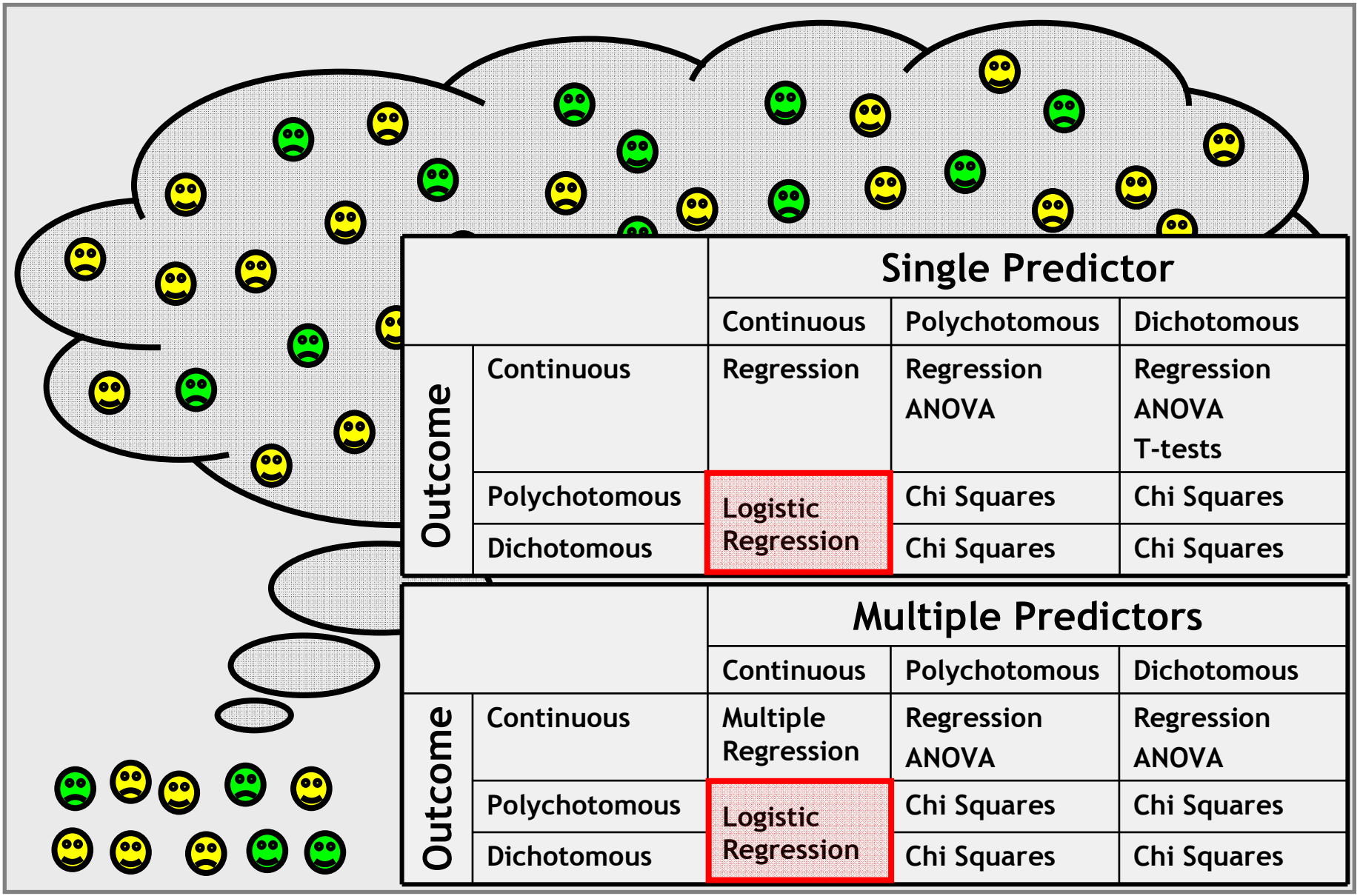
Appendix B: Road Map (Schematic)



		Single Predictor		
		Continuous	Polychotomous	Dichotomous
Outcome	Continuous	Regression	Regression ANOVA	Regression ANOVA T-tests
	Polychotomous	Logistic Regression	Chi Squares	Chi Squares
	Dichotomous		Chi Squares	Chi Squares

		Multiple Predictors		
		Continuous	Polychotomous	Dichotomous
Outcome	Continuous	Multiple Regression	Regression ANOVA	Regression ANOVA
	Polychotomous	Logistic Regression	Chi Squares	Chi Squares
	Dichotomous		Chi Squares	Chi Squares

Appendix B: Logistic Regression

Appendix B Post Hole:

Interpret a fitted simple logistic regression model, noting the statistical significance of the relationship, the direction of the relationship, and the magnitude of the relationship by comparing two fitted probabilities (or fitted percentages).

Appendix B Technical Memo and School Board Memo:

Conduct a logistic regression analysis with a dichotomous outcome and a continuous predictor. Generate and discuss a plot of prototypical fitted probabilities (or fitted percentages).

Appendix B: Research Question

Theory: In the 9th grade, the more mathematically "advanced" students are placed (i.e., tracked) into "college-prep math" courses, while the other students will take lower level "business math" courses in the 9th grade. Due to course-prerequisite structures and perhaps other factors, the placement decision has a huge impact on the academic future of 9th graders. The decision should be fair, but is the decision biased against academic minorities? If there were no bias, students with the same MCAS Math scores should have the same probability of college-prep placement, regardless of race/ethnicity. (Note that this is only a theory. Differences by race/ethnicity do not definitively prove bias on the part of the school. For example, students have some choice, and different students may choose differently.)

Research Question: Controlling for *8th Grade Math MCAS* scores, are Black and Hispanic students less likely to be placed in 9th-grade college preparation math courses?

Data Set: Anonymized data from 334 9th graders (class of 2009) from Riverside High School.

Variables:

Outcomes: 9th grade math placement (***PLACEMENT***) where a value of 1 indicates placement into 9h grade college-prep math, and 0 indicates non-college-prep.

Predictors: Scaled score on the 8th grade Math MCAS exam (***MCAS***)

A set of race/ethnicity indicators: ***ASIAN***, ***BLACK***, ***LATINO*** and ***MIXED***, where ***WHITE*** is reserved as a reference category.

Two Algebraically Equivalent Models:

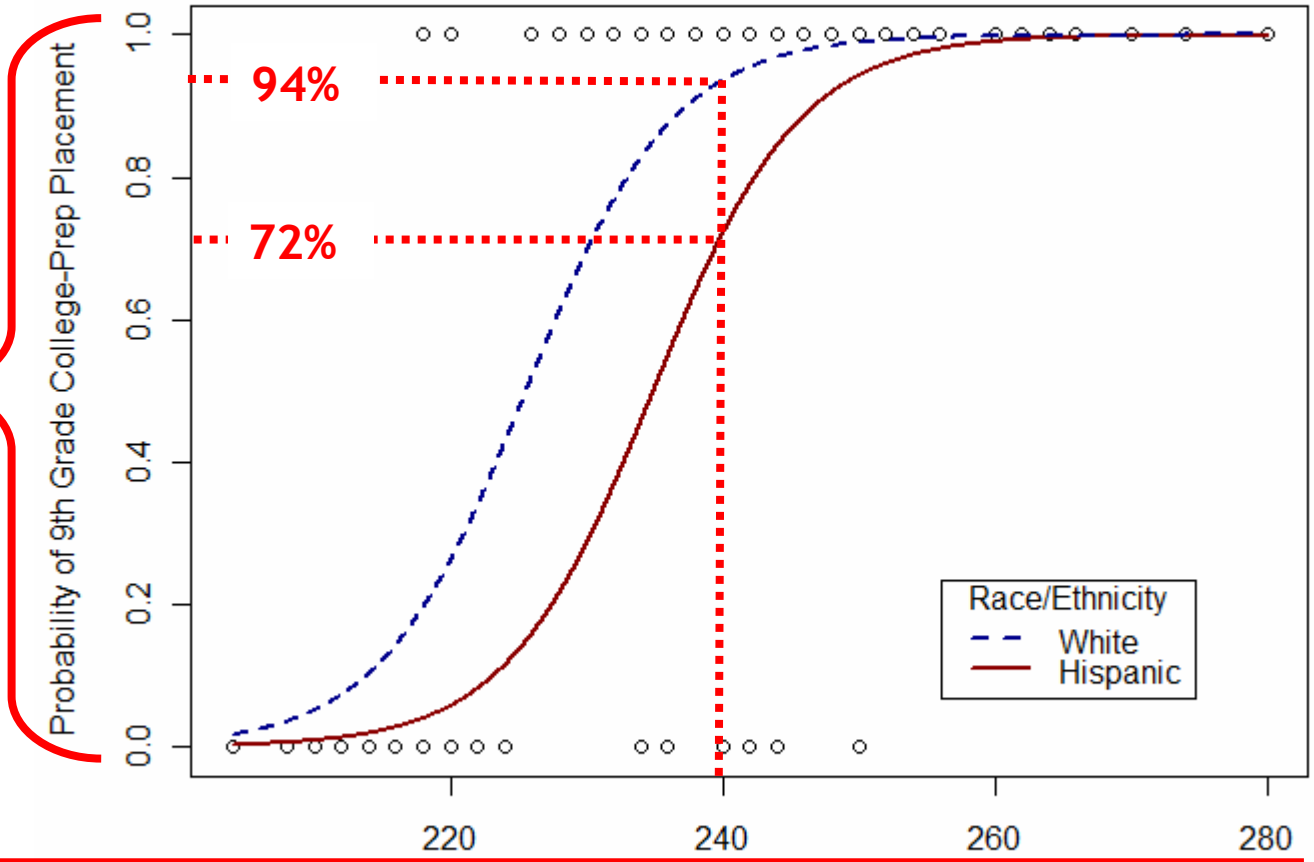
$$\log_e\left(\frac{p(PLACEMENT=1)}{1-p(PLACEMENT=1)}\right)=\beta_0+\beta_1 MCAS+\beta_2 ASIAN+\beta_3 BLACK+\beta_4 HISPANIC+\beta_5 MIXED$$

$$p(PLACEMENT=1)=\frac{1}{1+e^{-(\beta_0+\beta_1 MCAS+\beta_2 ASIAN+\beta_3 BLACK+\beta_4 HISPANIC+\beta_5 MIXED)}}$$

Beginning at the End: The Answer

Figure B.1. Fitted probabilities of placement in 9th grade college-prep math by race/ethnicity and Math MCAS scaled scores from the 8th grade (n= 334).

On the Y axis, we have fitted probabilities. By definition, a probability is between 0 and 1 (inclusive). If you like, you can multiply the probabilities by 100 to get percentage chances. For example, a probability of .90 is a 90% chance. The probabilities are "fitted" because we estimated them based on our fitted model.



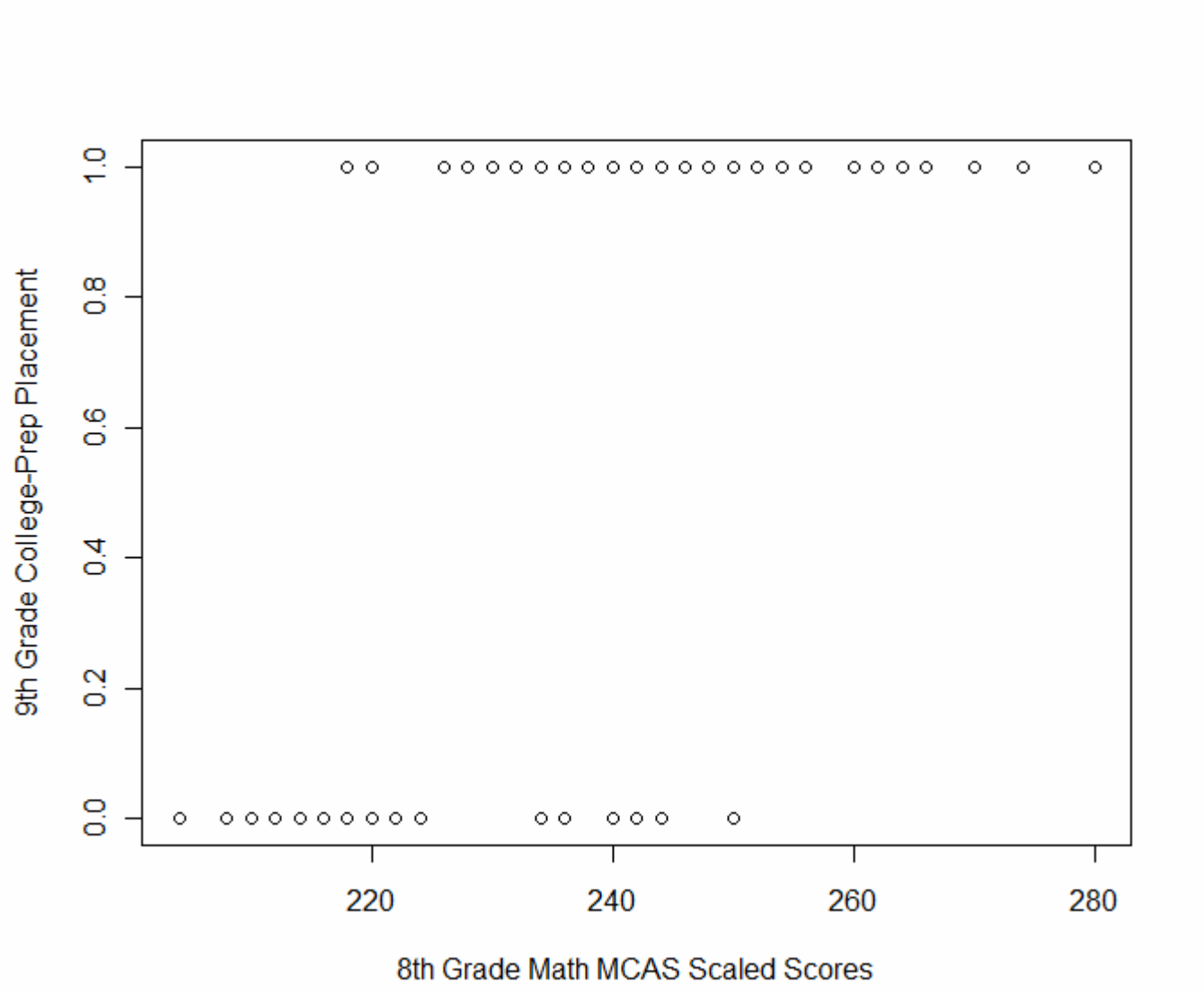
On the X axis, we have 8th grade Math MCAS scores. We see that ***higher*** scores are associated with ***higher*** probabilities of placement in 8th grade algebra (for all racial/ethnic groups). To see the racial/ethnic differences, let's focus on students who scored a 240 on the 8th grade Math MCAS:

Among student who scored 240, White students are 94% likely to be placed in college-prep math, and Hispanic students are 72% likely to be placed in college-prep math.

