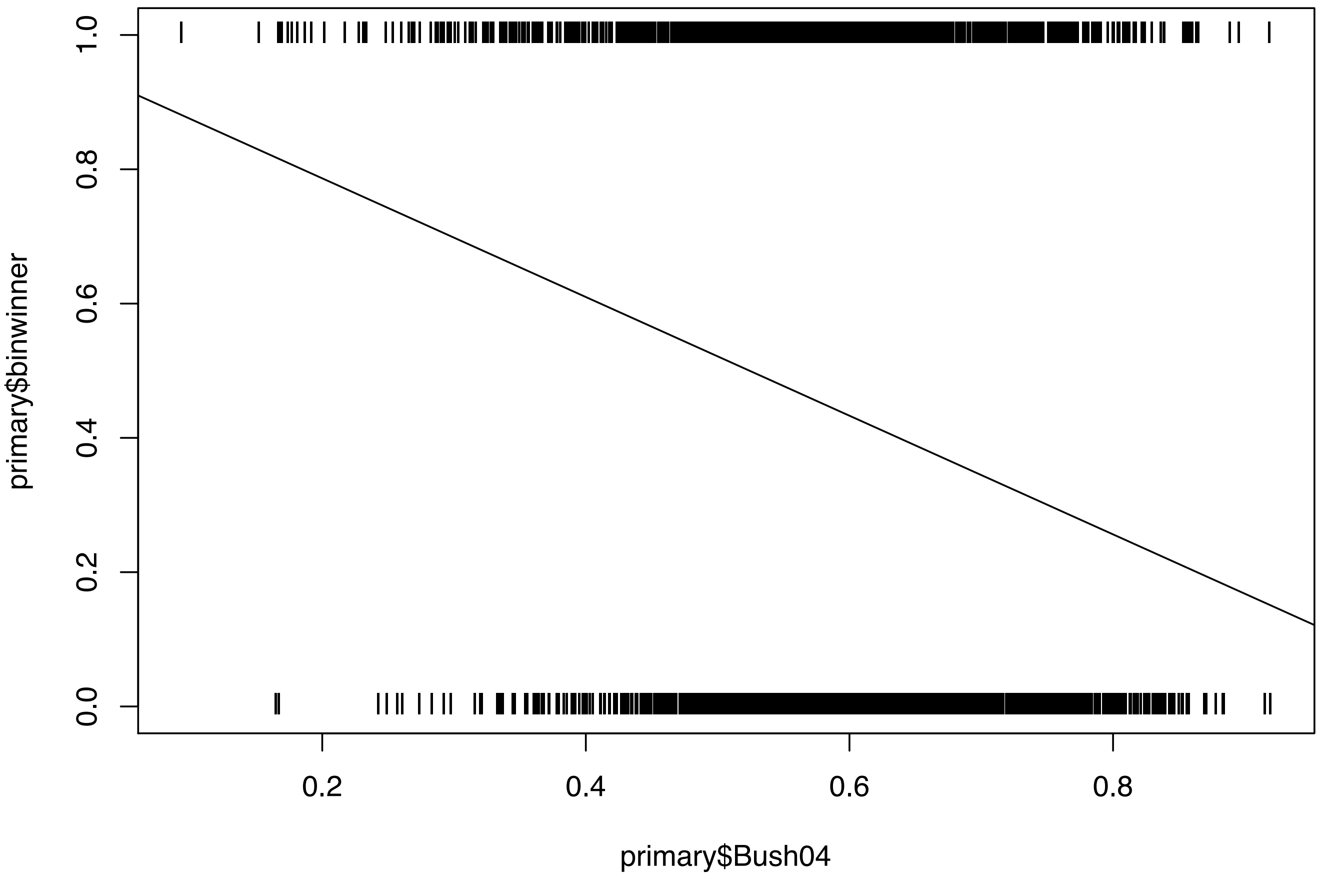
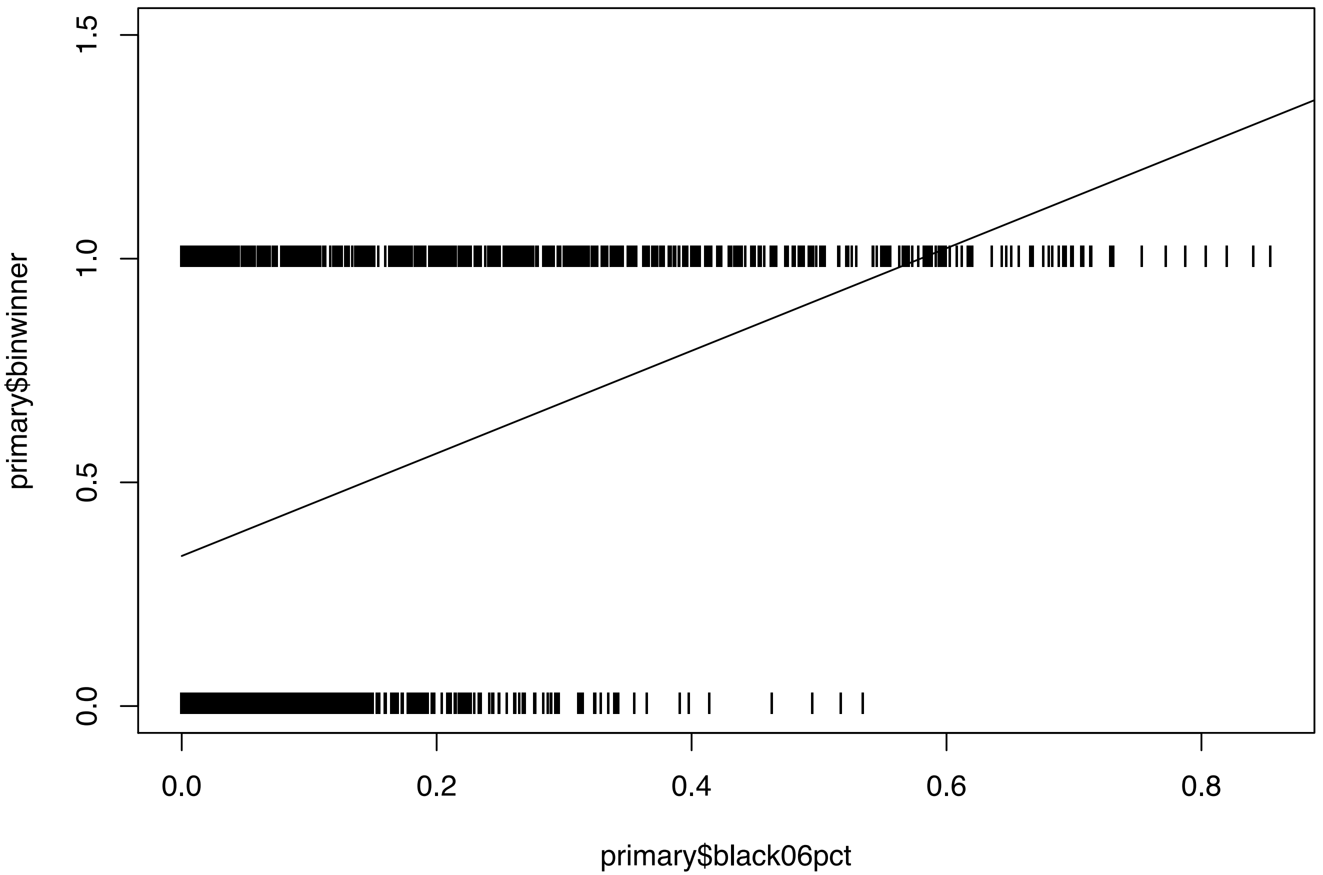
The Democratic primary

As a first pass at a model, we can try simply fitting a regression to 1/0 data; that is, use OLS to estimate the coefficients in a linear model where Obamma/ Clinton is recoded 1/0

In the case of the margin by which Bush won in 04, the predictions don’t, on the face of it, look miserable; we should probably try to see if there are ways to probe the fit and see if it, in fact, makes sense

For the percentage of the county that was African American, however, we have a different picture...


primary$binwinner
0.0
0.2
0.4
0.6
0.8
1.0
0.2
0.4
0.6
0.8
primary$Bush04


primary$binwinner
primary$black06pct
0.0
0.5
1.0
1.5
0.0
0.2
0.4
0.6
0.8

Some formality

A statistical model begins with a (hypothetical) description of how our data were generated; we use probability to express the uncertainties involved

In both of our examples (the primary and the NYT data set), our response is binary, taking only the values 0 and 1; let’s imagine that our observations are the result of a coin toss, where the properties of the coin depend on our covariates

