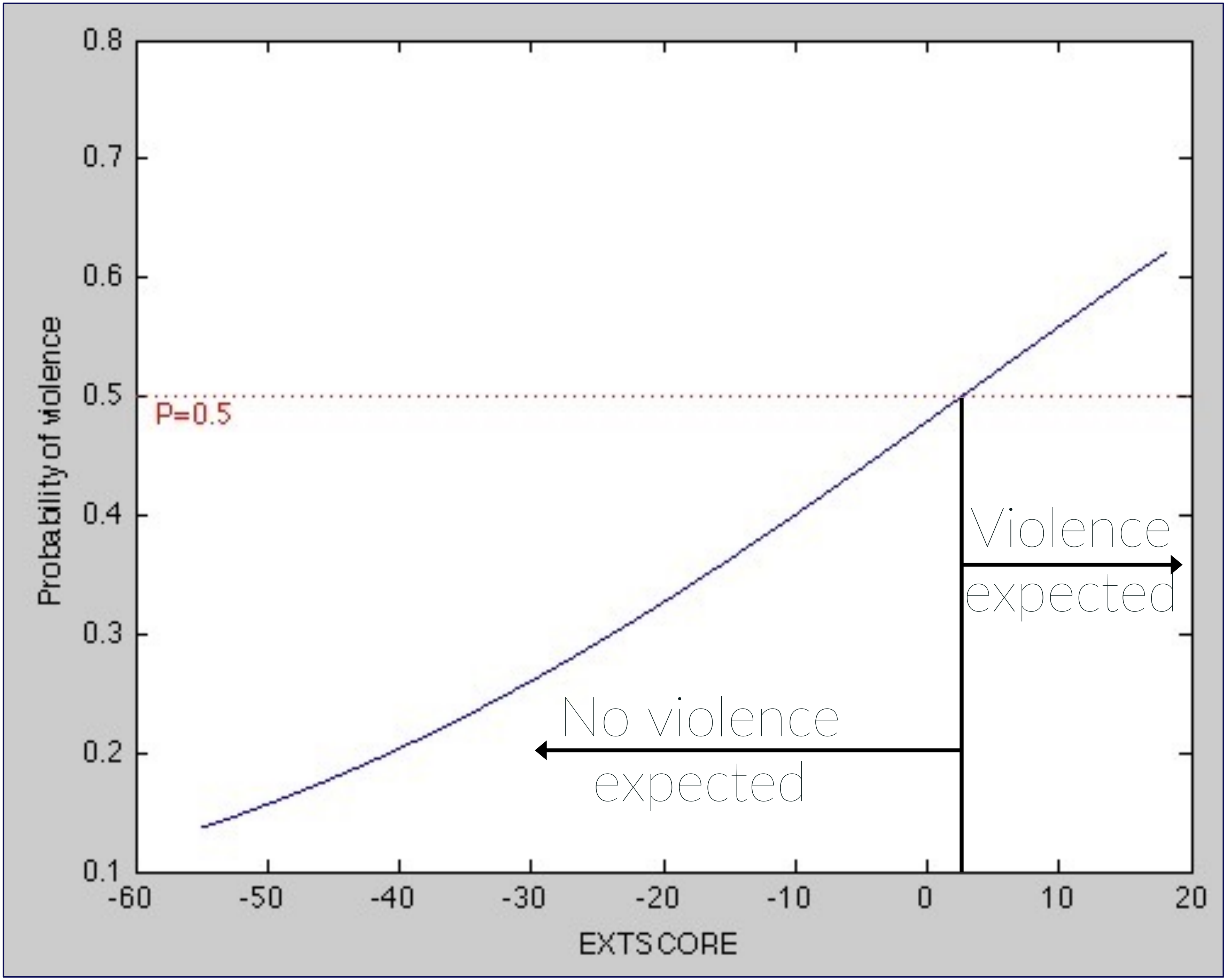
In terms of *probabilities*, $$P = \frac{e^{b_0 + b_1 X}}{1 + e^{b_0 + b_1 X}}$$

…so we can generate a plot based on this eqn.