Beginning at the Beginning: The Data Set

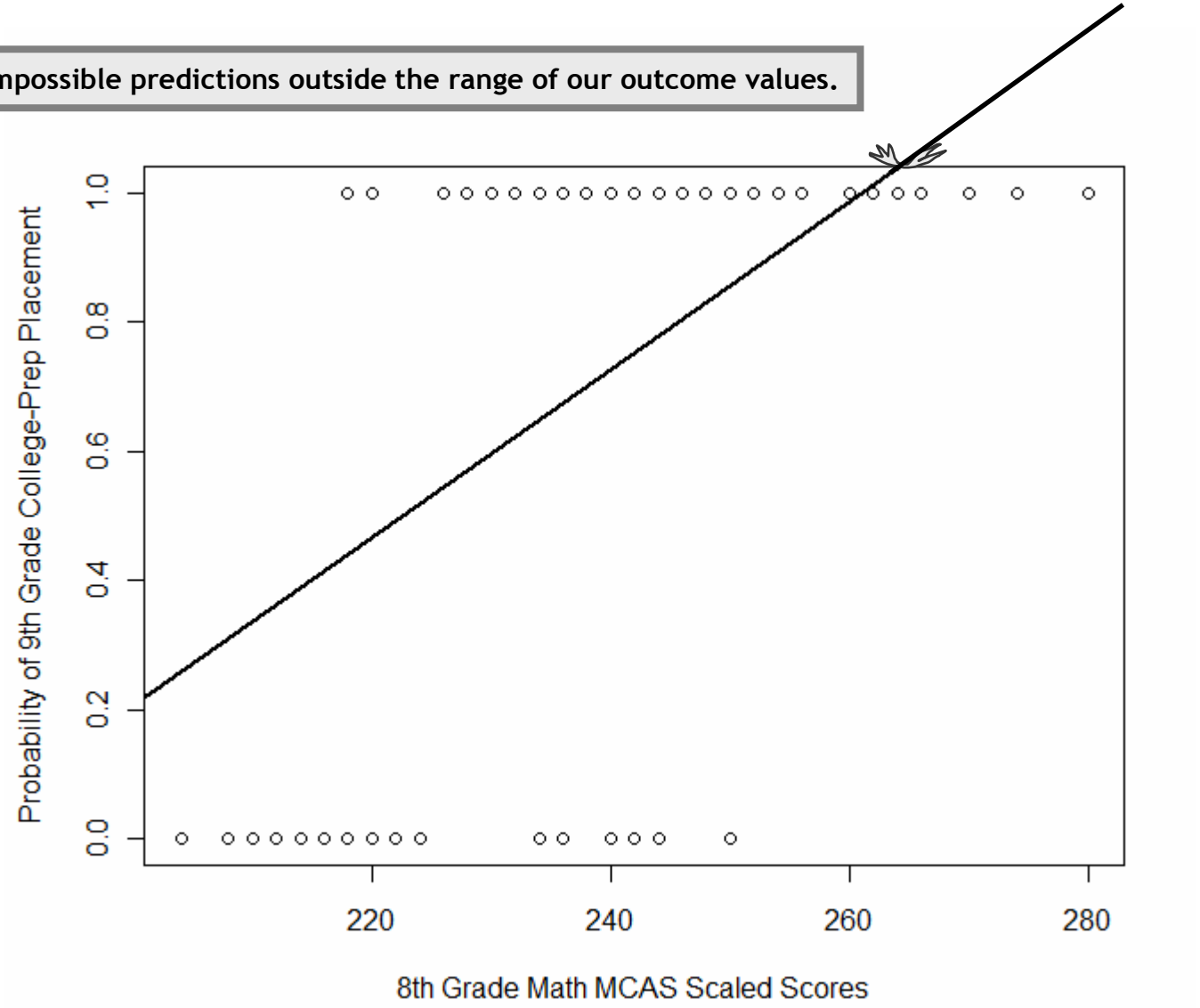
placement	mcas	race	asian	black	hispanic	mixed	white
1	260	White	0	0	0	0	1
1	254	Hispanic	0	0	1	0	0
1	252	White	0	0	0	0	1
1	254	White	0	0	0	0	1
1	260	White	0	0	0	0	1
0	218	White	0	0	0	0	1
1	256	White	0	0	0	0	1
1	266	Hispanic	0	0	1	0	0
0	222	Asian	1	0	0	0	0
1	242	White	0	0	0	0	1
0	216	Black	0	1	0	0	0
1	260	White	0	0	0	0	1
1	250	Asian	1	0	0	0	0
1	262	Hispanic	0	0	1	0	0
0	218	White	0	0	0	0	1
1	262	Asian	1	0	0	0	0
1	256	Asian	1	0	0	0	0
1	266	White	0	0	0	0	1
1	264	Asian	1	0	0	0	0
1	256	White	0	0	0	0	1
1	254	White	0	0	0	0	1

Bivariate Scatterplot of *PLACEMENT* vs. *MCAS*



A Simple Linear Regression of *CollegePrep* vs. *MCAS*

Problem #1: Impossible predictions outside the range of our outcome values.



Problem #2: The distributions of the residuals (conditional on the predictor) are neither normal nor homoscedastic.

The Mean For Dichotomies: Reprise From Unit 3, Post Hole 3

Let's say our sample (N=12) is 1/3 boys:

MALE: 1 0 0 1 0 0 1 0 1 0 0 0
Please show your work:

Raw	Mean	Mean Deviation	Square Mean Deviation	Z-Score
1	.33	.67	.449	1.37
0	.33	-.33	.109	.67
0	.33	-.33	.109	.67
1	.33	.67	.449	1.37
0	.33	-.33	.109	.67
0	.33	-.33	.109	.67
1	.33	.67	.449	1.37
0	.33	-.33	.109	.67
1	.33	.67	.449	1.37
0	.33	-.33	.109	.67
0	.33	-.33	.109	.67
0	.33	-.33	.109	.67

Please note the **mean** of the raw distribution:	**.5**
Please note the **sum of squared mean deviations**:	**2.67**
Please note the **variance** of the raw distribution:	**.24**
Please note the **standard deviation** of the raw distribution:	**.49**

Why is the mean the proportion of ones?

In our sample of 12, we have 4 students who are boys. Each of those 4 students gets a 1 for *MALE*, and everybody else (each of the girls) gets a 0. When we add up the values for *MALE*, we are actually counting the number of boys. (That is the beauty of 0/1 coding for dichotomies, also unfortunately known as "dummy coding.") When we take the average, we are dividing the total number of boys in our sample by the total number of students in our sample, thus we get the proportion of boys in our sample: .33, or 1/3, or 33%.

The trick to naming dummy variables:

You can name variables anything you want, but there is an especially helpful naming convention for dummy variables. You should name the variable after the thing that gets a 1, so that the 1 stands for "Yes" and the 0 stands for "No."

Good Practice:
MALE, a variable where 1 = male and 0 = female
FEMALE, a variable where 1 = female and 0 = male

Bad Practice:
GENDER, a variable where 1 = male and 0 = female
GENDER, a variable where 1 = male and 2 = female

Linear Probability Model

Good Old General Linear Model:

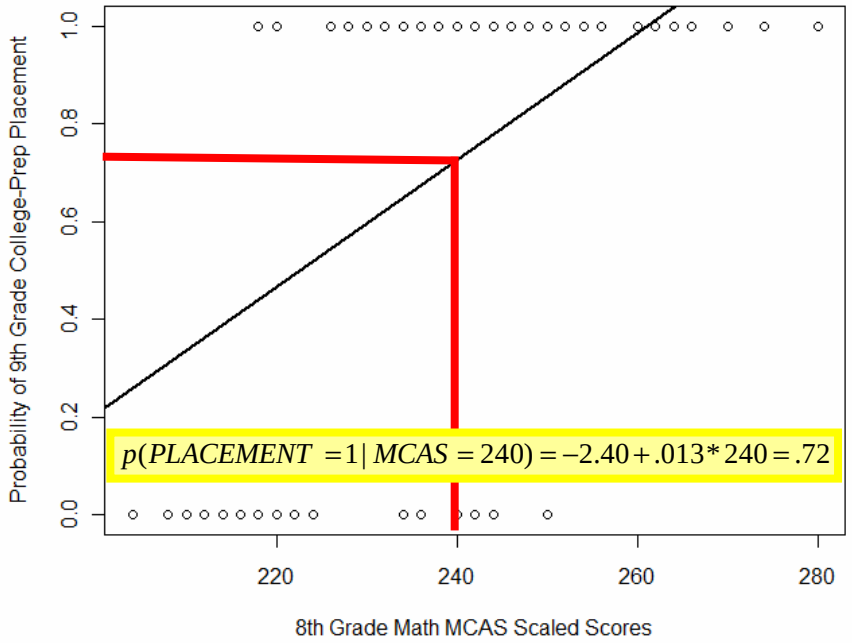
$$PLACEMENT = \beta_0 + \beta_1 MCAS + \varepsilon$$

```
> linear.probability.model <- lm(placement~mcas)
> summary(linear.probability.model)

Call:
lm(formula = placement ~ mcas)

Residuals:
     Min       1Q   Median       3Q      Max
-0.85643 -0.11672  0.01342  0.19563  0.56003

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.3971866  0.2123293  -11.29   <2e-16 ***
mcas         0.0130145  0.0008458   15.39   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 0.2641 on 332 degrees of freedom
Multiple R-squared: 0.4163,     Adjusted R-squared: 0.4145
F-statistic: 236.7 on 1 and 332 DF,  p-value: < 2.2e-16
```

Linear Probability Model:

$$p(PLACEMENT = 1) = \beta_0 + \beta_1 MCAS$$

$$p(PLACEMENT = 1) = -2.40 + .013 MCAS$$

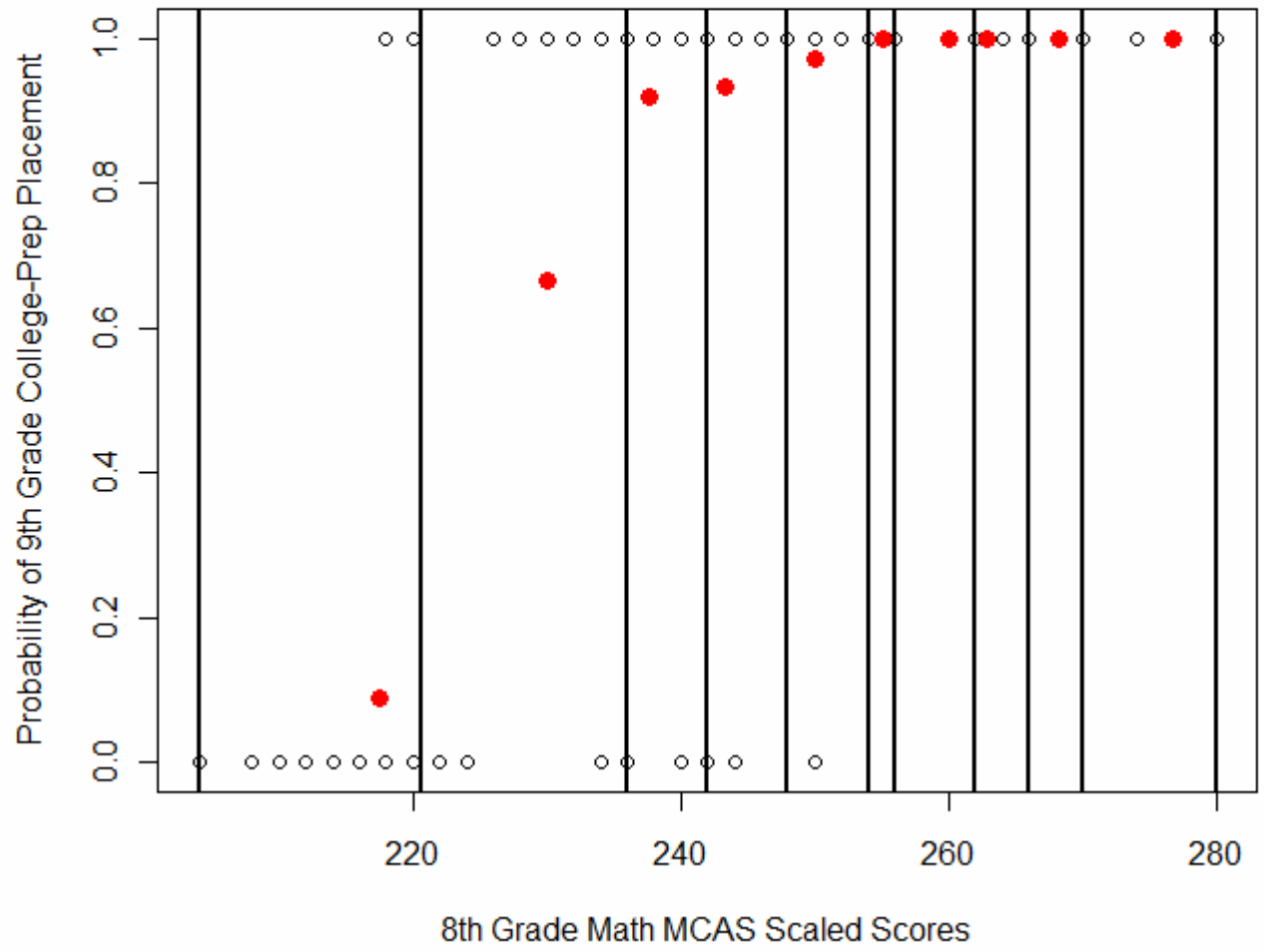


Throughout the course, we start by trying to predict individuals, but then in light of the difficulty, we end up adopting the more tractable goal of predicting averages. Since the average of a dichotomy is a probability, we can predict probabilities with our linear model. With a linear probability model, we recognize we are predicting probabilities, and we give up right away on predicting individuals by jettisoning the error term.

Exploratory Data Analysis

Create ten equal-sized bins of *MCAS* scores.

Within each bin, place a point at the mean *PLACEMENT* and the mean *MCAS* score.



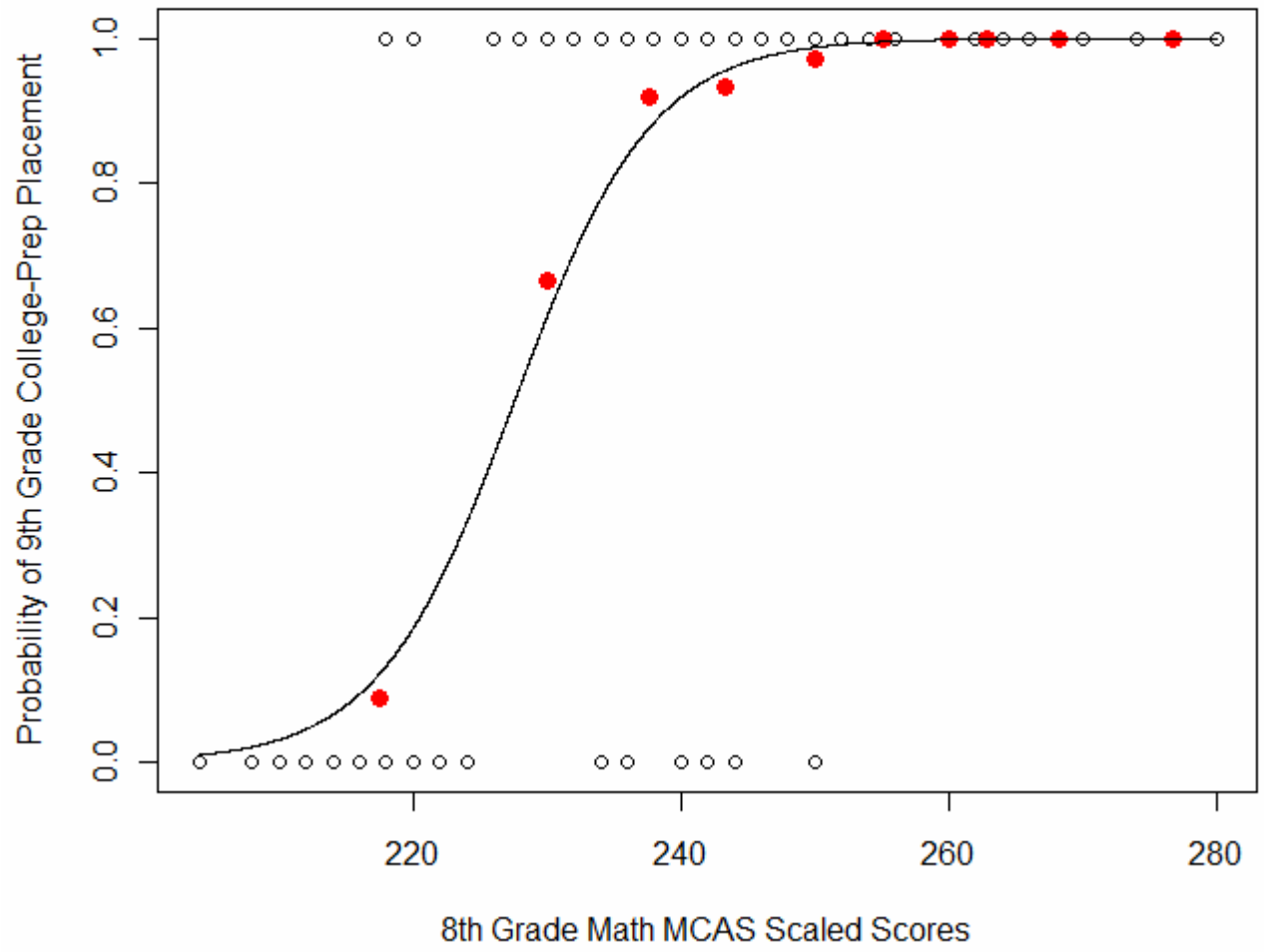
Notice that the relationship is non-linear!