Keeping (largely) the notation from the linear model, we could say that our observation y has a Bernoulli distribution with “success” probability p(x), where x represents our covariates

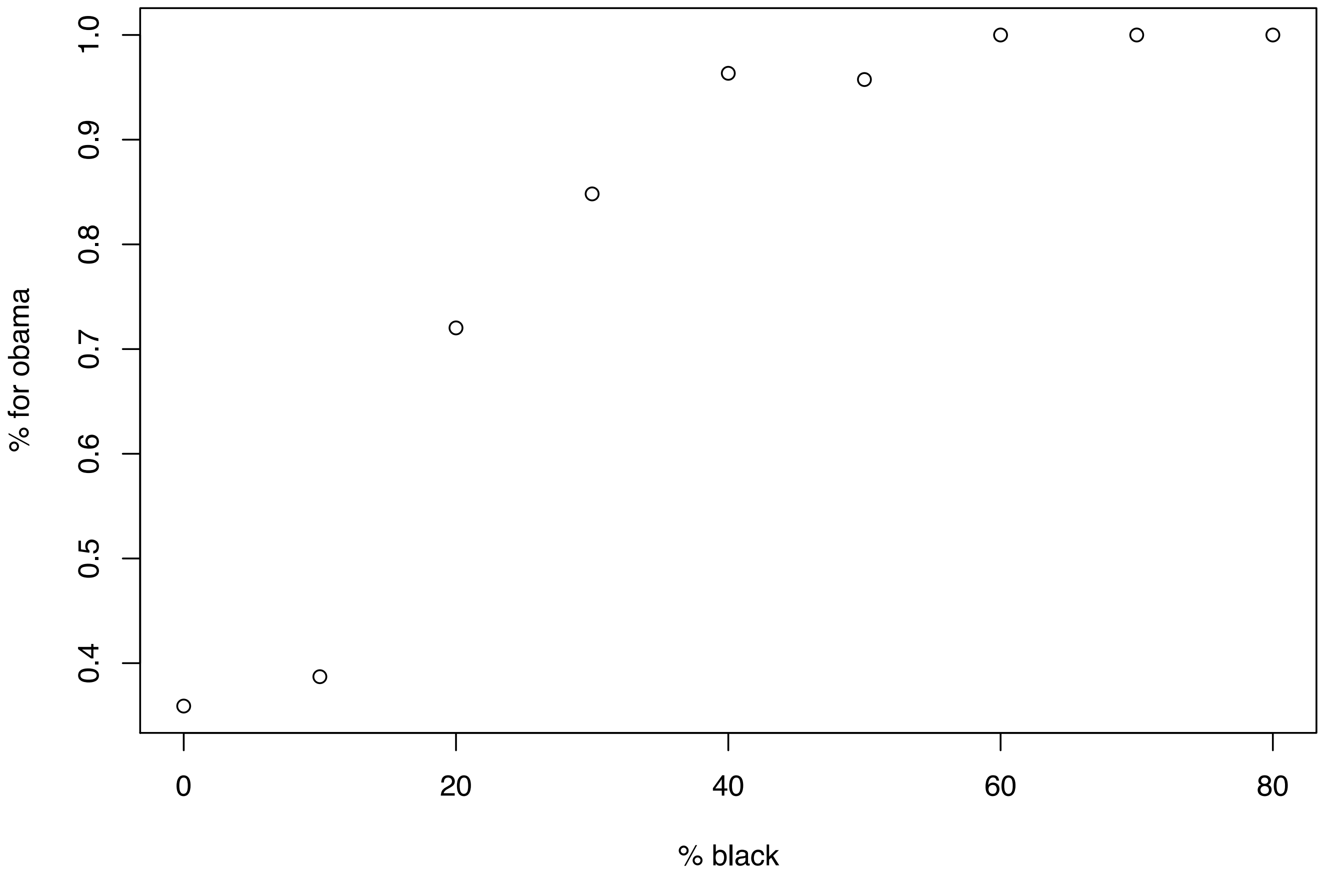
Implications

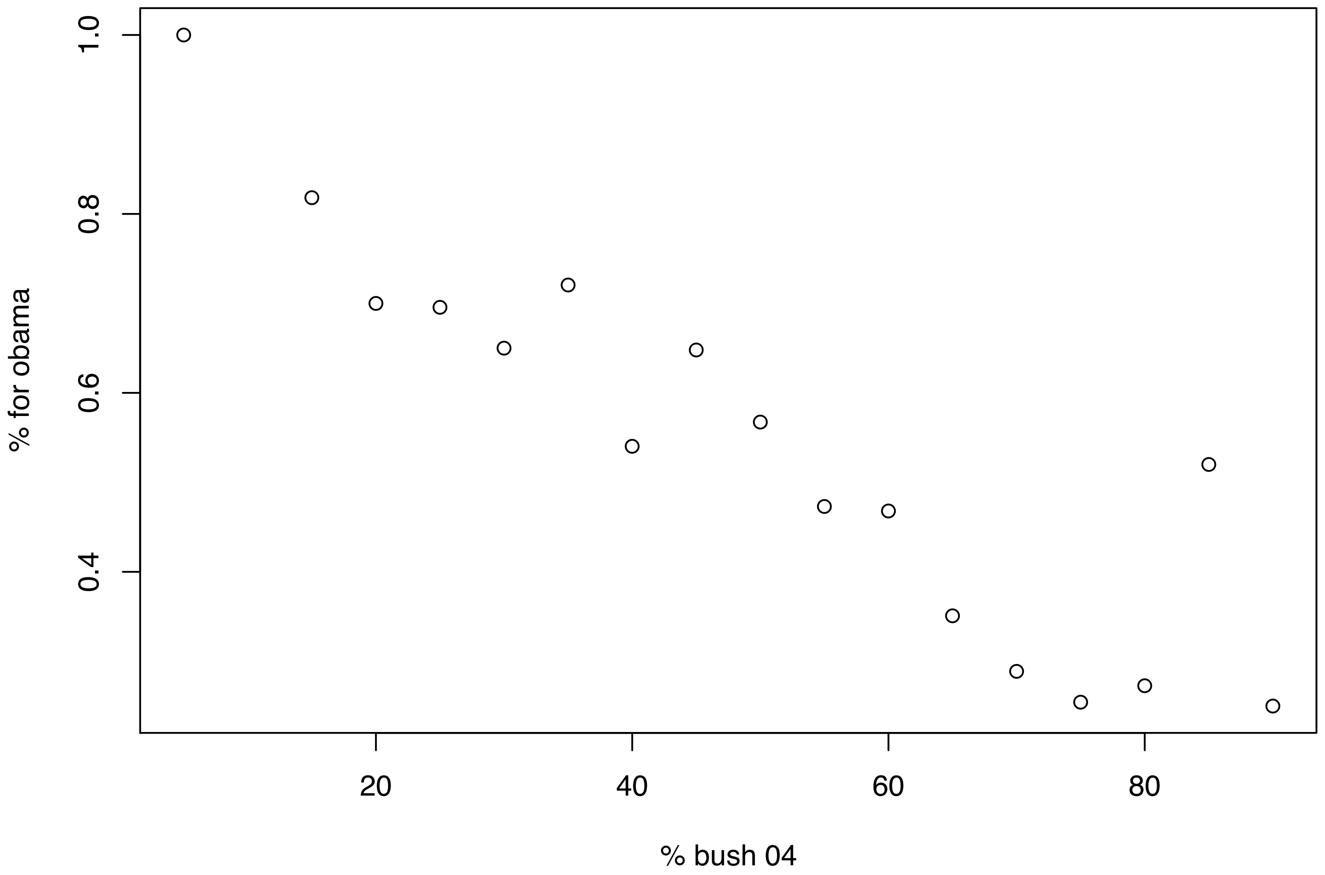
With that single modeling decision, we can now have a better look at our data