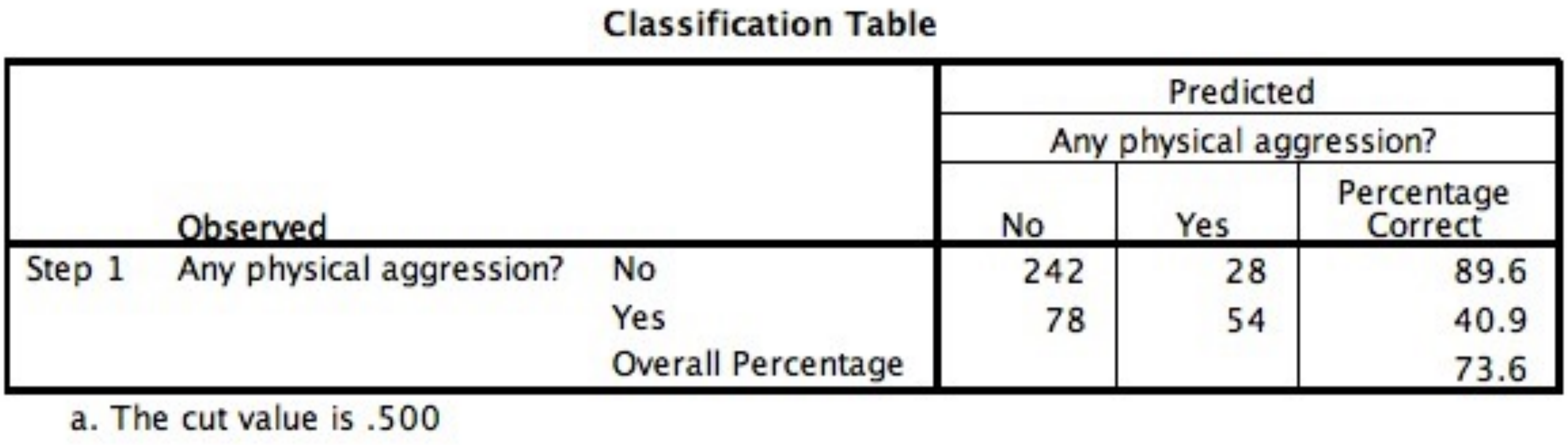


```
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.086317   0.189062  -0.457 0.647994
EXTSCORE     0.032332   0.008336   3.878 0.000105 ***
---
```

```
> extscore <- -55:.01:18;
> p_violence <- exp(-.086+.032*extscore) / (1+exp(-
.086+.032*extscore))
> plot(extscore,p_violence)
```


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Probability of violence
P=0.5
-60
-50
-40
-30
-20
-10
0
10
20
EXTSCORE
Violence expected
No violence expected

Interpretation



Classification Table

			Predicted		
			Any physical aggression?		
	Observed		No	Yes	Percentage Correct
Step 1	Any physical aggression?	No	242	28	89.6
		Yes	78	54	40.9
		Overall Percentage			73.6

a. The cut value is .500

Prediction based on EXTSCORE > cutoff of P>0.5, about at EXTSCORE=3.

Interpretation

At a specific EXTSCORE, e.g., 10:

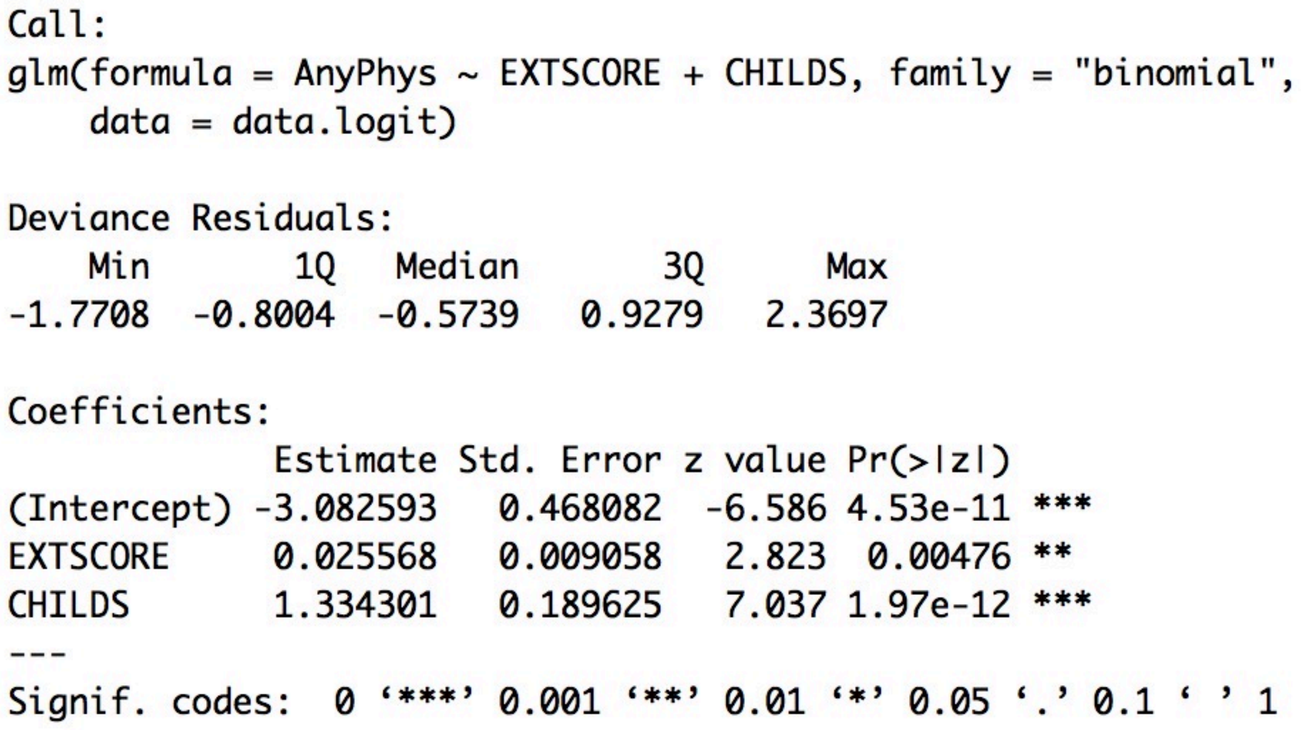
$$P = \frac{e^{b_0 + b_1 X}}{1 + e^{b_0 + b_1 X}} = \frac{e^{-.086 + .032 * 10}}{1 + e^{-.086 + .032 * 10}} = \frac{e^{.234}}{1 + e^{.234}} = 0.558$$