R Script For Logistic Exploratory Data Analysis

```
# a function for logistic exploratory work
logistic.explore <- function(outcome, predictor, my.y.axis.label='', my.x.axis.label='') {
  # pairwise deletion of missing data
  outcome <- outcome[is.na(outcome)==FALSE & is.na(predictor)==FALSE]
  predictor <- predictor[is.na(outcome)==FALSE & is.na(predictor)==FALSE]
  # create basic scatterplot
  plot(outcome~predictor, ylab=my.y.axis.label, xlab=my.x.axis.label)
  # create ten equal-sized bins based on the predictor
  ten.bins <- ceiling(rank(predictor)/(length(predictor)/10))
  # calculate the mean outcome for each bin
  outcome.bins <- aggregate(outcome, by=list(ten.bins), FUN=mean)
  # calculate the mean predictor for each bin
  predictor.bins <- aggregate(predictor, by=list(ten.bins), FUN=mean)
  # in each of the ten bins, plot the mean outcome vs. the mean predictor
  points(predictor.bins[,2], outcome.bins[,2], pch=16, col='red')
  # fit the model for the sake of adding a fitted logistic curve
  logistic.model <- glm(outcome ~ predictor, family=binomial("logit"))
  # create protypical predictor values for the fitted logistic curve
  proto.pred <- seq(min(predictor), max(predictor), by=.001)
  # generate predicted values for the fitted logistic curve
  pred <- predict(logistic.model, newdata=data.frame(predictor=proto.pred))
  # create an inverse logit function
  inv.logit <- function(x) {exp(x)/(1+exp(x))}
  # plot the fitted logistic curve
  lines(prototemp, inv.logit(pred))
}


logistic.explore(placement, mcas, 'Probability of Placement', '8th Grade Math MCAS Score')
```

A Fitted Logistic Ogive

A logistic ogive is an S-shaped curve. The logistic ogive is a member of the exponential family of curves.



Not bad!

Linear Models vs. Non-Linear Models

Some Linear Models:

$$PLACEMENT = \beta_0 + \beta_1 MCAS + \varepsilon$$

$$p(PLACEMENT = 1) = \beta_0 + \beta_1 MCAS$$

What makes these models linear is that the right-hand side of the equation is simply a variable (or variables) times something plus something. Recall from 8th grade algebra that *y* = *mx* + *b* is the equation for a straight line in two dimension. Recall (perhaps) from multivariable calculus or linear algebra that *z* = *mx* + ny+ *b* is the equation for a straight plane in three dimensions.

$$PLACEMENT = \beta_0 + \beta_1 MCAS + \beta_2 ASIAN + \beta_3 BLACK + \beta_4 HISPANIC + \beta_5 MIXED + \varepsilon$$

Not all models are linear! To predict probabilities with a logistic ogive, we can go non-linear. (But we need not!)

A Non-Linear Formulation of the Logistic Model:

$$p(PLACEMENT = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 MCAS)}}$$

In this formulation, you can see the "linear component", but it's in the exponent of *e* (Euler's number, approximately 2.7). And, all that is in the denominator of a strange fraction. Thus, the modeled outcome is clearly more than "a variable times something plus something."

An Algebraically Equivalent Linear Formulation of the Logistic Model:

$$\log_e\left(\frac{p(PLACEMENT = 1)}{1 - p(PLACEMENT = 1)}\right) = \beta_0 + \beta_1 MCAS$$

The left-hand side of the equation is in log odds (or "logits"). The right-hand side is linear. This model is "linear in the log odds" or "linear in the logits." (More about odds in the next slide.)

Hitherto in this course, all our models have been so-called "general linear models." Our outcomes and predictor(s) have been linked by an "identity" where the conditional distributions of residuals were "normal." Now, we are entering the world of "general*ized* linear models." We can use links other than "identity." Here, we use a "logit" link where the conditional distribution of residuals is "binomial."

Understanding "Log Odds": Odds and Logs

Odds: The Wizard of Odds

Have you have said of a possible event, "The odds are 50/50" or "There's a 50/50 chance"? You were using odds. You were saying, "it would be a fair bet, if I were to wager $50 that the event ***will*** happened, and you were to wager $50 that the event ***won't*** happen, and winner takes all." Gamblers like odds. Note that 50/50 is a fraction that simplifies to 1/1 which simplifies to 1, but it's always best to think of odds as a fraction, no matter how they are reported.

Probability is one way to quantify chance. Odds are another way to quantify chance. Probabilities, which are bounded from 0 to 1, and odds, which have a lower bound of 0 but no upper bound as they can go to positive infinity, have a one-to-one relationship. Iff the probability is 0, then the odds are 0. Iff the probability is 1, then the odds are infinite. If you tell me the probability, I can tell you the odds, and vice versa. Here are the rules:

$$odds(EVENT=1)=\frac{p(EVENT=1)}{p(EVENT=0)}=\frac{p(EVENT=1)}{1-p(EVENT=1)} \qquad p(EVENT=1)=\frac{odds(EVENT=1)}{1+odds(EVENT=1)}$$

Here's my trick for converting odds to probabilities: Turn the odds to a fraction in which the numerator and denominator sum to 100. Then, the numerator is the probability in percentage form.

$$\text{iff } odds=1=\frac{50}{50}, \text{ then } p=.50 \quad \text{iff } odds=3=\frac{75}{25}, \text{ then } p=.75 \quad \text{iff } odds=.25=\frac{20}{80}, \text{ then } p=.20 \quad \text{iff } odds=99=\frac{99}{1}, \text{ then } p=.99$$

When the probability is greater than .50 (or 50%), the odds are greater than 1, and vice versa.
When the probability is less than .50 (or 50%), the odds are between 0 and 1, and vice versa.

Logs: I'm a Lumberjack, and I'm Okay

Logs (or logarithms) are related to exponents (i.e., powers).
The base of the "natural log" is *e* (Euler's number, about 2.7). Euler's number is one of the five most important numbers: 0, 1, π, *i* and *e*.
(Some more explanatiion goes here.)

Log odds (or logits) are unbounded (with a domain of negative infinity to positive infinity). This is important because our predictors are not necessarily bounded, so our outcomes should be not necessarily bounded.

Fitting The Logistic Model: Maximum Likelihood Estimation (MLE)

Quick Introduction to MLE: The classic role-playing game, *Dungeons and Dragons*, is famous for using funky dice to generate random numbers. In addition to your typical 6-sided dice (which randomly generate numbers from 1-6), *D&D* uses 4-sided dice (1-4), 8-sided dice (1-8), 10-sided dice (0-9), 12-sided dice (1-12), and 20-sided dice (1-20). Thus there are six types of dice. Suppose I role a single die behind my dungeon-master screen, and you have to guess which die (4-, 6-, 8-, 10-, 12-, or 20-sided). I rolled a 7. What is your guess? What is the probability of rolling a 7 given that I rolled a 4-sided die? $p(7 \mid 4\text{-sided}) = 0$, $p(7 \mid 6\text{-sided}) = 0$, $p(7 \mid 8\text{-sided}) = .125$, $p(7 \mid 10\text{-sided}) = .1$, $p(7 \mid 12\text{-sided}) = .083$, and $p(7 \mid 20\text{-sided}) = .05$. Note that the 8-sided die gives us the maximum likelihood.

Hitherto in this course, we have estimated the parameters for our models using the method of ordinary least squares (OLS, see Unit 1). With OLS, we choose the parameters (the intercept (β_0) and slope (β_1)) that minimized the sum of squared residuals. As it happens, when we *minimized* the sum of squared residuals, we *maximized* the likelihood of the data given the parameters (β_0 and β_1), *when* our residuals are normally distributed. When our outcome is dichotomous, our residuals are no longer normally distributed; they are binomially distributed.

When we ask about a likelihood, we ask, given parameters (e.g., intercept and slope), how likely is the data that we observe? No matter what the parameters, it's always unlikely that we observe exactly the data that we observe. Nevertheless, the data are more likely under some parameters than under other parameters. When we ask about MAXIMUM likelihood, we ask what parameters make the data that we observe MOST likely.

Example Data Set

ID	PLACEMENT	MCAS
1	0	218
2	0	228
3	1	248
4	1	250
5	0	254
6	1	260
7	1	268
8	1	270
9	1	279

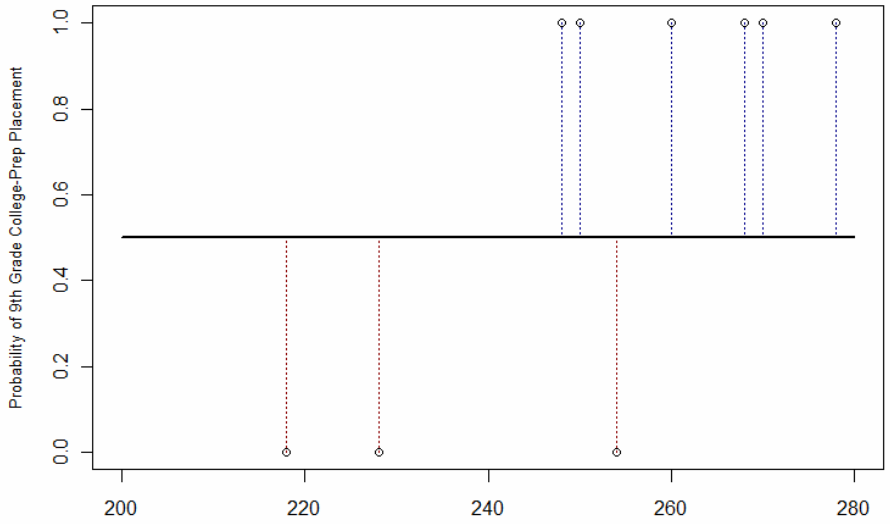
For the sake of illustration, let's work with a small data set. Let's calculate the likelihood that we would observe this data if students were placed into college-prep math based on a coin toss, thus our intercept and slope for our logistic model are both zero. Recall from grade-school probability that the probability of calling a coin flipped in the air is .5. The probability of calling two coin flips in a row is .25 (.5*.5). The probability of calling three coin flips in a row is .125 (.5*.5*.5). Now, if students are placed by coin flip, the likelihood that we observe *PLACEMENT* values of 0, 0, 1, 1, 0, 1, 1, 1 and 1, for subjects 1-9 respectively, is the same as the likelihood of tossing a coin tails, tails, heads, heads, tails, heads, heads, heads and then heads.

$$\text{Likelihood}(\text{data} \mid \text{intercept} = 0, \text{slope} = 0) = .5*.5*.5*.5*.5*.5*.5*.5*.5 = .00195$$

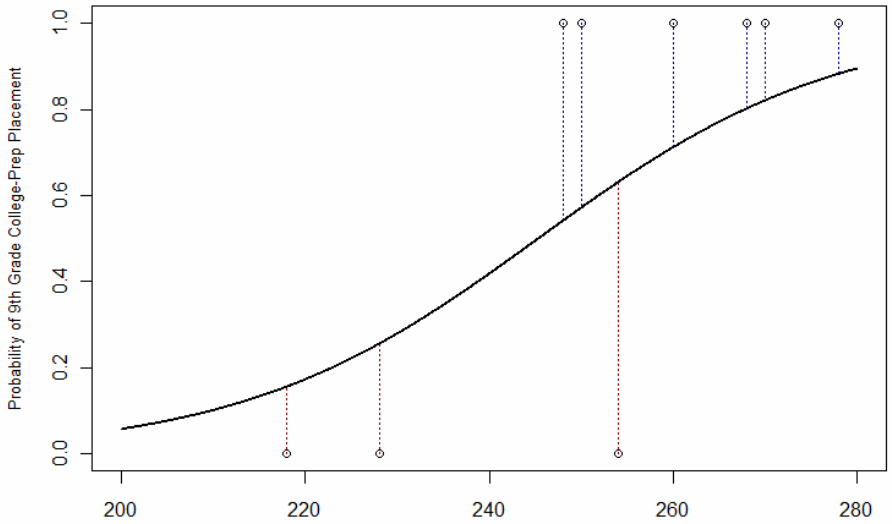
We can calculate the likelihoods for all combinations of intercepts and slopes and observe the MAXIMUM likelihood. (Or we can use calculus!) On the next slide are four graphical examples with this data.