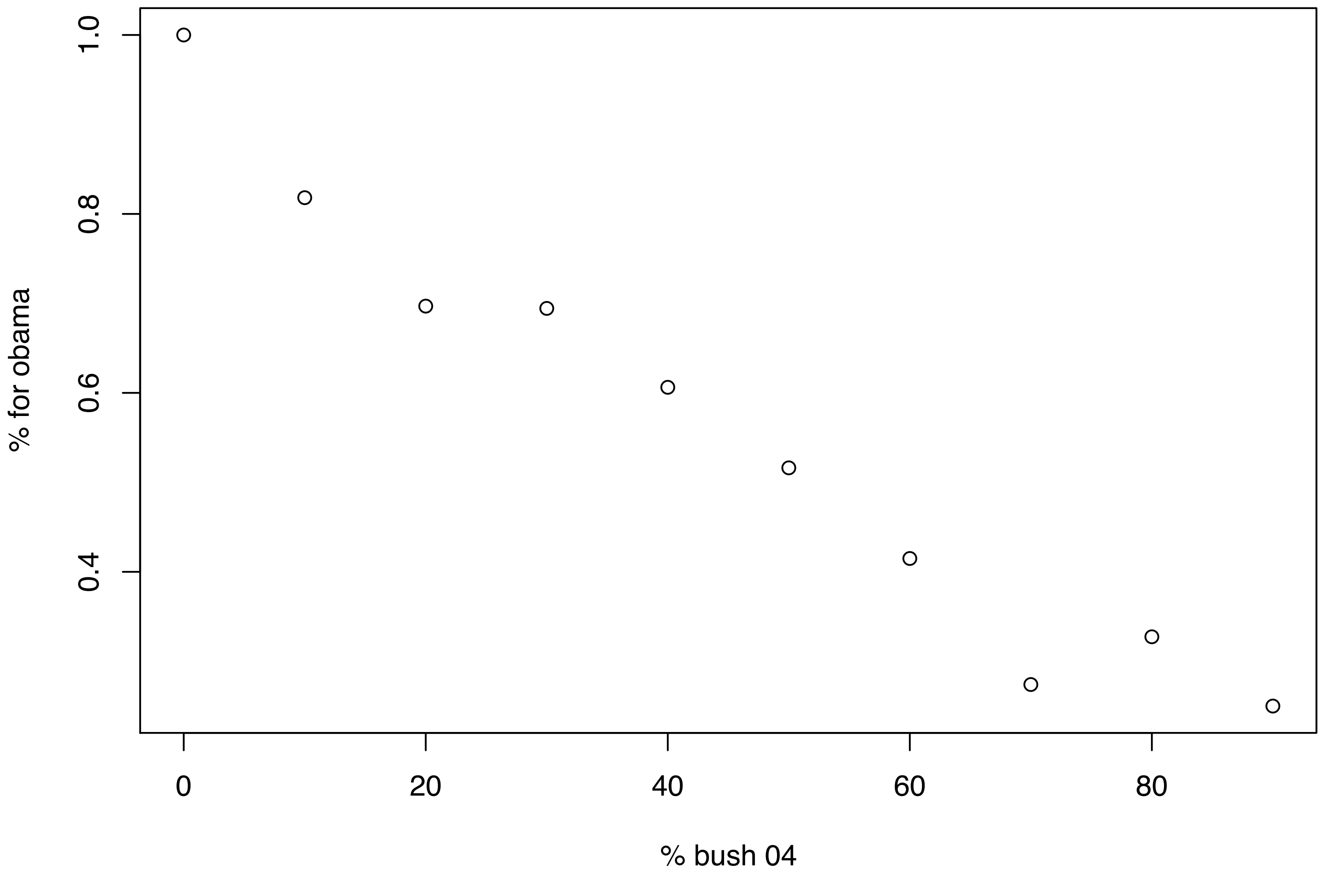
1. Divide the input data into a series of equally-sized bins

2. Within each bin, compute the proportion of successes

3. Plot the proportions against the midpoints of the bins


% for obama
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0
20
40
60
80
% black

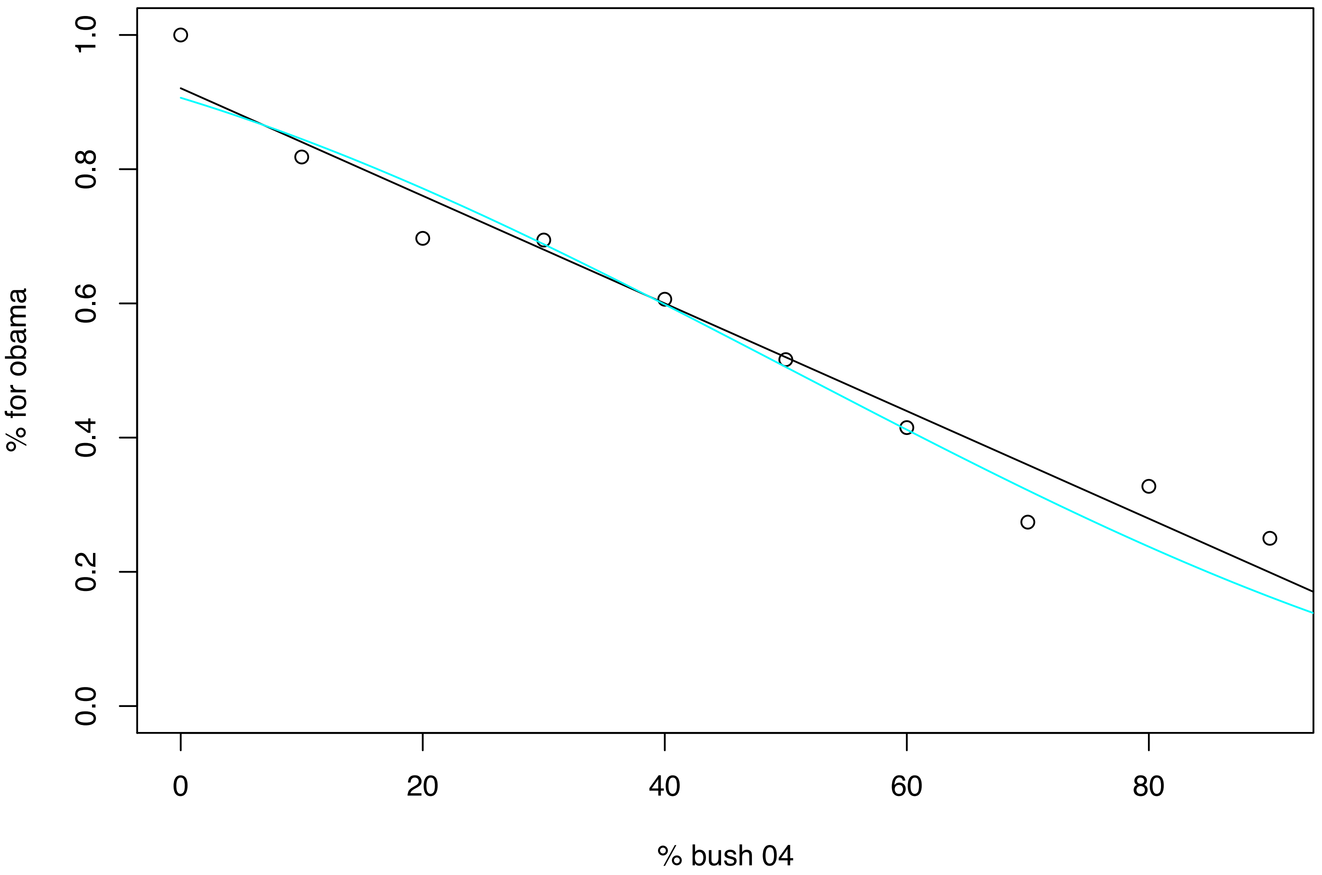

% for obama
1.0
0.8
0.6
0.4
20
40
60
80
% bush 04


% for obama
1.0
0.8
0.6
0.4
0
20
40
60
80
% bush 04

Binning

If we assume that p(x) is not varying very much within each bin, then the data points we are seeing are each of the form X/n where X has a binomial distribution with n (the number of points in the bin) and p

An obvious idea is to try to model these using a regression equation - -What does that amount to? Do we anticipate any problems here?


% for obama
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
% bush 04

```
table(primary$winner)

#     0    1
#  1210 1031

tbl = table(primary2$winner,primary2$Bush04,dnn=c("winner","bush"))

tbl

#              bush
#  winner    0    1
#       0  171 1039
#       1  302  728
```

Starting over

Let’s start over, this time working a bit more closely with the underlying probabilities

In our data set, 1031/2241 = 0.46 counties were won by Obama (keep in mind that he is ahead in both the popular vote and in the number of states won)

We define the odds of Obama winning a county to be

$$\mathrm{odds} = \frac{\mathrm{p}}{1-\mathrm{p}} = \frac{0.46}{0.54} = 0.85$$

Odds

Now let’s consider separately the odds for those counties that George Bush won in 2004

Among those counties that Bush won in 2004, the probability of an Obama win is 728/(728+1039) = 0.41; the associated odds are

$$\text{odds} = \frac{p}{1 - p} = \frac{0.41}{0.59} = 0.69$$

Among those counties that Bush lost in 2004, the probability of an Obama win is 302/(171+302) = 0.64; the associated odds are

$$\text{odds} = \frac{p}{1 - p} = \frac{0.64}{0.36} = 1.78$$

Odds ratio

To compare the odds associated with two populations, it is common to consider the odds ratio

In our primary example, we compute

$$\mathsf{oddsratio} = \frac{\mathsf{odds_l}}{\mathsf{odds_w}} = \frac{1.78}{0.69} = 2.58$$

In other words, the odds of Obama winning a county is over 2 and a half times greater if Bush lost that county in 2004

Odds ratio

Essentially, an odds ratio quantifies whether one group or experimental condition has higher or lower odds for achieving some binary outcome

In terms of regression terminology...