The probability of Y=1 (i.e., violence) is 56%

And predicted odds = .558 / (1-.558) = 1.26

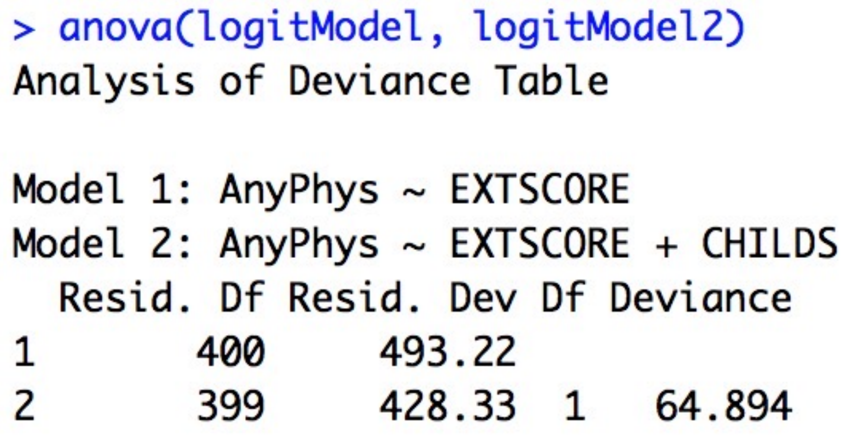
Hierarchical logistic regression

Add a second predictor: CHILDS (childish behavior)



```
Call:
glm(formula = AnyPhys ~ EXTSCORE + CHILDS, family = "binomial",
    data = data.logit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7708  -0.8004  -0.5739   0.9279   2.3697

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.082593   0.468082  -6.586 4.53e-11 ***
EXTSCORE     0.025568   0.009058   2.823  0.00476 **
CHILDS       1.334301   0.189625   7.037 1.97e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

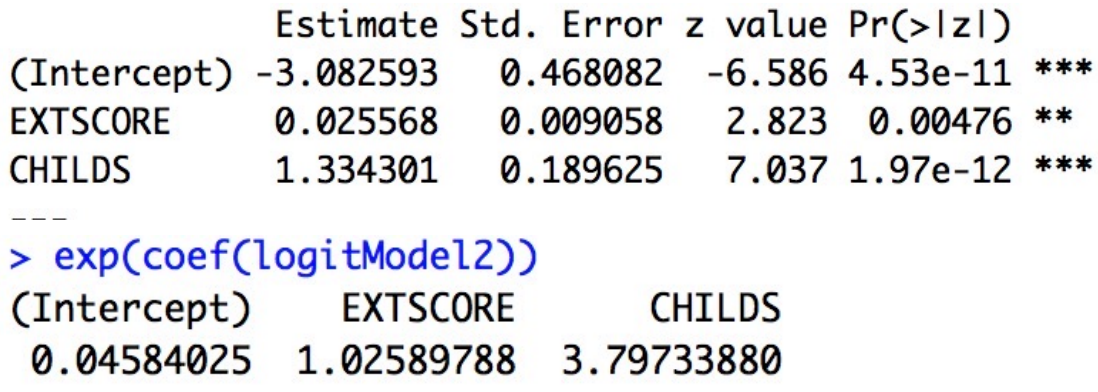
Hierarchical logistic regression



```
> anova(logitModel, logitModel2)
Analysis of Deviance Table

Model 1: AnyPhys ~ EXTSCORE
Model 2: AnyPhys ~ EXTSCORE + CHILDS
  Resid. Df Resid. Dev Df Deviance
1       400     493.22
2       399     428.33  1   64.894
```

Chi-square for this step is the change from the previous one.

Model refers to both predictors together.



```
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.082593   0.468082  -6.586 4.53e-11 ***
EXTSCORE     0.025568   0.009058   2.823  0.00476 **
CHILDS       1.334301   0.189625   7.037 1.97e-12 ***
---
> exp(coef(logitModel2))
(Intercept)    EXTSCORE      CHILDS
 0.04584025  1.02589788  3.79733880
```

Likelihood ratio test of a predictor

Effect of adding a variable is conducted using the LR (likelihood ratio), which is based on the difference between deviance values:

Reduced model:	Deviance =	493.22
Full model:	Deviance =	428.33
	Change =	64.89

Deviance change is distributed as chi-square with df = difference in # of parameters

Interactions in logistic regression

Compute the (cent.) interaction vectors as usual:

$$y = b_0 + b_1X_1 + b_2X_2 + b_3X_1X_2$$

Examine LR test of deviance change associated with the interaction to determine significance (also requires fitting: $y=b_0 + b_1X_1+b_2X_2$)

If so, follow up with tests of simple effects.

Models for multi-category DVs

3+ ordered categories: *Ordinal* logistic regression

- e.g.: None, Some, Most All
- Predicts probabilities of being at or below a particular level
- R: model <- polr(DV ~ IV1 + IV2 ..., data = dat, Hess = TRUE)
- polr() is from the **MASS** package

3+ non-ordered categories: *Multinomial* regression

- Predicts probabilities of being in a particular category relative to a reference category
- R: model <- multinom(DV ~ IV1 + IV2 ..., data = dat)
- multinom() is from the **nnet** package