Four Graphical Examples of Likelihoods

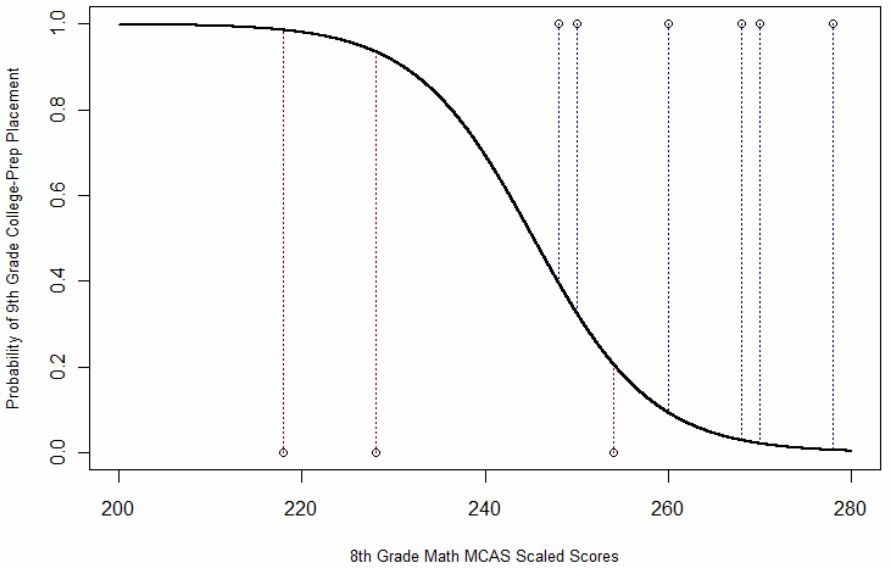
Likelihood (data | intercept = 0.00 , slope = 0.000) = 0.50 * 0.50 * 0.50 * 0.50 * 0.50 * 0.50 * 0.50 * 0.50 * 0.50 = 0.00195



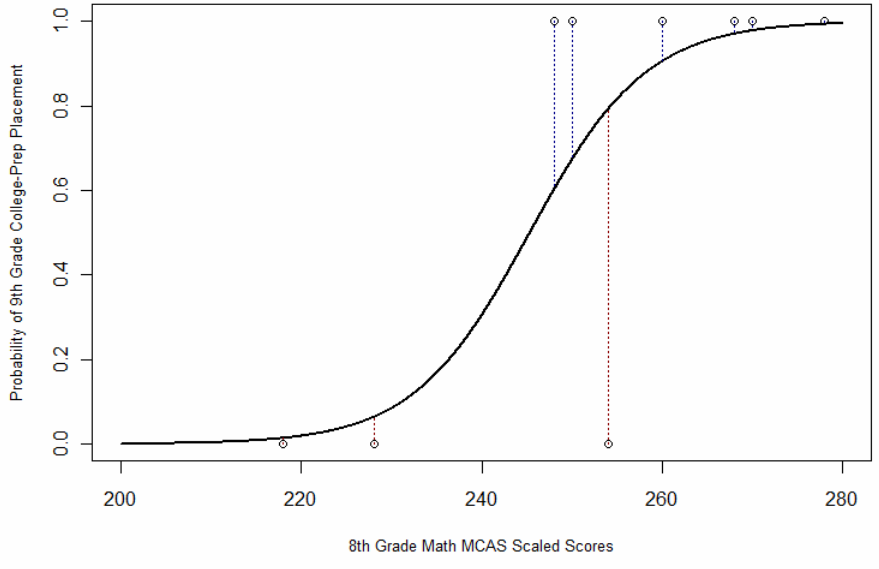
Likelihood (data | intercept = -15.14 , slope = 0.062) = 0.84 * 0.74 * 0.54 * 0.57 * 0.37 * 0.71 * 0.80 * 0.82 * 0.88 = 0.02971



Likelihood (data | intercept = 37.85 , slope = -0.154) = 0.01 * 0.06 * 0.40 * 0.33 * 0.79 * 0.09 * 0.03 * 0.02 * 0.01 = 0.00000



Likelihood (data | intercept = -37.85 , slope = 0.154) = 0.99 * 0.94 * 0.60 * 0.67 * 0.21 * 0.91 * 0.97 * 0.98 * 0.99 = 0.06628



R Script For Maximum Likelihood Animation

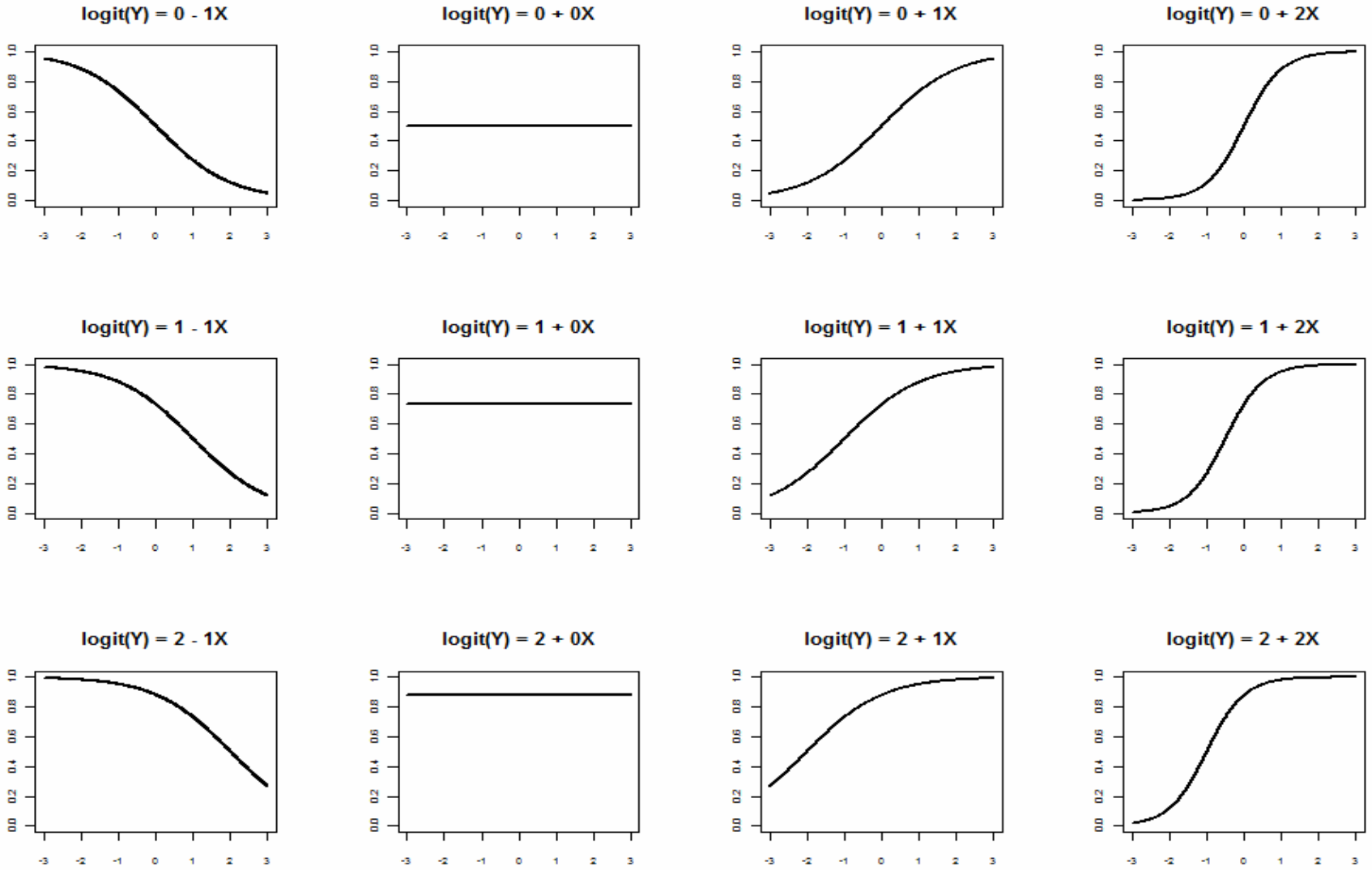
```
maximum.likelihood.animation <- function(intercept=0, slope=0) {
  old.par <- par(pty='s', lend=1)   # save current graphical paramters
    on.exit(par(old.par))            # reset graphical parameters when done
  ID <- 1:9                         # generate data
  PLACEMENT <- c(0,0,1,1,0,1,1,1,1)
  MCAS <- c(218,228,248,250,254,260,268,270,278)
  graphics.off()                    # turn off graphics.
  windows(8, 6)                     # open a carefully sized window to fit the bins as per next.
  par(mar=c(6, 4, 4, 2) + 0.1)
  inv.logit <- function(x) {exp(x)/(1+exp(x))}    # create an inverse logit function
  proto.MCAS <- seq(200, 280, .01)  # create protypical values for the fitted probability curve
  intercept.diff <- ((-37.8524)-intercept)/50     # generate the 60 steps to a perfect fit
  new.intercept <- c(rep(intercept, 10), intercept+(1:50)*intercept.diff)
  slope.diff <- ((0.1543)-slope)/50
  new.slope <- c(rep(slope, 10), slope+(1:50)*slope.diff)
  for (i in 1:60) {                 # begin animation
  plot(PLACEMENT~MCAS, ylab='Probability of 9th Grade College-Prep Placement',  # draw the plot
    xlab='8th Grade Math MCAS Scaled Scores', ylim=c(0,1), xlim=c(200,280), cex.lab=.8)
  proto.fitted.prob <- inv.logit(new.intercept[i] + new.slope[i]*proto.MCAS)    # create protypical values
  fitted.prob <- inv.logit(new.intercept[i] + new.slope[i]*MCAS)
  lines(proto.MCAS, proto.fitted.prob, lwd=2)     # draw the fitted probability curve
  segments(MCAS[PLACEMENT==0], PLACEMENT[PLACEMENT==0], MCAS[PLACEMENT==0],
    fitted.prob[PLACEMENT==0], lty='dotted', col="darkred")   # highlight the fitted probability for each observation
  segments(MCAS[PLACEMENT==1], PLACEMENT[PLACEMENT==1], MCAS[PLACEMENT==1],
    fitted.prob[PLACEMENT==1], lty='dotted', col="darkblue")
  prob.given <- (fitted.prob^PLACEMENT)*((1-fitted.prob)^(1-PLACEMENT))   # show the likelihood math
  prob.given.r <- formatC(prob.given, digits=2, format='f')
  {title(main=bquote('Likelihood (data | intercept ='
    ~.(formatC(new.intercept[i], digits=2, format='f'))~', slope ='
    ~.(formatC(new.slope[i], digits=3, format='f'))~') ='
    ~.(prob.given.r[1])~'*'
    ~.(prob.given.r[2])~'*'
    ~.(prob.given.r[3])~'*'
    ~.(prob.given.r[4])~'*'
    ~.(prob.given.r[5])~'*'
    ~.(prob.given.r[6])~'*'
    ~.(prob.given.r[7])~'*'
    ~.(prob.given.r[8])~'*'
    ~.(prob.given.r[9])~'='
    ~.(formatC(prod(prob.given), digits=5, format='f'))), cex.main=.8)}
  Sys.sleep(1.00)}
}
maximum.likelihood.animation(intercept=0, slope=0)
```

You set the parameters with the function below (default is intercept=0 and slope=0). A window will appear with a plot of the data and the logistic ogive based on your parameters. Above will be the likelihood. After ten seconds, the parameters will begin seeking the maximum likelihood.

The Y-Intercept and Slope Of the Logistic Model

The slope's sign determines whether the curve is monotonic decreasing or monotonic increasing. The slope's magnitude determines the steepness of the slope, where the greater the magnitude, the greater the steepness.

The y-intercept sends the curve up and to the right.



Interpreting the Fitted Logistic Model

```
> model.1 <- glm(placement ~ mcas, family=binomial("logit"))
> summary(model.1)

Call:
glm(formula = placement ~ mcas, family = binomial("logit"))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.96229   0.02252   0.05975   0.23287   2.00780

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -44.43515    6.03228  -7.366 1.76e-13 ***
mcas          0.19524    0.02609   7.484 7.21e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 267.74  on 333  degrees of freedom
Residual deviance: 102.81  on 332  degrees of freedom
AIC: 106.81

Number of Fisher Scoring iterations: 7
```

$$\log_e\left(\frac{\hat{p}(PLACEMENT=1)}{1-\hat{p}(PLACEMENT=1)}\right) = -44 + 0.20MCAS$$

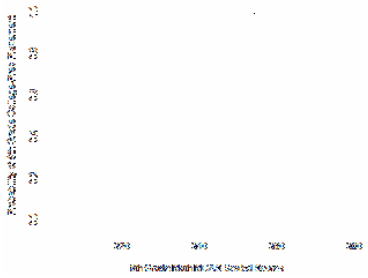
True but useless interpretation:
Comparing two students who differ by 1 point on the MCAS, we estimate that the log odds of placement in college-prep math are 0.20 logits greater for the higher scoring student. (Remember, "logits" means "log odds.")

For a better (but still very opaque) interpretation, we can convert from log odds to odds by antilogging (i.e., exponentiating) both sides of the fitted model. In particular, we can exponentiate the slope to obtain the odds ratio. E.g., exp(0.19524) = 1.215603

An odds ratio compares two odds. The baseline odds for comparison are the odds associated with a given level (any level!) of the predictor. We compare to those baseline odds the odds associated with one unit greater of the predictor. The odds ratio tells us how many times greater the odds are for one unit greater of the predictor.

Comparing two students who differ by 1 point on the MCAS, we estimate that the odds of placement in college-prep math for the higher scoring student are 1.22 times greater than the odds of placement for the lower scorer.

We found a statistically significant positive relationship between placement in college-prep math and 8th grade MCAS scores ($p < .001$). Students who score higher on the MCAS are more likely to be placed in college-prep math.



Pick a few key values of the predictor and discuss the associated fitted probabilities for them.

Bordeline failing/ni students with an MCAS score of 220 have a 19% chance of placing in college-prep math. Bordeline ni/proficient students with an MCAS score of 240 have a 92% chance.

$$\hat{p}(PLACEMENT=1) = \frac{1}{1+e^{-(-44+0.20MCAS)}}$$

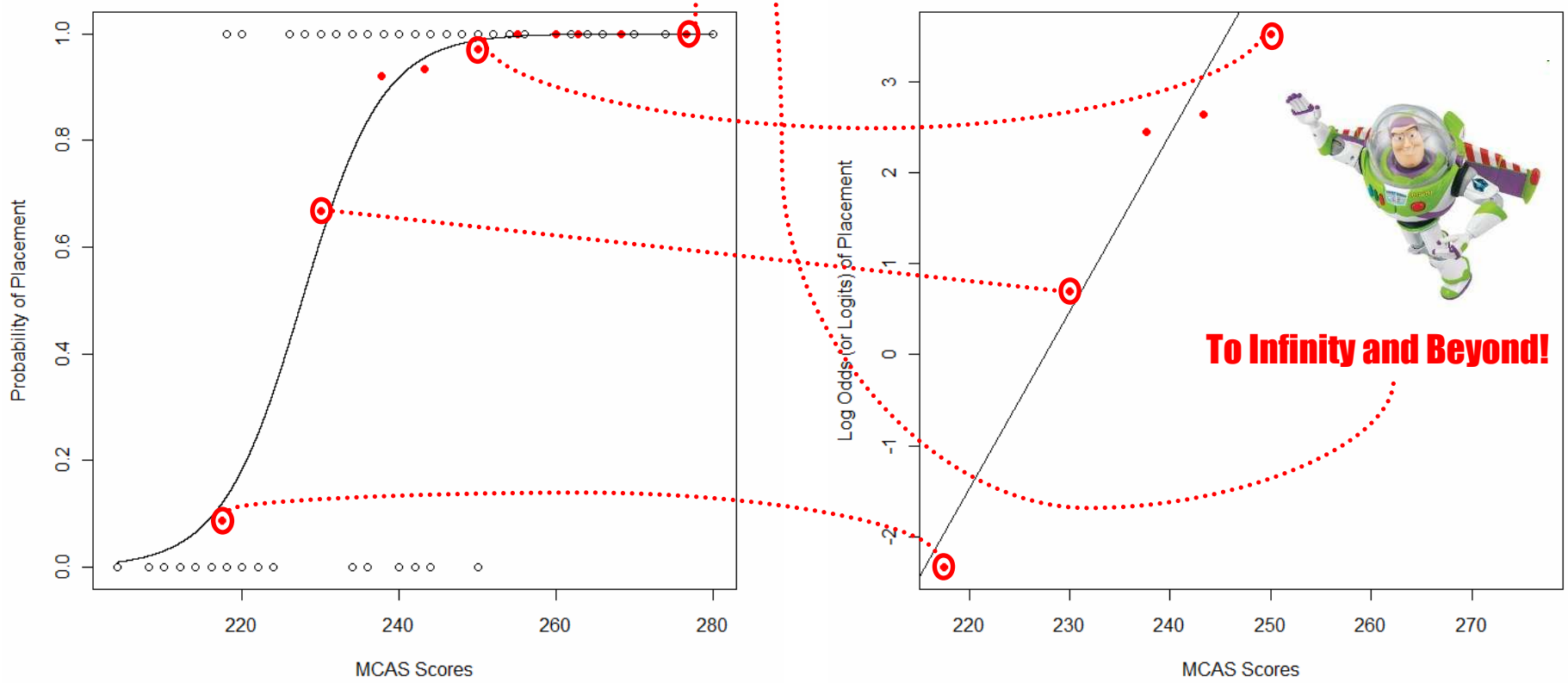
Although this formulation of the fitted logistic model is not directly interpretable, it allows us to think in terms of probabilities (or percentages). Graph it!