A number greater than one means positive association between the “independent” variable and the “dependent” variable (Obama winning in 2008 versus Bush in 2004)

A number between zero and one means negative association between independent and dependent variables

If the odds ratio equals one, the independent variable carries no information about the dependent variable

Log-odds

While odds are fairly direct quantities, it is often desirable to work with something a little more symmetric

$$\text{log-odds} = \log \frac{p}{1-p}$$

This transformation is also known as the logit; unlike the original probabilities, the logit can take any value between minus infinity and infinity

In the same spirit, it is often common to work with the log-odds ratio; assume we have probabilities of an event under two conditions p_0 and p_1, then

$$\text{log-odds ratio} = \log \frac{\frac{p_0}{1-p_0}}{\frac{p_1}{1-p_1}} = \operatorname{logit} p_0 - \operatorname{logit} p_1$$

Following our noses

Suppose we consider the following simple model for the probability of whether or not Obama wins a county

$$\text{logit}\,\mathsf{p} = \beta_0 + \beta_1\,\mathsf{bush}$$

where we take `bush` to be zero if Bush lost in 2004 and one otherwise; now we can relate this model to our log-odds ratio

$$\begin{aligned}\text{log-odds ratio} &= \text{logit}\,\mathsf{p}_1 - \text{logit}\,\mathsf{p}_0\\ &= (\beta_0 + \beta_1 * 1) - (\beta_0 + \beta_1 * 0)\\ &= \beta_1\end{aligned}$$

Logistic regression

And before you know it, we've arrived; if we choose to express the dependence of p on our covariate bush, p = p(bush), through the model

$$\text{logit}\,\text{p} = \beta_0 + \beta_1\,\text{bush}$$

then we can interpret the coefficient β_1 as a log-odds ratio, measuring the effect of a Bush win in 2004 on the odds that Obama will win the same county in 2008

To sum up, we started by considering the odds and ultimately odds ratios and were led to their "logged" counterparts which gave rise to the logistic model

The interpretation of the coefficients in a logistic model comes directly from the log-odds ratios and is a natural scale for working with probabilities

Logistic regression

There is nothing stopping us from extending our simple model, from the “factor” involving Bush’s win to other independent variables; say, from

$$\mathsf{logit\ p} = \beta_0 + \beta_1\,\mathsf{bush}$$

to

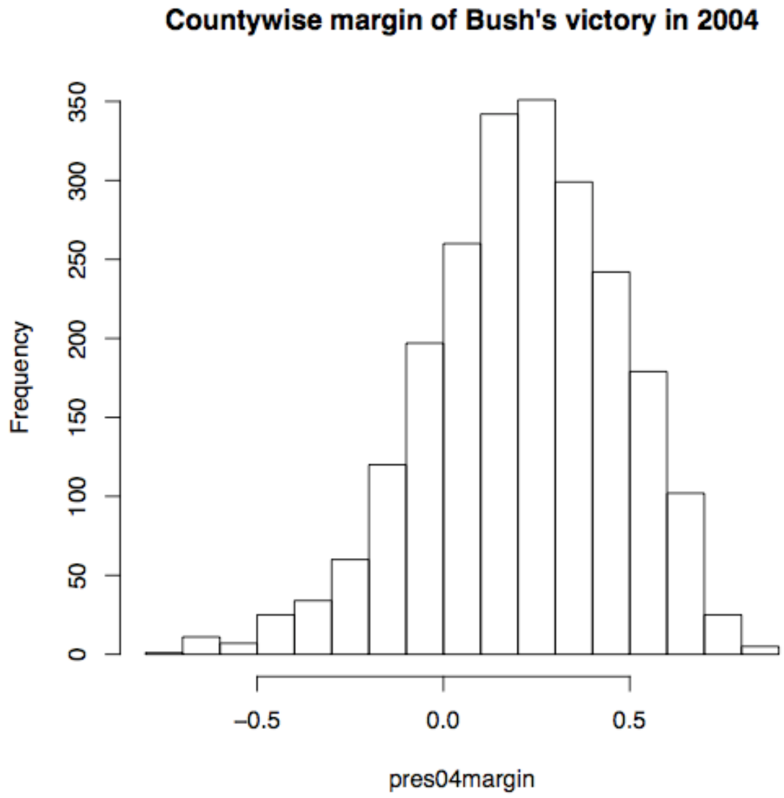
$$\mathsf{logit\ p} = \beta_0 + \beta_1\,\mathsf{bush} + \beta_2\,\mathsf{pct\_hs\_grad}$$

and so on...

Logistic regression

The variable `bush` was derived from the (continuous) variable pres04margin which holds the margin by which Bush won each county in 2004

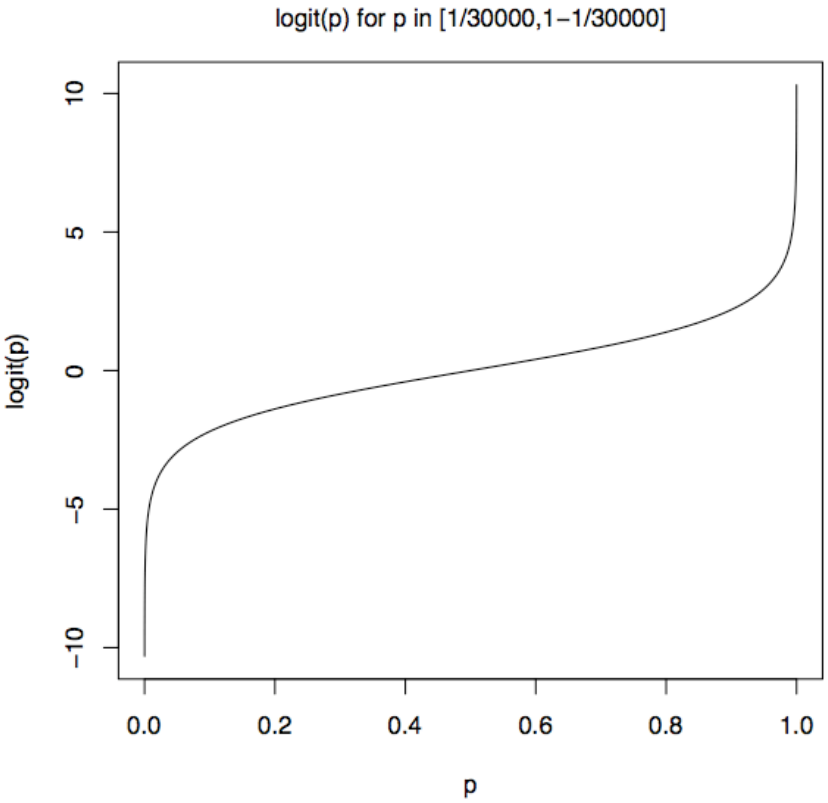
Let's now consider what our logit model implies for this continuous covariate; to get there, we need to know a little more about the logit



Logistic regression

At the right we plot the logit(p) as a function of p for a wide range of values; notice that the function has asymptotes at 0 and 1 (tending to $-\infty$ and ∞, respectively)

It should be clear from this plot that logit(p) is an invertible function...



The logit transform

... that is, for any value v $\in (-\infty, \infty)$, the p such that logit(p) = v can be found by first writing

$$\mathsf{logit}\,\mathsf{p} = \log \frac{\mathsf{p}}{1 - \mathsf{p}} = \mathsf{v}$$

and then taking the logarithm of both sides we have p/(1-p) = exp(v) or p = (1-p) exp(v); then, collecting terms p(1+exp v) = exp v, or

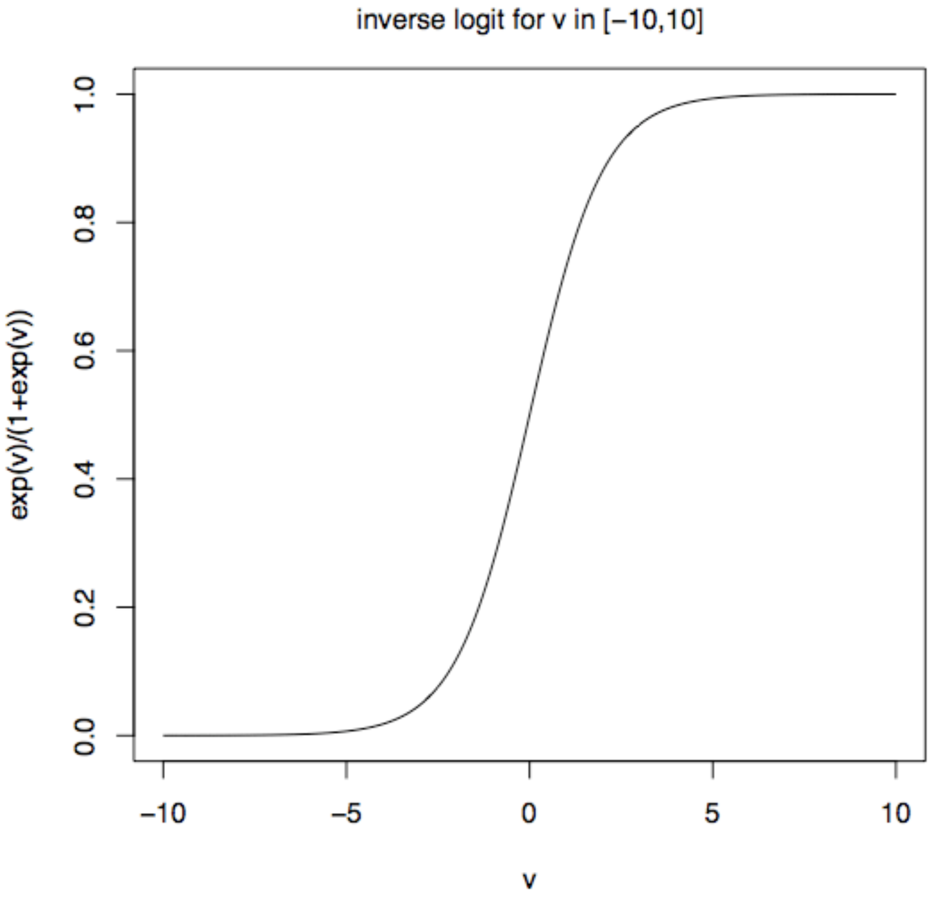
$$\mathsf{p} = \frac{\exp \mathsf{v}}{1 + \exp \mathsf{v}}$$

The logit transform

At the right we plot the inverse of the logit function; for very small (negative) values, it tends to zero, while for large values it tends to 1

A function that looks like this is often called a sigmoid function in some branches of mathematics and computer science; we might also recognize it as another kind of function... any guesses?

Now, let’s apply this to our “regression” context with a continuous covariate...



The logit transform

With these results in mind, suppose we choose to replace our model

$$\text{logit}\,\text{p}(\text{bush}) = \log\frac{\text{p}(\text{bush})}{1-\text{p}(\text{bush})} = \beta_0 + \beta_1\,\text{bush}$$

where bush was a binary covariate, with the expression

$$\text{logit}\,\text{p}(\text{pres04margin}) = \beta_0 + \beta_1\,\text{pres04margin}$$

where pres04margin is now continuous; for any values of the coefficients and the covariate, we invert the logit to give us

$$\text{p}(\text{pres04margin}) = \frac{\exp(\beta_0 + \beta_1\,\text{pres04margin})}{1+\exp(\beta_0 + \beta_1\,\text{pres04margin})}$$

The link

As an aside, in the context of logistic regression, we refer to the logit as a “link” function; it is the link that takes us from the scale of the covariates into the scale of our quantity of interest p

This idea of a link wasn’t really necessary when we studied the normal linear model; the nature of the variation in our data (y = f(x) + e) made it sensible to directly model

By comparison, when we used OLS to try to estimate the dependence of p on covariates, the lack of a link meant that our estimates were not constrained to be between 0 and 1; we’ll see this again in a few slides

Logistic regression

Therefore, we will imagine that in the primary race between Obama and Clinton, the probability that Obama wins a given county is determined by the toss of a coin

The probability that the coin is heads (indicating an Obama win and producing a 1 in our data set) is modeled by

$$\mathsf{p}(\mathsf{pres04margin}) = \frac{\exp(\beta_0 + \beta_1\,\mathsf{pres04margin})}{1 + \exp(\beta_0 + \beta_1\,\mathsf{pres04margin})}$$

Our data set consists of over 2,000 counties, where in each case we have a binary response $\mathsf{y} \in \{0, 1\}$ paired with the value of the covariate, pres04margin, which indicates the margin of Bush's victory in the county in 2004

We then form estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$, chosen that the resulting fit "captures the patterns" of wins and losses we see in our data; this is vague, but for now it's enough to say that there is an underlying measure that $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are chosen to minimize

Logistic regression

On the next page we present a plot of the 2008 county election results (1 if Obama won, 0 if Clinton won) against the margin for Bush in 04

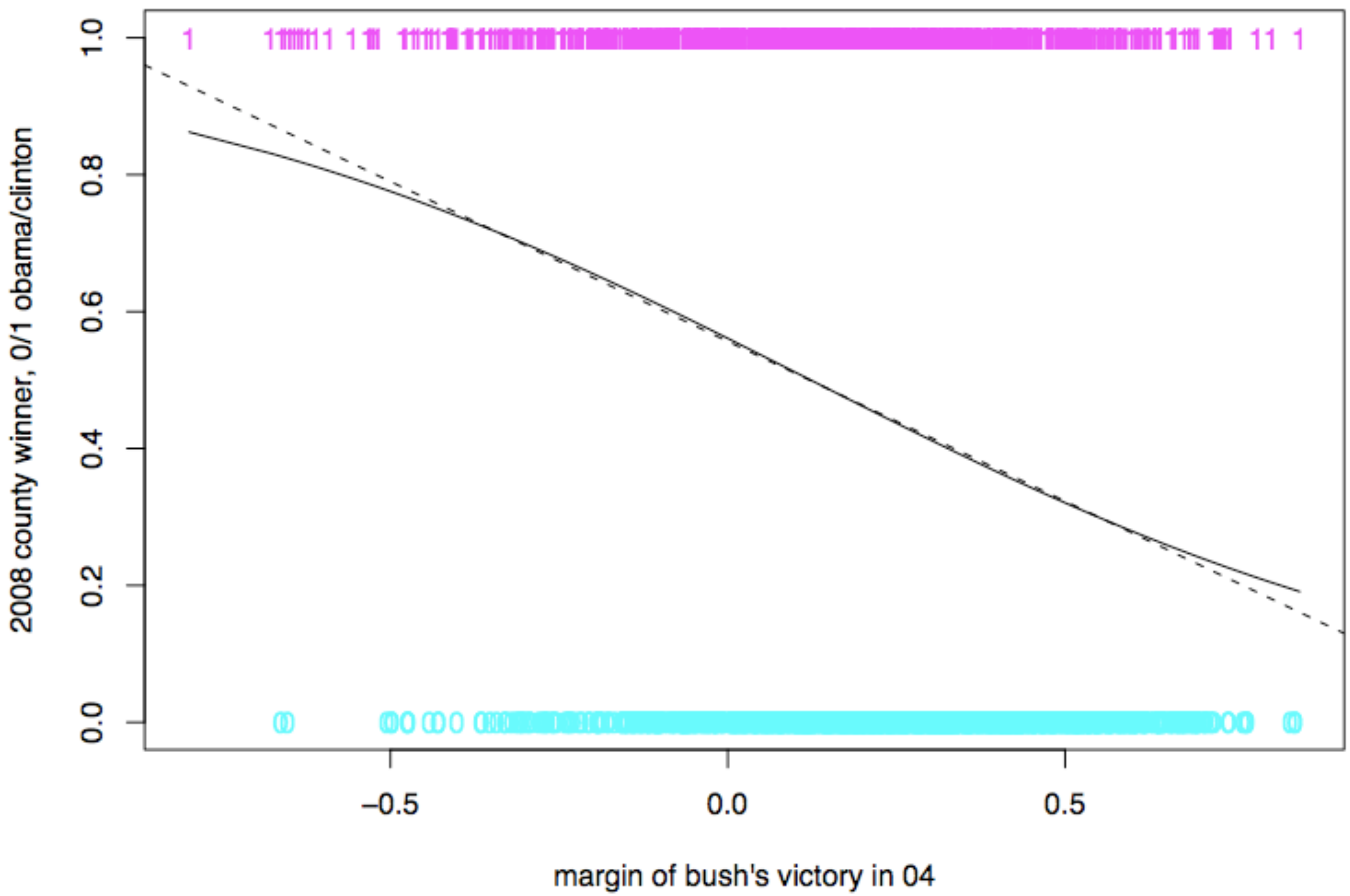
The solid curve in the middle is the result of “fitting” the relationship