Checking the "Linear In The Log Odds" Assumption

There are only two assumptions that you need to check: independence and linearity. Linearity!?! Yes! Remember that our model is a generalized linear model (with a logit link). We are assuming that the logits (or log odds) of the outcome are linearly related to the predictor.

I recommend that you check the functional form in terms of probability. Do the red means/probabilities hug the logistic ogive? They do here.

Alternatively, you can logit transform all the probabilities (log(p/(1-p))). Annoyingly, for bins in which the mean/probability is 1, the logit is infinity.



Your method of inspection does not matter. If the dots hug the line in probability land, they will hug the line in logit land, and vice versa. The relationship is more intuitive in probability land. (In logit land, however, if the relationship is non-linear, a fix may be easier to see.)

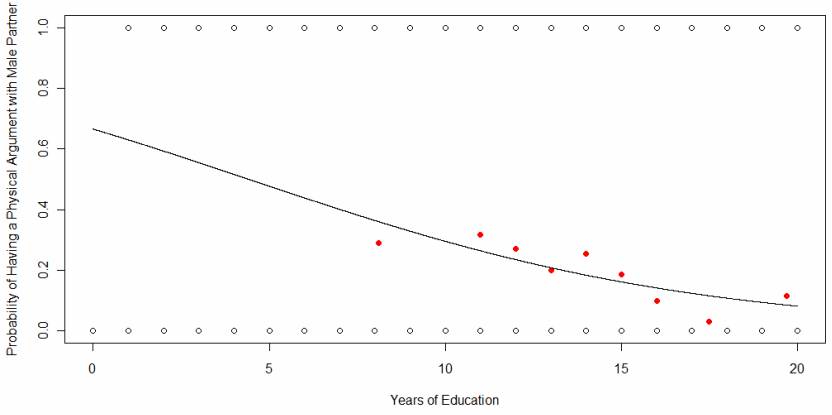
Dig the Post Hole

Appendix B Post Hole:

Interpret a fitted simple logistic regression model, noting the direction, magnitude and statistical significance of the relationship. For magnitude, compare two fitted probabilities. As always, check the assumptions.

Evidentiary Material: Regression output and exploratory plot.

```
Call:
glm(formula = physical.argument.with.male.partner ~ years.of.education,
    family = binomial("logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4799  -0.7311  -0.5926  -0.4425   2.2426

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        0.68772    0.24826   2.770   0.0056 **
years.of.education -0.15589    0.01862  -8.372   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1971.9  on 1976  degrees of freedom
Residual deviance: 1897.1  on 1975  degrees of freedom
  (41 observations deleted due to missingness)
AIC: 1901.1
```



Sample: 2018 women in heterosexual relationships. National Couples Survey, 2005-2006. Physical argument: pushing, shoving, biting, pulling hair, hitting, throwing things, or using weapons.

Here is the answer blank:

Here is my answer:

We found a statistically significant negative relationship between physical arguments and years of education ($p < .001$). Women with 12 years of schooling have about a 25% chance of having had a physical argument. Women with 16 years of schooling have about a 15% chance. The linear-in-the-logits assumption appears reasonable. As for independence, there may be clustering by community.

Tips:

- Report direction and statistical significance as usual.
- For magnitude, choose any two probabilities to compare.
- Use the plot to gauge roughly the probabilities.
- Check for linearity and independence.

Multiple Logistic Regression

$$\log_e\left(\frac{\hat{p}(PLACEMENT=1)}{1-\hat{p}(PLACEMENT=1)}\right)=-42+.19MCAS-.68ASIAN-.77BLACK-1.7HISPANIC+13MIXED$$

$$\hat{p}(PLACEMENT=1)=\frac{1}{1+e^{-(-42+.19MCAS-.68ASIAN-.77BLACK-1.7HISPANIC+13MIXED)}}$$

```
Call:
glm(formula = placement ~ mcas + asian + black + hispanic + mixed,
    family = binomial("logit"))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-3.01818   0.02268   0.05795   0.21037   2.38268

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -41.97550    6.07595  -6.908 4.90e-12 ***
mcas          0.18608    0.02623   7.093 1.31e-12 ***
asian        -0.68116    0.82493  -0.826   0.4090
black        -0.77702    0.85553  -0.908   0.3638
hispanic     -1.74008    0.91215  -1.908   0.0564 .
mixed        12.87381 1322.71936   0.010   0.9922
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 267.74  on 333  degrees of freedom
Residual deviance:  98.13  on 328  degrees of freedom
AIC: 110.13

Number of Fisher Scoring iterations: 16
```

Note: exp(-1.74008) = 0.1755064, thus the odds ratio is 0.1755064.

Controlling for 8th grade math MCAS scores, the odds of placement in college-prep math for a Hispanic student are 0.176 times the odds of placement for a White student.

Instead of saying that Hispanic is 0.176 times White, we can say that White is 5.68 times Hispanic. 1/.0176 = 5.68

Controlling for 8th grade math MCAS scores, the odds of placement in college-prep math for a White student are 5.68 times greater than the odds of placement for a Hispanic student.

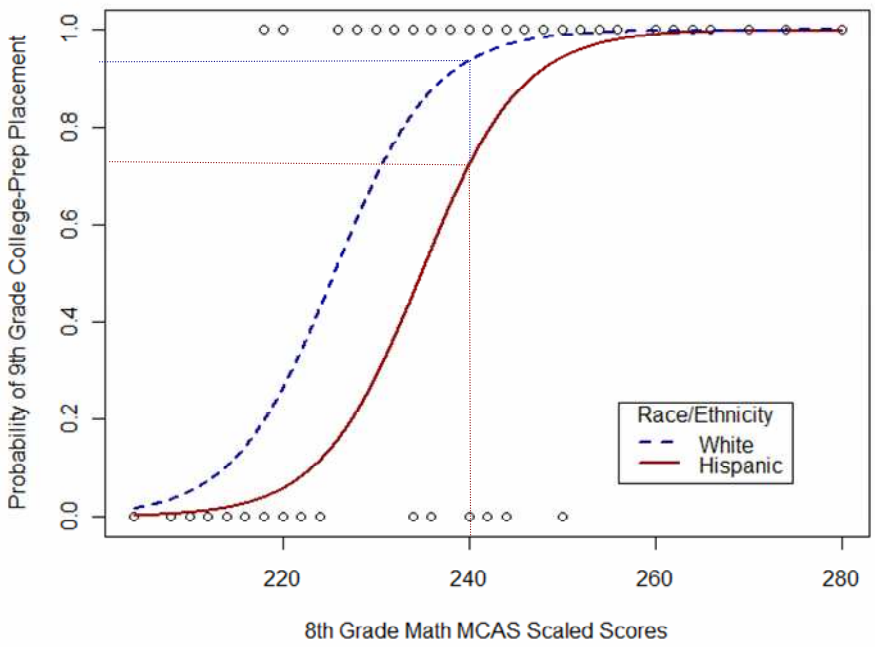
We can also compare probabilities.

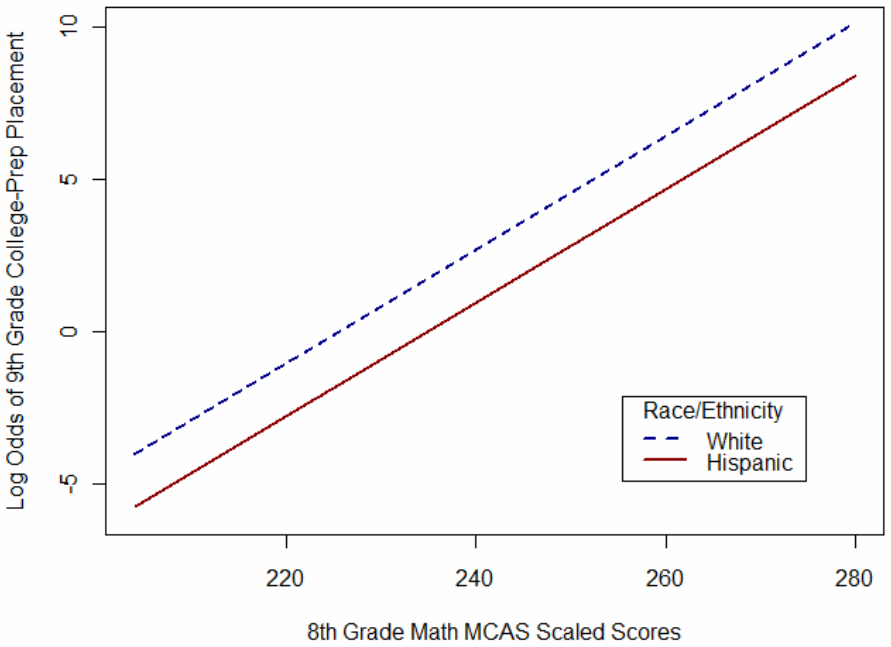
$$\hat{p}(PLACEMENT=1|WHITE=1,MCAS=240)=\frac{1}{1+e^{-(-42+.19*240)}}=.936$$

$$\hat{p}(PLACEMENT=1|HISPANIC=1,MCAS=240)=\frac{1}{1+e^{-(-42+.19*240-1.7*1)}}=.720$$

Comparing two students who scored just Proficient (MCAS = 240), the White student has a 94% chance of placement, whereas the Hispanic student has a 72% chance.

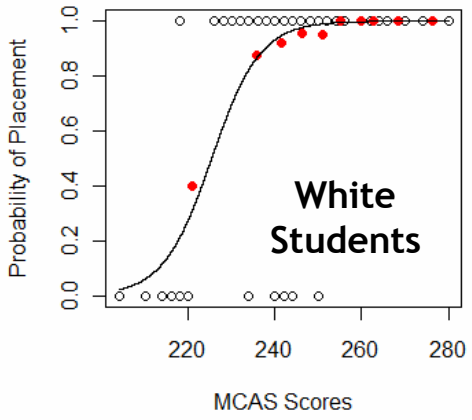
Graphing the Fitted Multiple Logistic Regression Model

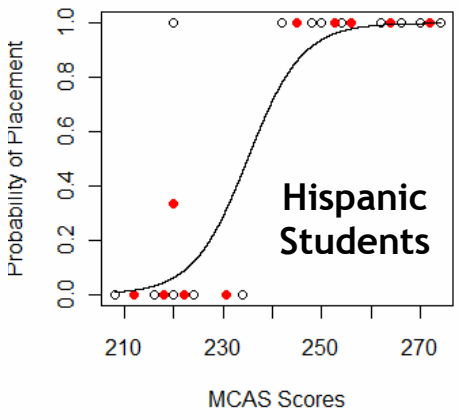




Notice the difference in percentages (i.e., the "effect" of ethnicity) varies by level of MCAS. This sounds like an interaction, but it's not. In probability land, the relationship between White/Latino and placement differs by MCAS level, but not in logit land, with parallel trend lines.

We can check the linear-in-the-logits assumption with exploratory plots. There are so few Hispanic students, however, that it's difficult to gauge whether the linear-in-the-log-odds assumption is met in the population. But, remember, the assumption is about the population! Since the assumption appears so solid for White students, and it does not seem to be contradicted by Hispanic students, we will go with it. This is one of the many spots in data analysis where we perhaps engage the art more than the science. (Also note that there may be clustering within classrooms.)





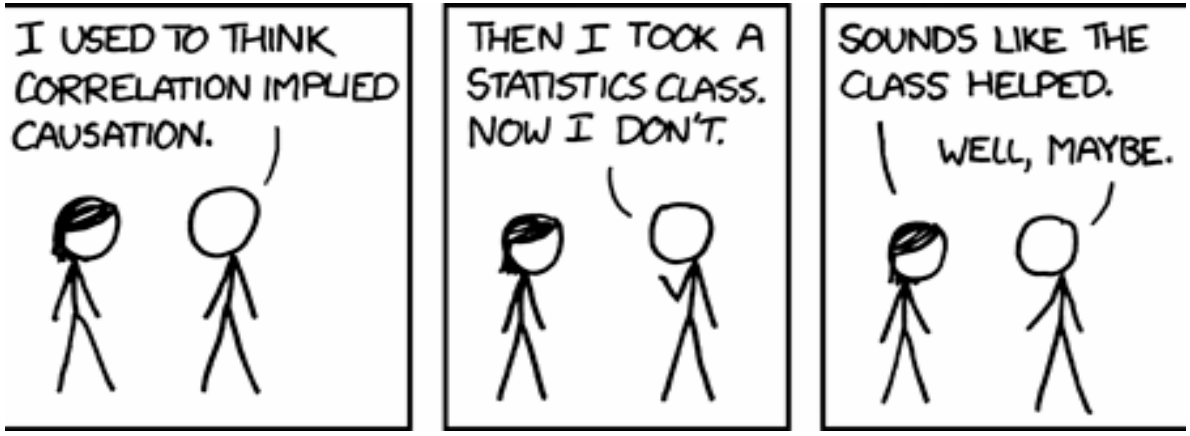
Correlation (Even Fancy Correlation) Does Not Imply Causation

This is not experimental data. We did not randomly assign students to MCAS scores. We did not randomly assign students to race/ethnicity. Logistic regression tells us that there is a relationship, but it does NOT tell us WHY there is a relationship.

Possible Causal Explanations:

- Bias in Placement
- Tracking Since 6th Grade
- Differences in Parental Advocacy
- MCAS Does Not Fully Capture Proficiency

Review the 3 Causal Rules of 3 as reprised from Unit 4.
Findings like this can tear a community apart. Handle with care.



Also note that we have yet to draw an inference from the sample to the population regarding *PLACEMENT* and *RACE*.

3 Causal Rules of 3

I. Causal conclusions require 3 conditions:
   A. Correlation
   B. Succession
   C. "Necessary Connexion"

II. In addition to your Predictor and Outcome, always consider the possible influence of a 3rd Hidden Confounding Variable.

III. When presenting your pet causal conclusion, present 2 other plausible causal conclusions for the sake of balance.

Taxonomy of Fitted Binomial Logistic Regression Models

Table B.1. Taxonomy of nested logistic regression models that display the fitted relationship between whether a student is placed in 9th grade college-prep mathematics as a function of 8th grade Math MCAS scaled score and race/ethnicity, for 334 Riverside High School students, class of 2009.

	Models		
	Null	#1	#2
Intercept	1.834***	-44.435***	-41.976***
	(0.159)	(6.032)	(6.076)
MCAS		0.195***	0.186***
		(0.026)	(0.026)
ASIAN			-0.681
			(0.825)
BLACK			-0.777
			(0.856)
HISPANIC			-1.740~
			(0.912)
MIXED			12.875
			(1322.719)
-2LL	267.74	102.81	98.13
df	333	332	328

Key: ~p<.10; *p<.05; **p<.01; ***p<.001

What is the Null Model?
What is that -2LL thing?