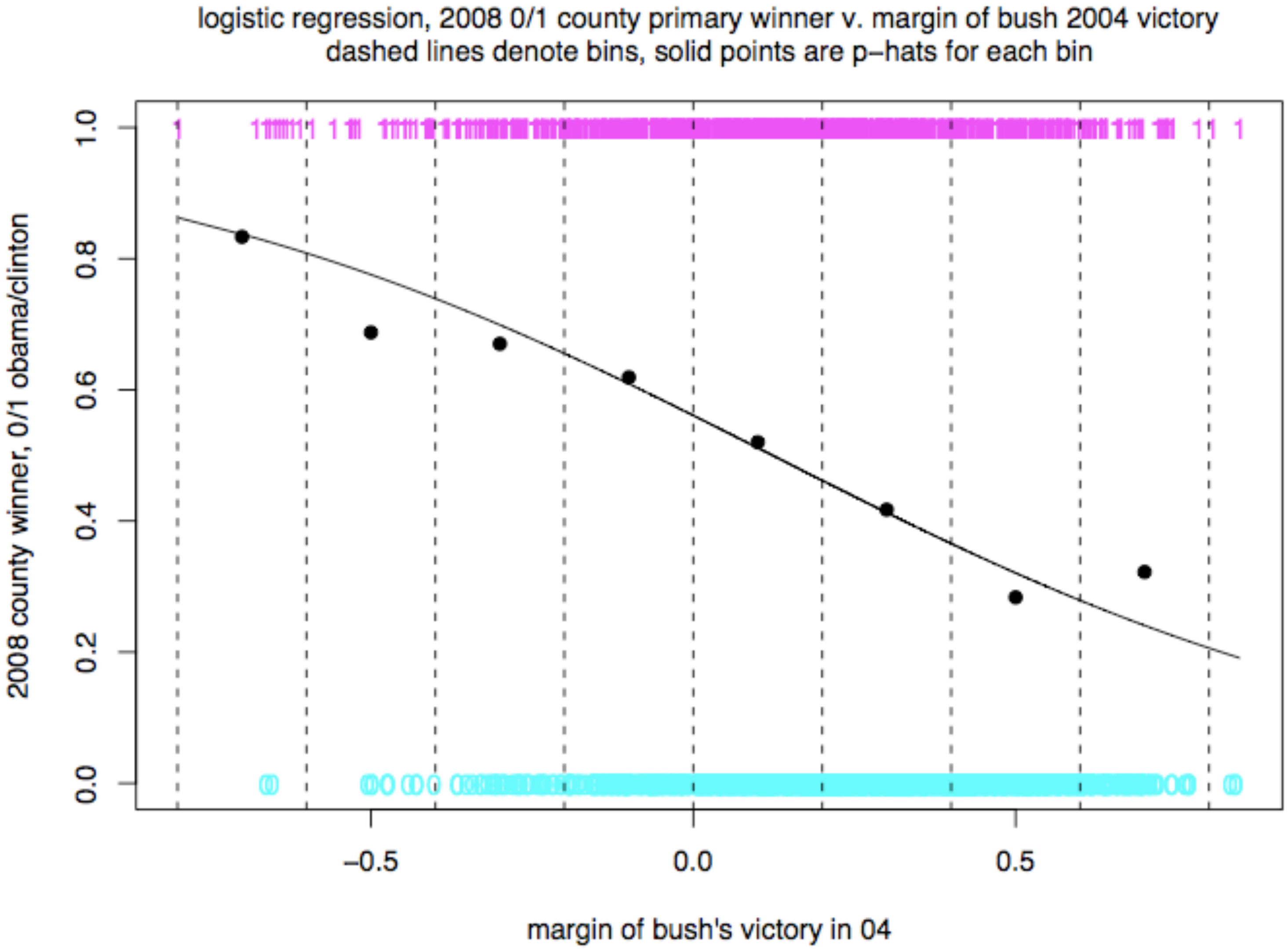
$$\mathsf{logit\ p}(\mathsf{pres04margin}) = \beta_0 + \beta_1\, \mathsf{pres04margin}$$

based on the county data in primary; the dashed line is the OLS fit to the 0/1 data (this is, in some sense, the limit of the binning we did last time)

What can you say about the fits?

2008 0/1 county primary winner v. margin of bush 2004 victory
logistic regression, solid line; ols, dashed





logistic regression, 2008 0/1 county primary winner v. margin of bush 2004 victory
dashed lines denote bins, solid points are p−hats for each bin

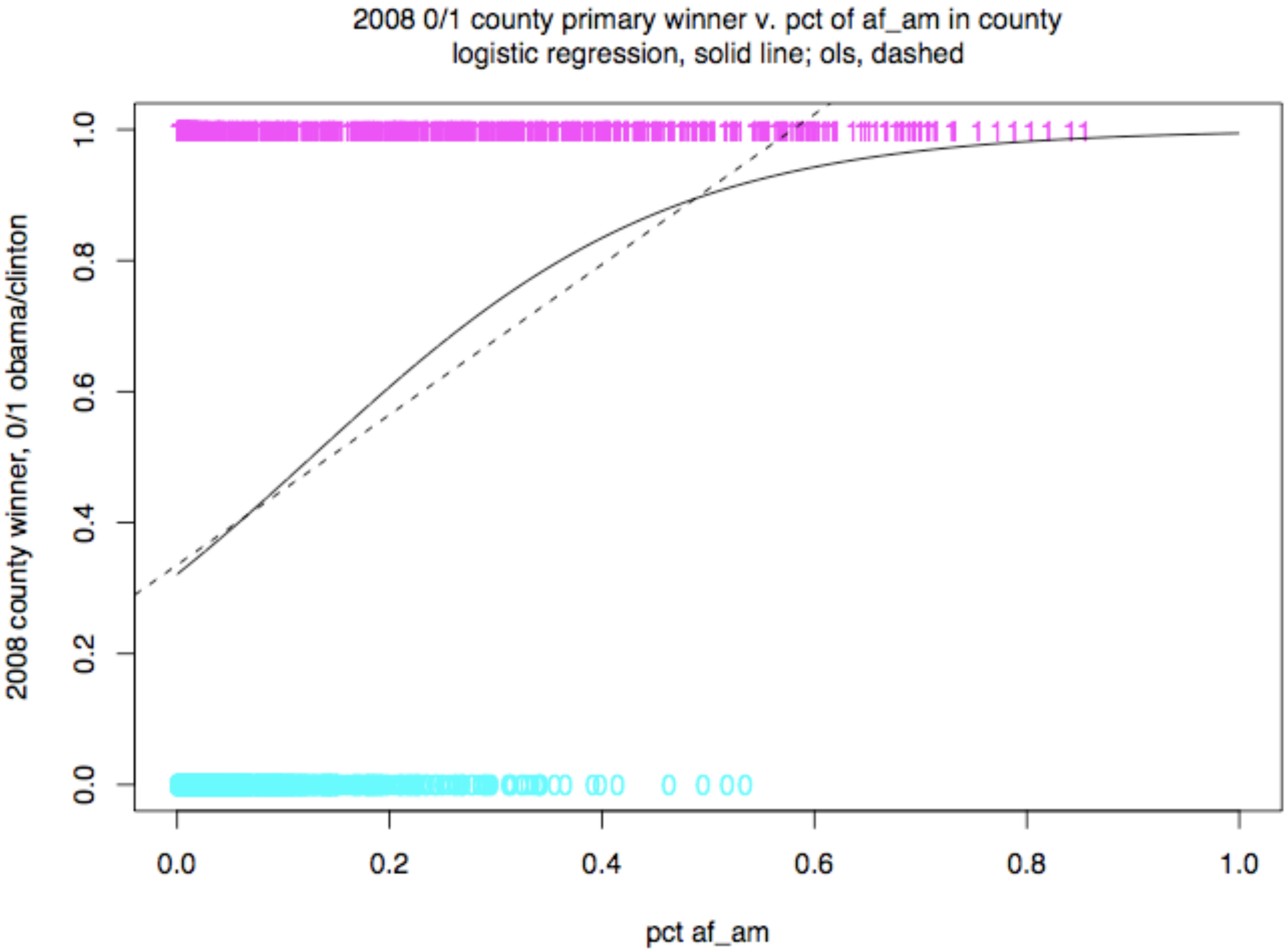
Logistic regression

Fitting produces estimates $\widehat{\beta}_0 = 0.24$ and $\widehat{\beta}_1 = -1.99$; the negative coefficient on pres04margin means that if we compare a margin of x to a margin of x + 0.1, the odds of an Obama win drop by a factor of

$$\text{exp}(\widehat{\beta}_1 * 0.1) = \text{exp}(-0.2) = 0.8$$

Perhaps not surprisingly, we can compute standard errors for these estimates (if there's one thing we're good at, it's assessing uncertainty); for $\widehat{\beta}_0 = 0.24$, the SE is 0.06 and for $\widehat{\beta}_1 = -1.99$ the SE is 0.18, implying that both effects are statistically significant