The Null Model is a model with NO predictors! We have seen the Null Model before, but not necessarily by that name. Recall from Unit 5 that, to calculate the R-square statistic, we compare the sum of squares error/residual (SSE) to the sum of squares total (SST). The SSE is a "badness of fit" statistic for our model and its predictors. The SST is a "badness of fit" statistic for the mean. You can think of the mean as the intercept of a null model, a model with no predictors. Recall from Unit 3 that I describe the mean as our "best guess" in absence of further information: NOT a "good guess," BUT a "bad guess that happens to be the best we can do without further information." In general, if we want to know if our model fits well, we need a baseline. The null model provides our baseline. Does our model improve our guesses over and above the null model?

In Units 1-10, we fit our general linear by minimizing the sum of squared error/residuals, so sums of squares supplied useful "badness of fit" statistics for model comparison. In Appendix B, we fit our general*ized* linear model by maximizing the likelihood. To get a juicy "badness of fit statistic," we log the likelihood and multiply it by -2 to get the -2 log-likelihood (-2LL).

```
model.0 <- glm(placement ~ NULL, family=binomial("logit"))
summary(model.0)
model.1 <- glm(placement ~ mcas, family=binomial("logit"))
summary(model.1)
model.2 <- glm(placement ~ mcas + asian + black + hispanic + mixed,family=binomial("logit"))
summary(model.2)
```

Likelihood Ratio Tests (LRTs)

Why is the -2 log-likelihood (-2LL) a "badness of fit" statistic? Why is the -2LL juicy?

Recall from earlier in this Appendix B that we estimate the parameters for our model (i.e., fit our model) by choosing the regression coefficients that maximize the likelihood of the data. So, for likelihood, the bigger, the better. Likelihood is a "goodness of fit" statistic. The likelihood is a probability (the probability of the data given the parameters), so it is bounded by 0 and 1 (but it's usually very close to 0, even when we maximize it). What happens when we log the likelihood? If we log 0, we get negative infinity. If we log 1, we get 0. Therefore, when we log our maximum likelihood, we will get a very large negative number! The log-likelihood is a "badness of fit" statistic; the bigger (the more negative), the worse, because the closer our log-likelihood is to negative infinity, the closer our likelihood was to 0. When we multiply our log-likelihood by -2, we get a doubly large positive number. Thus, the -2LL is a "badness of fit" statistic.

The -2LL is important because it allows us to test whether more complex models are statistically significantly better than their simpler counterparts. Is the likelihood of the data given Model 2 statistically significantly greater than the likelihood of the data given Model 1? Is the likelihood of the data given Model 1 statistically significantly greater than the likelihood of the data given the Null Model? This test is called the *Likelihood Ratio Test (LRT)*. Differences in -2LL between ***nested*** models follow a known sampling distribution, the chi-square distribution with degrees of freedom equal to the difference in degrees of freedom between the nested models. Hurray, for known sampling distributions!

$$\chi^2(df_{\text{simpler model}} - df_{\text{more complex model}}) = (-2ll_{\text{simpler model}}) - (-2ll_{\text{more complex model}}), \quad p = ???$$

Is Model 2 statistical significantly better than Model 1?

$$\chi^2(4) = 4.68, \quad p = .322$$

4 = $df_{\text{simpler model}}$ - $df_{\text{more complex model}}$ = 332 - 328
4.68 = $(-2ll_{\text{simpler model}})$ - $(-2ll_{\text{more complex model}})$ = 102.81 - 98.13

Based on a likelihood ratio test, race/ethnicity is not a statistically significant predictor of placement, controlling for MCAS scores, $X^2(4)=4.68$, $p = .322$. It is plausible that the race/ethnicity differences we see are due to sampling error.

Is Model 1 statistical significantly better than the Null Model?

$$\chi^2(1) = 164.93, \quad p < .001$$

```
> pchisq(4.68, df=4, lower.tail=FALSE)
[1] 0.3217343
> pchisq(164.93, df=1, lower.tail=FALSE)
[1] 9.475483e-38
```

Logistic Regression with Ordinal Polychotomous Outcomes

So far, we've looked at logistic regression with a dichotomous outcome, but what if our outcome is polychotomous? As it happens, at Riverside High School, there are in fact three levels of math courses, not just "Business" and "College-Prep," but also "Honors." Up to this point, I have collapsed "College-Prep" and "Honors," because I wanted to introduce you to binary logistic regression (i.e., logistic regression with dichotomous outcomes), but now let's consider the polychotomous outcome. There are two types of polychotomies, nominal and ordinal. The outcome, ***LEVEL***, is an ***ordinal*** polychotomy, because 2 = "Honors" is more advanced than 1 = "College-Prep" is more advanced than 0 = "Business." (We will treat LEVEL as a polychotomous variable, as opposed to a continuous variable, because we do not trust that the jump from 0 to 1 is the same as the jump from 1 to 2 (i.e., we do not trust that the scale is interval). We will take a quick look at the fitted ordinal logistic regression model, by following the detailed steps in http://www.ats.ucla.edu/stat/r/dae/ologit.htm.

```
> library(Design)
> ddist<- datadist(mcas)
> options(datadist='ddist')
> ol.model <- lrm(level ~ mcas, na.action=na.pass)
> print(ol.model)

Logistic Regression Model

lrm(formula = level ~ mcas, na.action = na.pass)


Frequencies of Responses
  0   1   2
 46 205  83

       Obs  Max Deriv Model L.R.       d.f.            P
       334      1e-04     307.75          1            0
     Tau-a         R2      Brier
     0.418      0.716      0.044

      Coef     S.E.    Wald Z P
y>=1 -44.0917 4.09852 -10.76 0
y>=2 -51.2684 4.64364 -11.04 0
mcas   0.1938 0.01770  10.95 0
```

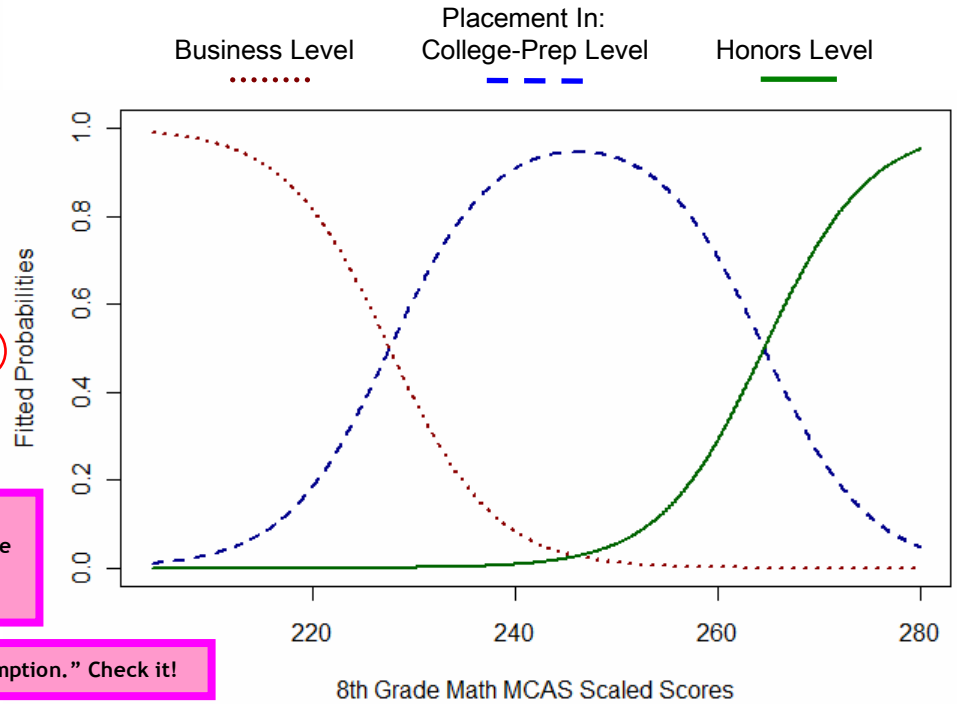
Is our ordinal logistic model (with MCAS as a predictor) statistical significantly better than the ordinal logistic model with no predictors (i.e., the null model)?

$\chi^2(1) = 307.75, \quad p < .001$

Notice that the parameter estimate for MCAS (.1938) is the same for both groups (*Honors* and *CollegePrep-or-Honors*).

$$\hat{\text{logit}}(Honors) = -51.27 + .1938MCAS$$

$$\hat{\text{logit}}(CollegePrep \text{ or } Honors) = -44.09 + .1938MCAS$$



This is by assumption of the ordinal logistic model, the "proportional odds assumption." Check it!

Checking the Proportional Odds Assumption

Note that http://www.ats.ucla.edu/stat/r/dae/ologit.htm recommends a method, but I think mine is more intuitive. I just adapt the exploratory logistic function from earlier so that we can SEE the assumption.

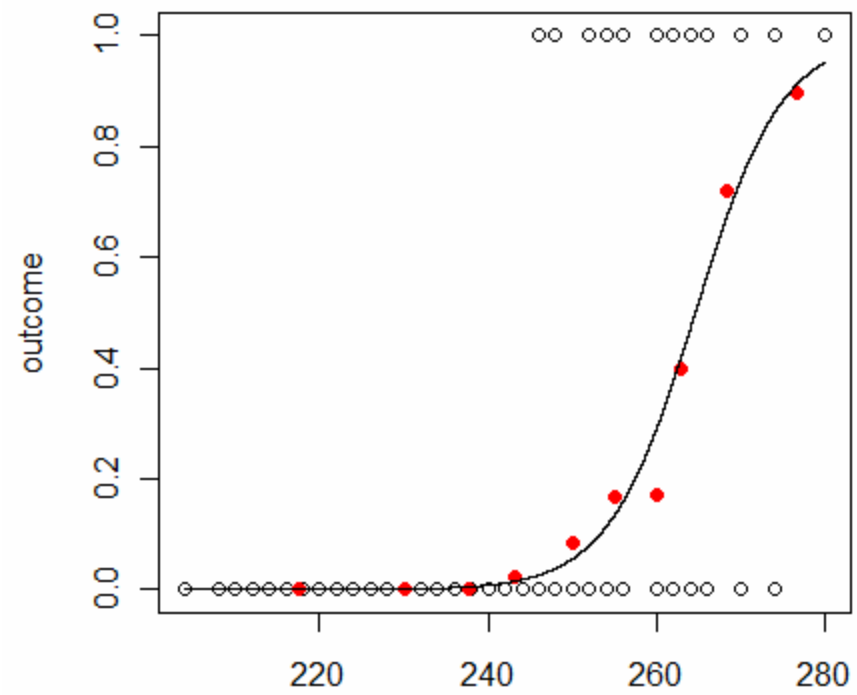
$$\hat{\text{logit}}(Honors) = -51.27 + .1938MCAS$$

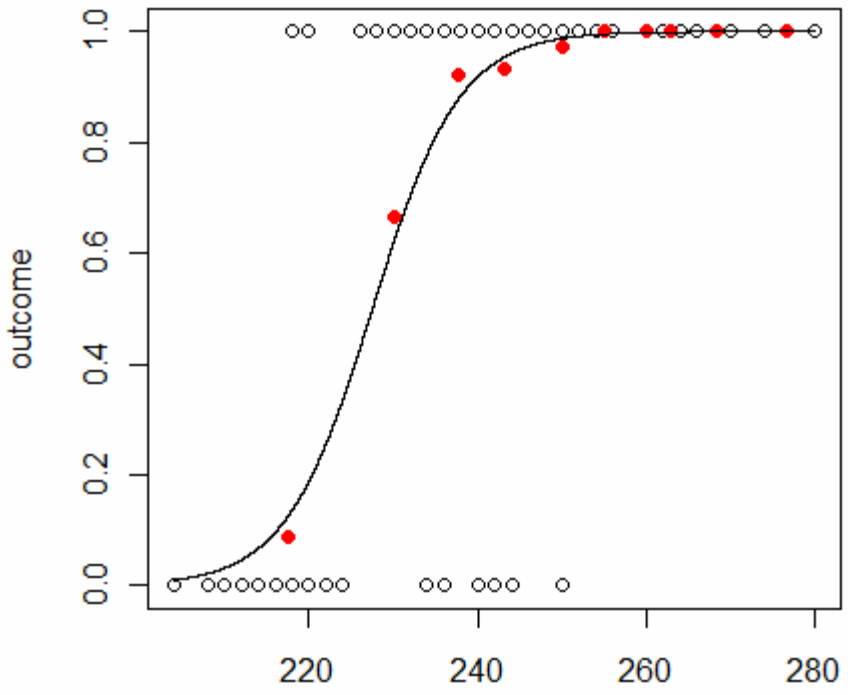
$$\log_e\left(\frac{\hat{p}(LEVEL >= 2)}{1-\hat{p}(LEVEL >= 2)}\right) = -51.37 + .1938MCAS$$

$$\hat{\text{logit}}(CollegePrep \text{ or } Honors) = -44.09 + .1938MCAS$$

$$\log_e\left(\frac{\hat{p}(LEVEL >= 1)}{1-\hat{p}(LEVEL >= 1)}\right) = -44.09 + .1938MCAS$$

```
check.proportional.odds(as.numeric(level>=2), mcas, logit.intercept=-51.27, logit.slope=.1938)
check.proportional.odds(as.numeric(level>=1), mcas, logit.intercept=-44.09, logit.slope=.1938)
```





To me, the proportional odds assumption looks pretty darn reasonable (as does the linear-in-the-log-odds assumption).

R Script For Checking The Proportional Odds Assumption

```
# a function for checking proportional odds assumption
check.proportional.odds <- function(outcome, predictor, logit.intercept, logit.slope) {
  # pairwise deletion of missing data
  outcome <- outcome[is.na(outcome)==FALSE & is.na(predictor)==FALSE]
  predictor <- predictor[is.na(outcome)==FALSE & is.na(predictor)==FALSE]
  # create basic scatterplot
  plot(outcome~predictor)
  # create ten equal-sized bins based on the predictor
  ten.bins <- ceiling(rank(predictor)/(length(predictor)/10))
  # calculate the mean outcome for each bin
  outcome.bins <- aggregate(outcome, by=list(ten.bins), FUN=mean)
  # calculate the mean predictor for each bin
  predictor.bins <- aggregate(predictor, by=list(ten.bins), FUN=mean)
  # in each of the ten bins, plot the mean outcome vs. the mean predictor
  points(predictor.bins[,2], outcome.bins[,2], pch=16, col='red')
  # create protypical predictor values for the fitted logistic curve
  proto.pred <- seq(min(predictor), max(predictor), by=.001)
  # generate predicted values for the fitted logistic curve
  fitted.logits <- logit.intercept + logit.slope*proto.pred
  # create an inverse logit function
  inv.logit <- function(x) {exp(x)/(1+exp(x))}
  # plot the fitted logistic curve
  lines(proto.pred, inv.logit(fitted.logits))
}
check.proportional.odds(as.numeric(level>=2), mcas, logit.intercept=-51.27, logit.slope=.1938)
check.proportional.odds(as.numeric(level>=1), mcas, logit.intercept=-44.09, logit.slope=.1938)
```

Logistic Regression with Nominal Polychotomous Outcomes

Your polychotomous outcome need not be ordinal to apply logistic regression technique. Even though out outcome, LEVEL, is ordinal, for the sake of this quick example we'll treat it as nominal. We know the order is informative, but I won't tell the computer if you won't. We will take a quick look at the fitted nominal logistic regression model, by following the detailed steps in http://www.ats.ucla.edu/stat/r/dae/mlogit.htm.

```
> library(mlogit)
> level.factor <- as.factor(level)
> mldata<-mlogit.data(data.frame(level.factor, mcas),
+   varying=NULL, choice="level.factor", shape="wide")
> mlogit.model<- mlogit(level.factor~0|mcas, data = mldata,
+   reflevel="0")
> summary(mlogit.model)

Call:
mlogit(formula = level.factor ~ 0 | mcas, data = mldata, reflevel = "0",
    method = "nr", print.level = 0)

Frequencies of alternatives:
      0       1       2
0.13772 0.61377 0.24850

nr method
7 iterations, 0h:0m:0s
g'(-H)^-1g = 5.86E-07
gradient close to zero

Coefficients :
             Estimate Std. Error  t-value  Pr(>|t|)
alt1       -43.973486   6.105432  -7.2024 5.917e-13 ***
alt2       -94.657982   8.917163 -10.6153 < 2.2e-16 ***
alt1:mcas    0.193199   0.026417   7.3136 2.602e-13 ***
alt2:mcas    0.384754   0.036239  10.6170 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Log-Likelihood: -152.89
McFadden R^2:  0.50168
Likelihood ratio test : chisq = 307.86 (p.value=< 2.22e-16)
```

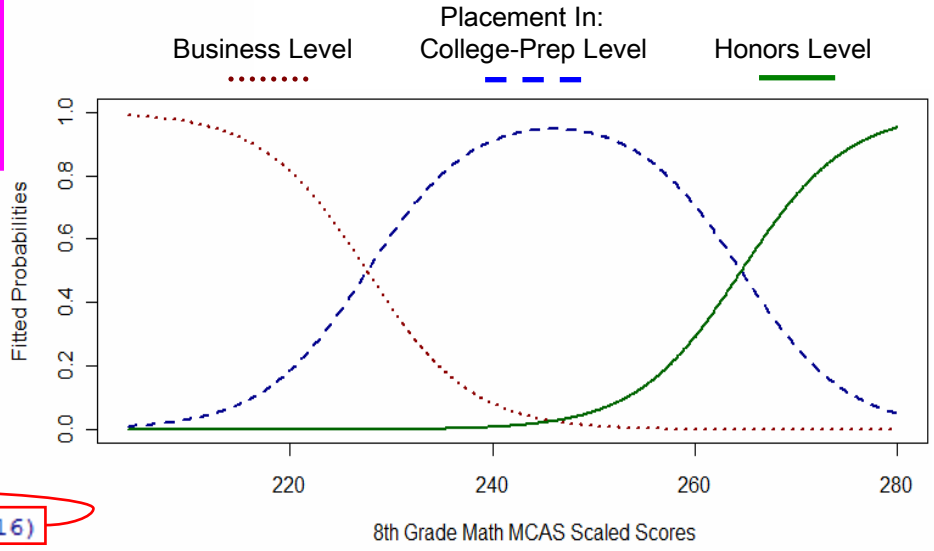
We could have chosen any level as the reference level. We chose LEVEL= 0, "Business Math."

$$\log_e\left(\frac{\hat{p}(LEVEL=1)}{\hat{p}(LEVEL=0)}\right) = -43.97 + .1931MCAS$$

$$\log_e\left(\frac{\hat{p}(LEVEL=2)}{\hat{p}(LEVEL=0)}\right) = -94.66 + .3848MCAS$$

There is no proportional-odds assumption, but there is an independence-of-irrelevant-alternatives assumption. Once I choose option A over option B, then any additional option C will be between option A and option C.

LRT



Special Property of Odds Ratios: Invariance Under Retrospective Sampling

As you can tell, I prefer probabilities to odds and odds ratios. The not-so-good reason for my preference is that I personally find odds and especially odds ratios unintuitive. The good reason for my preference is that my audiences find odds and especially odds ratios unintuitive. As a data-analyst, my job is to translate the results into language that my audience can understand.

Odds ratios, however, have a special property that may be useful when studying rare outcomes: invariance under retrospective sampling. What is prospective vs. retrospective sampling? Prospective sampling is the standard case in our field of education. With prospective sampling, we take a random sample from a (perhaps school) population, and we observe how the outcome happens to fall. Still prospectively, sometimes we take a *stratified* random sample, in which we oversample a minority group so the we have a sample size sufficient for precise estimates. This oversampling is fine as long as we include the group categories in our model as predictors in order to control for the potential confound. Such a stratified random sample is still a prospective sample, because we collected our data and let the outcome fall where it may. On the other hand, with retrospective sampling, we essentially stratify by the outcome! We do not let the outcome fall where it may; rather, we set the outcome by design and let the *predictors* fall where *they* may. This is weird, and usually a bad idea. Nevertheless, sometime we want to sample based on the outcome, especially when one level of the outcome is extremely rare. Odds-ratios permit retrospective sampling.

From our earlier prospective study based on a random sample of the entire Riverside High School population (N=334), we found that our dichotomous outcome, PLACEMENT, was 86% College-Prep (n=288) and 14% Business Math (n=46). This was a prospective study because we let the outcome fall where it may, but suppose we only had resources to collect data on 50 students. If we prospectively sampled, and let the outcome fall where it may, then we would have an N=50 broken down by about n=43 for College-Prep and n=7 for Business Math, based on the percentages in the population. That n=7 sadly does not give us enough statistical power to detect a reasonable effect size. Don't we wish we could sample n=25 and n=25 on the outcome (letting the predictors fall where they may)? We CAN because the odds ratio is invariant under retrospective sampling.

Logistic regression model fit based on a **prospective** sample (N=334, n=288, n=46):

$$\log_e\left(\frac{\hat{p}(LEVEL=1)}{\hat{p}(LEVEL=0)}\right) = -43.97 + .1931MCAS$$

Logistic regression model fit based on a **retrospective** sample (N=50, n=25, n=25):

$$\log_e\left(\frac{\hat{p}(LEVEL=1)}{\hat{p}(LEVEL=0)}\right) = -46.09 + .1939MCAS$$

The odds-ratios are equal in expectation. Remember to exponentiate (i.e., antilog) to get the odds-ratio!

The fitted intercepts are not equal, so neither are the fitted probabilities! NOT invariant.

See the ILLCAUSE practice for an example of a retrospective design. Note that, hitherto, the ILLCAUSE practice has been a prospective design based on a random sample stratified by a dichotomous predictor, *ChronicallyIll*. As long as we included *ChronicallyIll* as a predictor in our models of the ILLCAUSE data, our results were not confounded by the sampling method. For this practice example, however, *ChronicallyIll* is the outcome, not the predictor. All of a sudden our design switched from prospective (where we let the outcome fall as it may) to retrospective (where we sampled based on the outcome). Tricky!

Simulate Odds-Ratio Invariance Under Retrospective Sampling

```
# create data in case the real dataset is not handy
mcas <- rnorm(334, 240, 10)
p.placement.given.mcas <- function(x) 1/(1+exp(-(-43.97+ 0.1931*x)))
placement <- rbinom(334, 1, p.placement.given.mcas(mcas))
glm(placement ~ mcas, family=binomial("logit"))

retrospective.sample <- function(dichotomous.outcome, predictor, n.1=25, n.0=25){
  # pairwise deletion of missing data
  dichotomous.outcome <-
      dichotomous.outcome[is.na(dichotomous.outcome)==FALSE & is.na(predictor)==FALSE]
  predictor <- predictor[is.na(dichotomous.outcome)==FALSE & is.na(predictor)==FALSE]
  # take a random subsample of specified size
  subsampled.predictor <-
      c(sample(predictor[dichotomous.outcome==1], n.1),
      sample(predictor[dichotomous.outcome==0], n.0))
  subsampled.outcome <- c(rep(1, n.1), rep(0, n.0))
  # output a subsample
  data.frame(subsampled.outcome, subsampled.predictor)
}
pump.out.coefs <- function(o){
      glm(subsampled.outcome ~ subsampled.predictor,
      data=retrospective.sample(placement, mcas, n.1=25, n.0=25),
      family=binomial("logit"))$coefficients
}
simulation.results <- sapply(1:10000, pump.out.coefs)
simulation.results <- as.matrix(t(simulation.results))
# Use medians because some of the 10,000 samples don't allow convergence
   # so their results are crazy outliers
median.intercept <- median(simulation.results[,1])
median.slope <- median(simulation.results[,2])
median.intercept
median.slope
```

Appendix B Appendix: Key Concepts

3 Causal Rules of 3

- Causal conclusions require 3 conditions:
  - Correlation
  - Succession
  - "Necessary Connexion"
- In addition to your Predictor and Outcome, always consider the possible influence of a 3rd Hidden Confounding Variable.
- When presenting your pet causal conclusion, present 2 other plausible causal conclusions for the sake of balance.

Binary Logistic Regression:

There are only two assumptions that you need to check: independence and linearity. Linearity!?! Yes! Remember that our model is a generalized linear model (with a logit link). We are assuming that the logits (or log odds) of the outcome are linearly related to the predictor.

Ordinal Logistic Regression:

Notice that the parameter estimate for MCAS (.1938) is the same for both groups (*Honors* and *CollegePrep-or-Honors*). This is by assumption of the ordinal logistic model, the "proportional odds assumption." Check it!

Multinomial Logistic Regression:

There is no proportional-odds assumption, but there is an independence-of-irrelevant-alternatives assumption. Once I choose option A over option B, then any additional option C will be between option A and option C.

Appendix B Appendix: Key Interpretations

We found a statistically significant positive relationship between placement in college-prep math and 8th grade MCAS scores ($p < .001$). Students who score higher on the MCAS are more likely to be placed in college-prep math.

Comparing two students who differ by 1 point on the MCAS, we estimate that the odds of placement in college-prep math for the higher scoring student are 1.22 times greater than the odds of placement for the lower scorer.

Bordeline failing/ni students with an MCAS score of 220 have a 19% chance of placing in college-prep math. Bordeline ni/proficient students with an MCAS score of 240 have a 92% chance.

Controlling for 8th grade math MCAS scores, the odds of placement in college-prep math for a Hispanic student are 0.176 times the odds of placement for a White student.

Controlling for 8th grade math MCAS scores, the odds of placement in college-prep math for a White student are 5.68 times greater than the odds of placement for a Hispanic student.

Comparing two students who scored just Proficient (MCAS = 240), the White student has a 94% chance of placement, whereas the Hispanic student has a 72% chance.

Based on a likelihood ratio test, race/ethnicity is not a statistically significant predictor of placement, controlling for MCAS scores, $X2(4)=4.68$, $p = .322$. It is plausible that the race/ethnicity differences we see are due to sampling error.

Appendix B Appendix: Key Terminology

An odds ratio compares two odds. The baseline odds for comparison are the odds associated with a given level (any level!) of the predictor. We compare to those baseline odds the odds associated with one unit greater of the predictor. The odds ratio tells us how many times greater the odds are for one unit greater of the predictor.

Perceived Intimacy of Adolescent Girls (Intimacy.csv)

- Overview: Dataset contains self-ratings of the intimacy that adolescent girls perceive themselves as having with: (a) their mother and (b) their boyfriend.

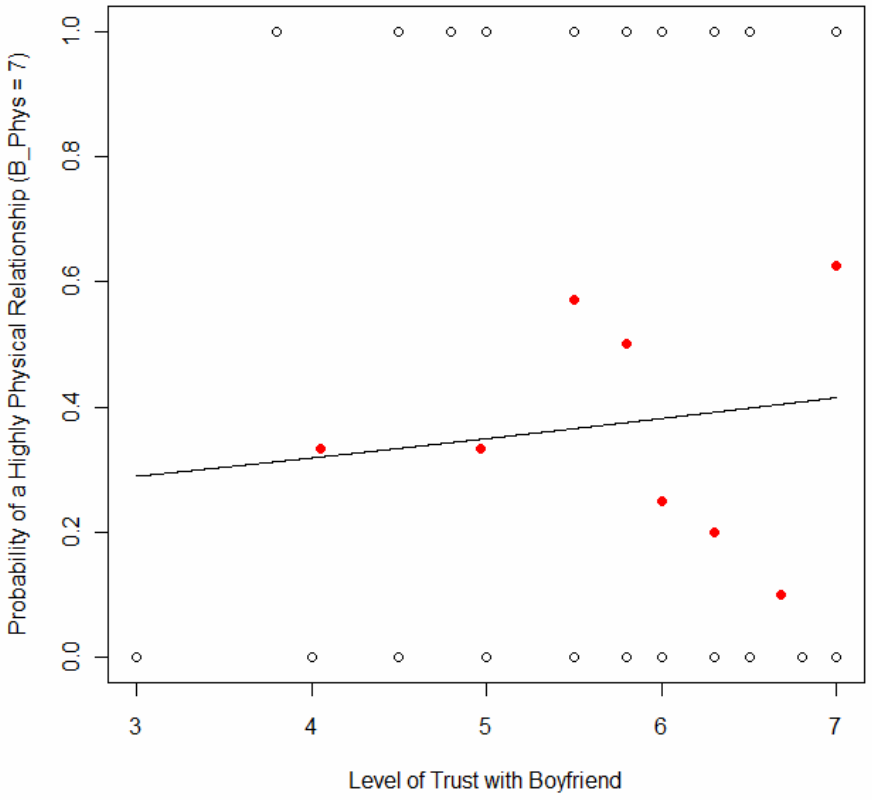
- Source: HGSE thesis by Dr. Linda Kilner entitled Intimacy in Female Adolescent's Relationships with Parents and Friends (1991). Kilner collected the ratings using the Adolescent Intimacy Scale.
- Sample: 64 adolescent girls in the sophomore, junior and senior classes of a local suburban public school system.
- Variables:

B_Physical 1 if the student hit the ceiling (7) of physical intimacy with boyfriend
0 else

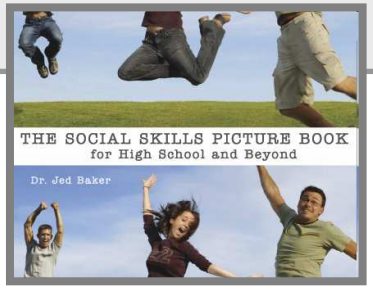
B_Trust A continuous variable (0-7) of trust in boyfriend

Perceived Intimacy of Adolescent Girls (Intimacy.csv)

```
> the.model <- glm(B.Physical ~ B.Trust, family=binomial("logit"))
> summary(the.model)

Call:
glm(formula = B.Physical ~ B.Trust, family = binomial("logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0351  -0.9992  -0.9269   1.3267   1.5256

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.318      1.758  -0.750    0.453
B.Trust        0.139      0.286   0.486    0.627

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 79.88  on 59  degrees of freedom
Residual deviance: 79.64  on 58  degrees of freedom
AIC: 83.64

Number of Fisher Scoring iterations: 4
```



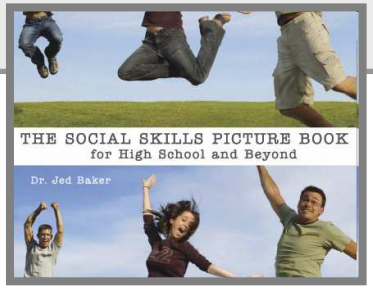
High School and Beyond (HSB.csv)