On the next page, we present some sample R code, although you don't have enough detail about what's going on to really grasp all of it yet; for the moment, you are meant to see how similar things are to the normal linear model


2008 0/1 county primary winner v. pct of af_am in county
logistic regression, solid line; ols, dashed
2008 county winner, 0/1 obama/clinton
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
pct af_am

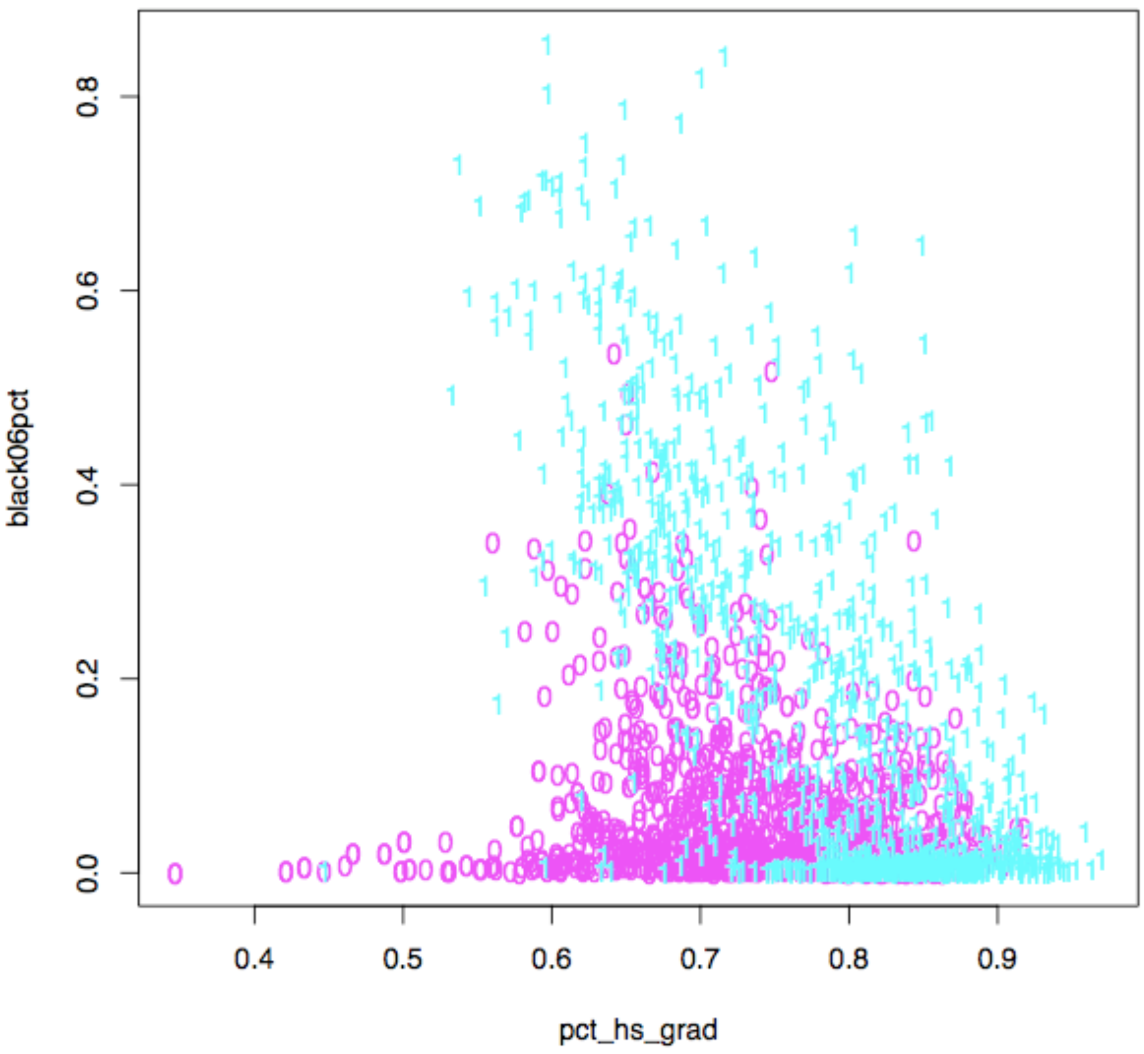
Logistic regression

Suppose we want to consider two variables simultaneously (and once we have seen two, there's nothing stopping us from getting into a lot of trouble)

We will consider the last two single-variable fits we performed; namely a model in which

$$\mathsf{logit}\,\mathsf{p} = \beta_0 + \beta_1\,\mathsf{pct\_hs\_grad} + \beta_2\,\mathsf{black06pct}$$

In the next slide, we plot the data with symbols that indicate who won each county, cyan for Obama and magenta for Clinton (for a very stupid reason, note that we have swapped colors in this plot)

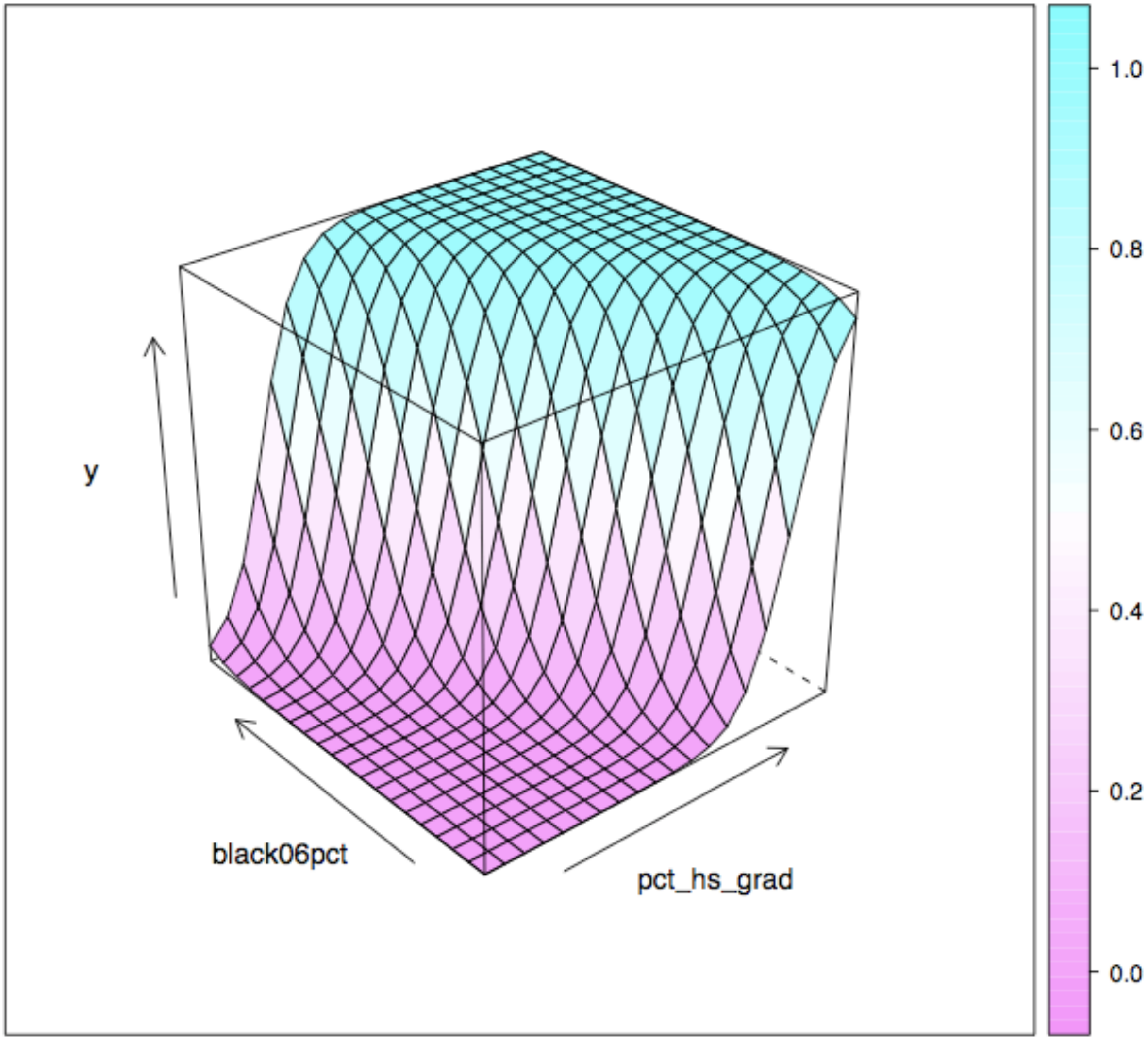

black06pct
0.0
0.2
0.4
0.6
0.8
pct_hs_grad
0.4
0.5
0.6
0.7
0.8
0.9
0
1

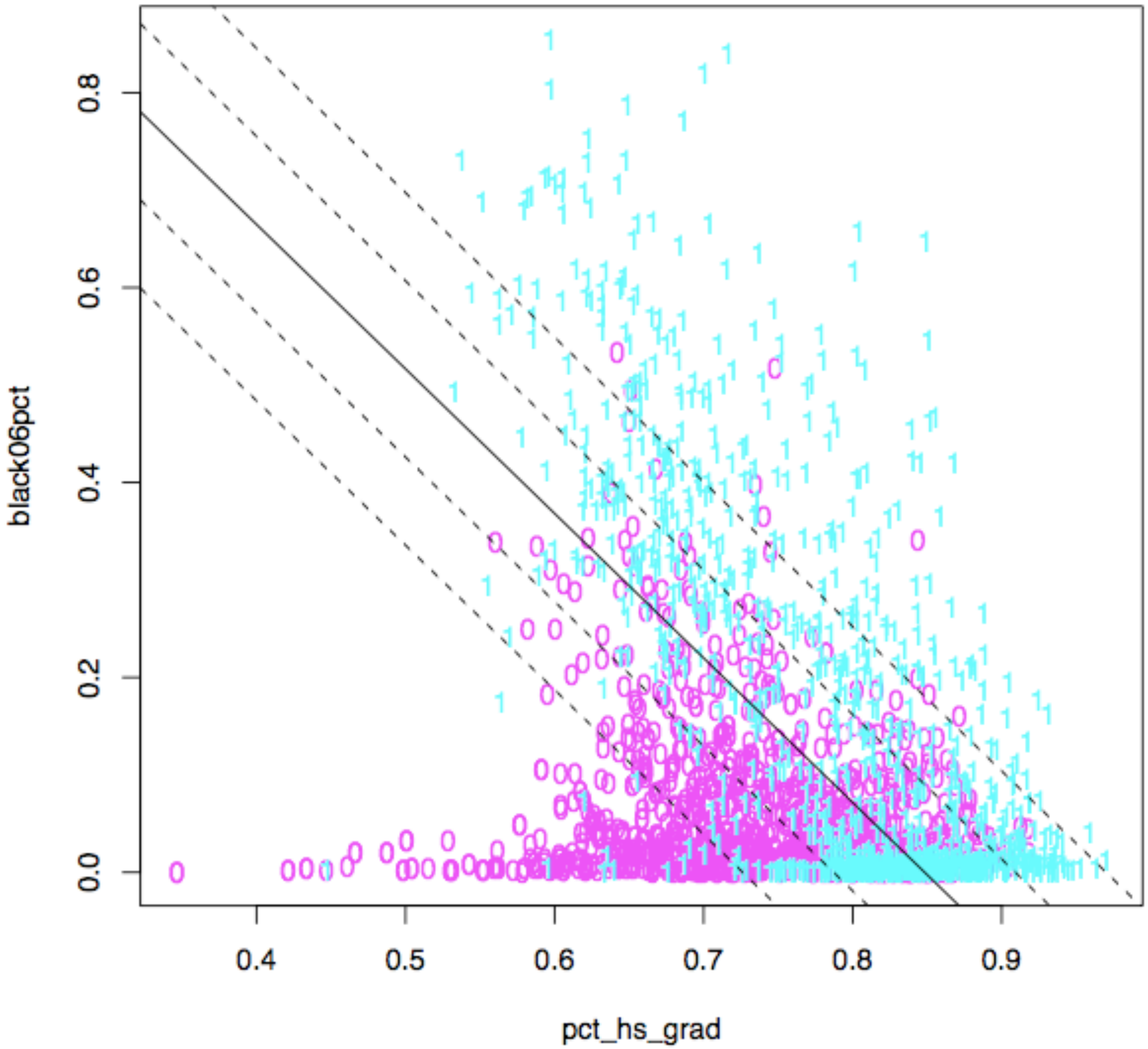
Logistic regression

Given the form of the model, we would expect the fitted probability surface to have contours that correspond to the lines

$$\mathsf{c} = \beta_1\, \mathsf{pct\_hs\_grad} + \beta_2\, \mathsf{black06pct}$$

In the next slide we present the fitted surface followed by a repeat of the covariate scatterplot with contour lines corresponding to the 0.1, 0.25, 0.5, 0.75 and 0.9 levels superimposed


y
black06pct
pct_hs_grad
1.0
0.8
0.6
0.4
0.2
0.0


0.0
0.2
0.4
0.6
0.8
black06pct
0.4
0.5
0.6
0.7
0.8
0.9
pct_hs_grad
0
1

Logistic regression

To compute the line associated with the 90% chance of an Obama victory, we computed

$$\operatorname{logit} 0.9 = \log \frac{0.9}{0.1} \approx 2.2$$

and then rewrote the following to express `pct_hs_grad` as a function of black06pct

$$2.2 = \widehat{\beta}_0 + \widehat{\beta}_1 \,\mathsf{pct\_hs\_grad} + \widehat{\beta}_2 \,\mathsf{black06pct}$$

Interpreting the fit

We began this part of our logistic regression adventure by considering odds and we saw that our fitted coefficients have interpretations as a log-odds ratio

Assuming our model is worth something, we can interpret the signs of the coefficients -- Positive coefficients mean increasing values of the covariates lead to larger odds and in turn larger probabilities

We can say a little more, at least approximately...

```
> source(url("http://www.stat.ucla.edu/~cocteau/primary3.R"))
> fit <- glm(winner~.,family="binomial",data=primary3)
> summary(fit)

Call:
glm(formula = winner ~ ., family = "binomial", data = primary3)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-3.3568  -0.6639  -0.1736   0.7064   3.7790  

Coefficients:
               Estimate Std. Error z value Pr(>|z|)    
(Intercept)   -12.03054    2.38004  -5.055 4.31e-07 ***
regionNE       -1.75933    0.24636  -7.141 9.25e-13 ***
regionS        -0.69357    0.20587  -3.369 0.000755 ***
regionW         1.51874    0.22424   6.773 1.26e-11 ***
pres04margin   -1.22034    0.41304  -2.955 0.003131 ** 
pct_less_30k   -7.22899    1.20294  -6.009 1.86e-09 ***
pct_more_100k  -8.21992    2.18772  -3.757 0.000172 ***
pct_hs_grad    14.11779    1.49400   9.450  < 2e-16 ***
pct_homeowner   1.65847    0.94561   1.754 0.079455 .  
black06pct     17.52698    1.71814  10.201  < 2e-16 ***
hisp06pct       3.03879    1.71333   1.774 0.076126 .  
white06pct      2.50931    1.63798   1.532 0.125533    
Bush04         -0.05407    0.21547  -0.251 0.801853    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3089.3  on 2238  degrees of freedom
Residual deviance: 1931.2  on 2226  degrees of freedom
AIC: 1957.2

Number of Fisher Scoring iterations: 5
```

Divide-by-four

To help interpret the coefficients of a logistic regression, suppose we have a simple linear model (on the logit-scale)

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

The point of greatest change occurs where $\beta_0 + \beta_1 x = 0$, or at p(x) = 0.5; it turns out that the slope of the curve is greatest here as well and that its maximum is

$$\frac{\beta_1 e^0}{(1+e^0)^2} = \frac{\beta_1}{4}$$

Therefore, $\beta_1/4$ gives you an upper bound to the change in probability you'll see after a 1-unit increase in the predictor -- The advantage here is that log-odds, for all the "naturalness" I've tried to convince you of, can be hard for people to interpret