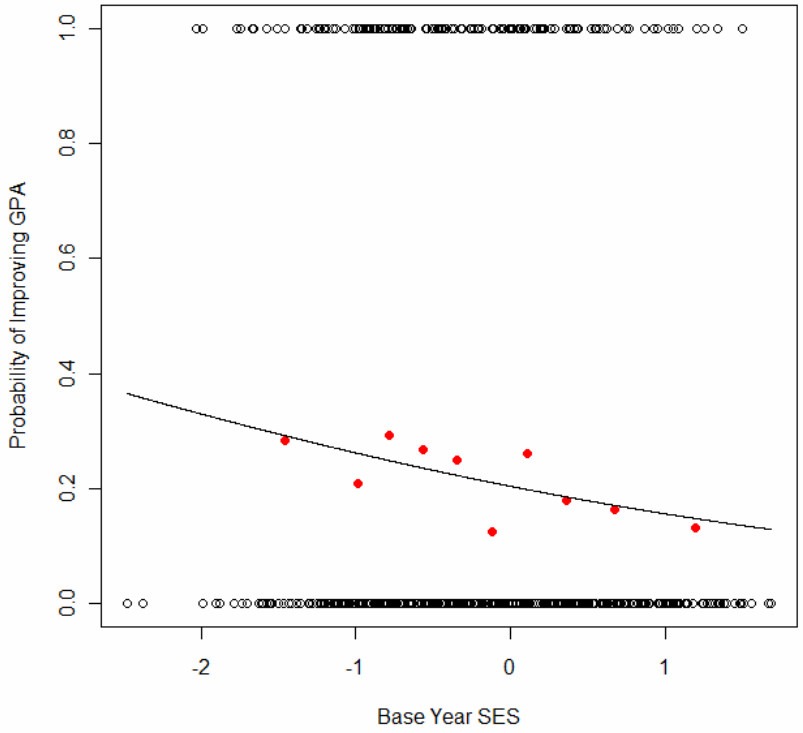
- Overview: High School & Beyond – Subset of data focused on selected student and school characteristics as predictors of academic achievement.
- Source: Subset of data graciously provided by Valerie Lee, University of Michigan.
- Sample: This subsample has 1044 students in 205 schools. Missing data on the outcome test score and family SES were eliminated. In addition, schools with fewer than 3 students included in this subset of data were excluded.
- Variables:

Improved.GPA	1 for students who improved their GPAs from 10th grade to 12th grade
	0 Else
BYSES	Base year SES

High School and Beyond (HSB.csv)




```
> the.model <- glm(Improved.GPA ~ BY.SES, family=binomial("logit"))
> summary(the.model)

Call:
glm(formula = Improved.GPA ~ BY.SES, family = binomial("logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9527  -0.7328  -0.6622  -0.5588   1.9993

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.3636     0.1011 -13.493  < 2e-16 ***
BY.SES       -0.3262     0.1237  -2.637  0.00836 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 703.97  on 672  degrees of freedom
Residual deviance: 696.85  on 671  degrees of freedom
AIC: 700.85

Number of Fisher Scoring iterations: 4
```



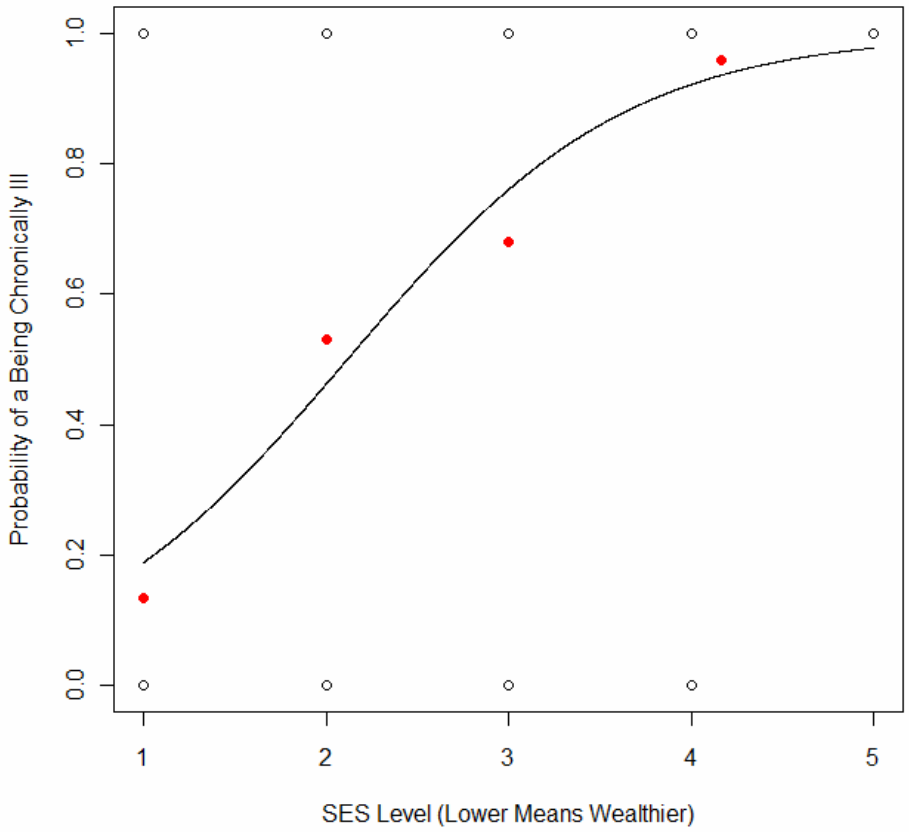
Understanding Causes of Illness (ILLCAUSE.csv)

- Overview: Data for investigating differences in children's understanding of the causes of illness, by their health status.
- Source: Perrin E.C., Sayer A.G., and Willett J.B. (1991). Sticks And Stones May Break My Bones: Reasoning About Illness Causality And Body Functioning In Children Who Have A Chronic Illness, *Pediatrics*, 88(3), 608-19.
- Sample: 301 children, including a sub-sample of 205 who were described as asthmatic, diabetic,or healthy. After further reductions due to the *list-wise deletion* of cases with missing data on one or more variables, the analytic sub-sample used in class ends up containing: 33 diabetic children, 68 asthmatic children and 93 healthy children.
- Variables:

ChronicallyIll	1 = Asthmatic or Diabetic, 0 = Health
SES	Child's SES (Note that a high score means low SES.)

Understanding Causes of Illness (ILLCAUSE.csv)

```
> the.model <- glm(ChronicallyIll ~ SES, family=binomial("logit"))
> summary(the.model)

Call:
glm(formula = ChronicallyIll ~ SES, family = binomial("logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2566  -1.1130   0.4040   0.7399   1.8280

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.7714     0.4839  -5.728 1.02e-08 ***
SES           1.3090     0.2130   6.146 7.94e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 283.37  on 204  degrees of freedom
Residual deviance: 228.30  on 203  degrees of freedom
AIC: 232.30

Number of Fisher Scoring iterations: 4
```



Warning! We cannot trust these probabilities. This sample is restrospective. We collected a random sample of chronically ill children and a random sample of healthy children. We collected observations based on the outcome and looked back. This retrospective design must be handled with care. Happily, logistic regression is specially capable for such designs. See SLIDE.

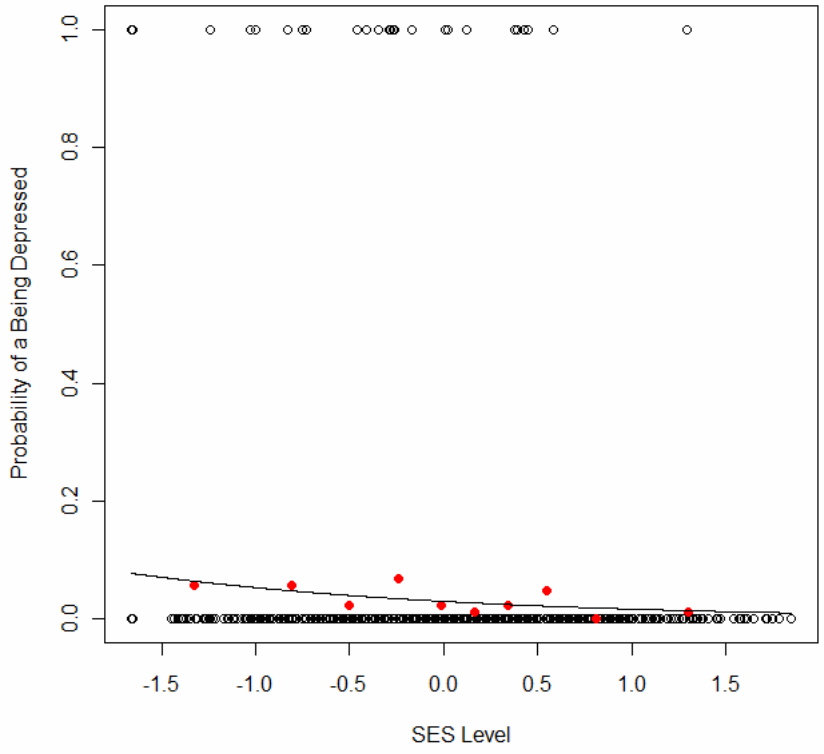
Children of Immigrants (ChildrenOfImmigrants.csv)

- Overview: "CILS is a longitudinal study designed to study the adaptation process of the immigrant second generation which is defined broadly as U.S.-born children with at least one foreign-born parent or children born abroad but brought at an early age to the United States. The original survey was conducted with large samples of second-generation children attending the 8th and 9th grades in public and private schools in the metropolitan areas of Miami/Ft. Lauderdale in Florida and San Diego, California" (from the website description of the data set).
- Source: Portes, Alejandro, & Ruben G. Rumbaut (2001). *Legacies: The Story of the Immigrant SecondGeneration*. Berkeley CA: University of California Press.
- Sample: Random sample of 880 participants obtained through the website.
- Variables:

Depressed	1 if the student scored 2 or more standard deviations above the mean depression level, 0 else
SES	Composite family SES score

Children of Immigrants (ChildrenOfImmigrants.csv)

```
> the.model <- glm(Depressed ~ SES, family=binomial("logit"))
> summary(the.model)

Call:
glm(formula = Depressed ~ SES, family = binomial("logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.3998  -0.2800  -0.2356  -0.2016   2.9334

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.5006     0.2076 -16.862   <2e-16 ***
SES          -0.6109     0.2565  -2.382   0.0172 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 248.17  on 879  degrees of freedom
Residual deviance: 242.43  on 878  degrees of freedom
AIC: 246.43

Number of Fisher Scoring iterations: 6
```



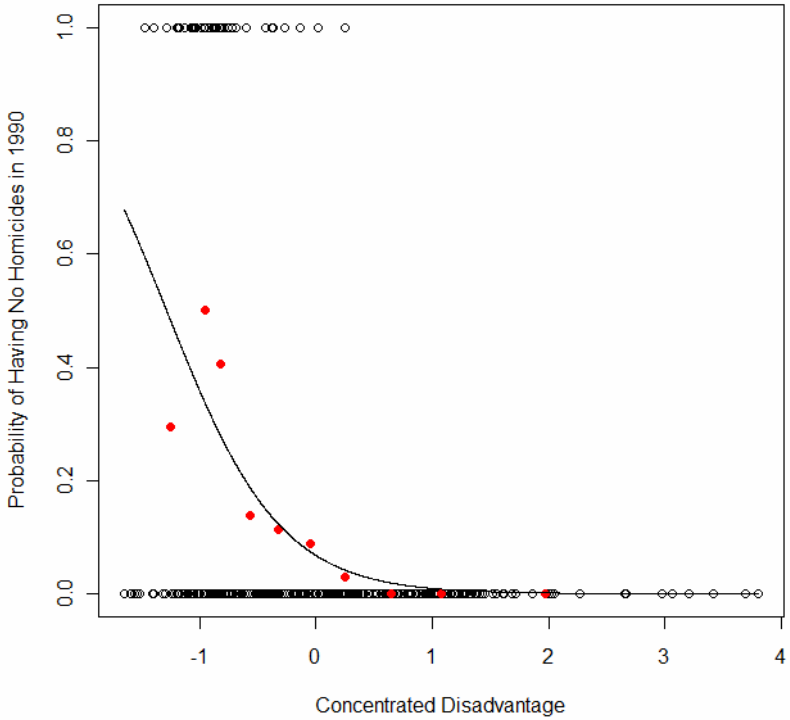
Human Development in Chicago Neighborhoods (Neighborhoods.csv)

- These data were collected as part of the Project on Human Development in Chicago Neighborhoods in 1995.
- Source: Sampson, R.J., Raudenbush, S.W., & Earls, F. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science*, 277, 918-924.
- Sample: The data described here consist of information from 343 Neighborhood Clusters in Chicago Illinois. Some of the variables were obtained by project staff from the 1990 Census and city records. Other variables were obtained through questionnaire interviews with 8782 Chicago residents who were interviewed in their homes.
- Variables:

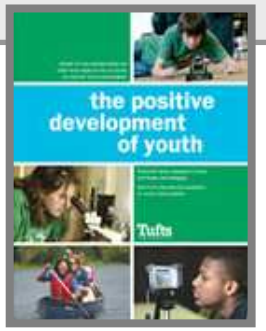
No.Homicides	1 if there were no homicides in the neighborhood in 1990 0 Else
Conc.Dis	A continuous composite variable measuring concentrated disadvantage

Human Development in Chicago Neighborhoods (Neighborhoods.csv)

```
> the.model <- glm(No.Homicides ~ Conc.Dis, family=binomial("logit"))
> summary(the.model)

Call:
glm(formula = No.Homicides ~ Conc.Dis, family = binomial("logit"))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.50515  -0.57661  -0.28944  -0.09909   2.52343

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.6303     0.2989  -8.800  < 2e-16 ***
Conc.Dis     -2.0450     0.3419  -5.981 2.22e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 294.96  on 341  degrees of freedom
Residual deviance: 228.88  on 340  degrees of freedom
AIC: 232.88

Number of Fisher Scoring iterations: 6
```



4-H Study of Positive Youth Development (4H.csv)

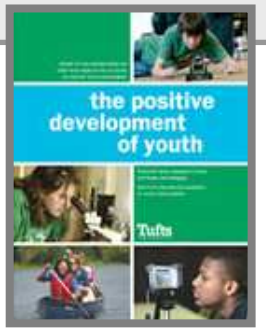


- 4-H Study of Positive Youth Development
- Source: Subset of data from IARYD, Tufts University

- Sample: These data consist of seventh graders who participated in Wave 3 of the 4-H Study of Positive Youth Development at Tufts University. This subfile is a substantially sampled-down version of the original file, as all the cases with any missing data on these selected variables were eliminated.
- Variables:

Depressed	0 = No (1-15 on Depression)
	1 = Yes (16+ on Depression)
PeerSupp	Peer Support

4-H Study of Positive Youth Development (4H.csv)




```
> the.model <- glm(Depressed ~ PeerSupp, family=binomial("logit"))
> summary(the.model)

Call:
glm(formula = Depressed ~ PeerSupp, family = binomial("logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3349  -0.8312  -0.6320   1.2224   1.8489

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.6113     0.6311   2.553   0.0107 *
PeerSupp     -0.6242     0.1512  -4.127 3.67e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 478.24  on 408  degrees of freedom
Residual deviance: 460.69  on 407  degrees of freedom
AIC: 464.69

Number of Fisher Scoring iterations: 4
```