

# Quantitative and Formal Methods in IR

## Lecture 6: Multiple Regression and Categorical Association

Dave Armstrong

University of Oxford  
Department of Politics and International Relations  
Center for Research Methods in the Social Sciences  
Nuffield College

February 9, 2008

1/36

## Outline

Model Fit for Multiple Regression

Relationships between Categorical Variables

2/36

## Adjusted $R^2$

When adding variables to a model (assuming the sample stays the same)  $R^2$  will never decrease and will usually increase, if only slightly. This makes the utility of  $R^2$  somewhat questionable.

What we would like is a measure of fit that got worse if we added a variable that didn't add anything to the model. Enter - adjusted  $R^2$ .

$$\text{Adjusted } R^2 = 1 - \frac{\frac{\text{Error Sum of Squares}}{n-1}}{\frac{\text{Total Sum of Squares}}{n-k-1}}$$

Where the total and error sums of squares are as defined previously. Here  $k$  refers to the total number of *independent variables* included in the model.

When a variable is added to the model that does not add any explanatory power, adjusted  $R^2$  will usually go down.

3/36

## F-Test

The F-test is another statistical test that allows us to test whether multiple regression coefficients are simultaneously zero. This is different from testing whether each one is statistically different from zero independently.

Stata reports an F-statistic and p-value. This is testing the null hypothesis that *all* of the regression coefficients are simultaneously zero.

$$H_0\beta_1 = \beta_2 = \dots = \beta_k = 0$$

It tests this against the alternative that at least one of the coefficients does not equal zero.

The F-statistic is generated the following way:

$$F = \frac{\frac{R^2}{k}}{\frac{1-R^2}{n-k-1}}$$

4/36

## Other uses of the F-test

The other main use of the F-test is to test multiple coefficients relating to a single concept.

For example, think back to when we added the the civilization dummy variables. If we want to know whether or not adding civilizations (as a concept) increased the fit of the model, we could use the F-test.

$$F = \frac{\frac{R_c^2 - R_r^2}{df1}}{\frac{1 - R_c^2}{df2}}$$

Where  $R_c^2$  is the  $R^2$  from the complete model (i.e., the one with all the variables included) and  $R_r^2$  is the  $R^2$  from the restricted model (i.e., the one with some of the variables excluded);  $df1$  is the number of variables excluded in the restricted model and  $df2$  is the residual degrees of freedom.

```
. reg new_polrts log_gdppc african-western
```

Source	SS	df	MS			
Model	73.4498214	6	12.2416369	Number of obs =	176	
Residual	84.5902338	169	.500533928	F( 6, 169) =	24.46	
Total	158.040055	175	.90308603	Prob > F =	0.0000	
				R-squared =	0.4648	
				Adj R-squared =	0.4458	
				Root MSE =	.70748	

new_polrts	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
log_gdppc	.161227	.0588314	2.74	0.007	.045088	.277366
african	-.292755	.1810181	-1.62	0.108	-.650103	.064593
islamic	-.9382975	.1592205	-5.89	0.000	-1.252615	-.6239803
latinam	.2534227	.1948103	1.30	0.195	-.1311524	.6379978
orthodox	-.2616364	.2284643	-1.15	0.254	-.7126478	.1893751
western	.6491538	.1847909	3.51	0.001	.284358	1.01395
_cons	-1.289391	.5107095	-2.52	0.013	-2.297583	-.2811991

```
. reg new_polrts log_gdppc
```

Source	SS	df	MS			
Model	31.0122955	1	31.0122955	Number of obs =	176	
Residual	127.02776	174	.730044596	F( 1, 174) =	42.48	
Total	158.040055	175	.90308603	Prob > F =	0.0000	
				R-squared =	0.1962	
				Adj R-squared =	0.1916	
				Root MSE =	.85443	

new_polrts	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
log_gdppc	.3593182	.0551299	6.52	0.000	.2505089	.4681276
_cons	-3.128798	.484976	-6.45	0.000	-4.085991	-2.171605

## Calculating F

$$F = \frac{\frac{R_c^2 - R_r^2}{df1}}{\frac{1 - R_c^2}{df2}}$$

$$= \frac{\frac{0.4648 - 0.1962}{5}}{\frac{1 - 0.4648}{169}}$$

$$= \frac{0.0537}{0.0032}$$

$$= 16.94$$

```
. dis Ftail(5, 169, 16.95)
1.449e-13
```

```
. test african islamic latinam orthodox western
```

- ( 1) african = 0
- ( 2) islamic = 0
- ( 3) latinam = 0
- ( 4) orthodox = 0
- ( 5) western = 0

```
F( 5, 169) = 16.96
Prob > F = 0.0000
```

## Contingency Tables

We cannot use regression to figure out whether or not there is a relationship between categorical variables (remember, we need a continuous dependent variable for that).

Rather, we can use a *contingency table* to figure out whether there is a relationship between two categorical variables. A contingency table is a cross-classification of observations. It shows the number of observations existing in every combination of values on two variables.

Example:

```
. tab mid4 dembin
```

mid4	dembin		Total
	0	1	
0	10,698	3,979	14,677
1	2,965	1,490	4,455
Total	13,663	5,469	19,132

This is the cross-tabulation of democracy and militarized interstate disputes from 1816-2001.

## Making the Cross-Tabulation

I'm going to depart a bit from the book because this way makes more sense to me. To remember which variable goes where, the dependent variable goes in the rows (on the  $y$ -axis) and the independent variable goes in the column (on the  $x$ -axis).

The language of dependent and independent variable is a bit misplaced here because in cross-tabs, the measure of association we're interested in is symmetric - meaning that it is the same regardless of which variable is in the rows and which is in the columns. However, the interpretation will be slightly different.

9/36

## Interpreting the Cross-Tab

```
. tab mid4 dembin
```

mid4	dembin		Total
	0	1	
0	10,698	3,979	14,677
1	2,965	1,490	4,455
Total	13,663	5,469	19,132

We percentage-ize down the columns and compare across the rows.

- Percentage-izing the columns means dividing each entry in the column by the column total. So the upper-left cell would be  $10698/13663$  and the lower-left cell entry would be  $2965/13663$ . The upper right and lower right would be  $3979/5469$  and  $1490/5469$ , respectively.

```
. tab mid4 dembin, col nofreq
```

mid4	dembin		Total
	0	1	
0	78.30	72.76	76.71
1	21.70	27.24	23.29
Total	100.00	100.00	100.00

10/36

## Comparing Across the Rows

```
. tab mid4 dembin, col nofreq
```

mid4	dembin		Total
	0	1	
0	78.30	72.76	76.71
1	21.70	27.24	23.29
Total	100.00	100.00	100.00

What we want to know is: Of the non-democracies, what percentage had MIDs and of the democracies, what percentage had MIDs.

- If those two percentages are the same, then whether or not a country has a MID does not depend on its level of democracy (i.e., the two variables are independent of each other).
- If the two percentages are not the same, then whether a country has a MID does depend on its level of democracy (i.e., the two variables are dependent).

You can probably imagine that we want to do inference on these percentages, much as we did with a difference of proportions test.

11/36

## Substantive Interpretation

The interpretation here is a bit less precise than in the linear regression model. It is rather more like correlation.

- One thing we can do is look for the biggest column percentage in each row. In the above table, a greater percentage of democracies, than non-democracies have had MIDs over the last 200(ish) years.
- This suggests that over this period, democracies are relatively more likely to have MIDs than non-democracies.
  - Notice, that this does not mean that democracies are more likely to have MIDs than not in the absolute sense, only relative to non-democracies.

12/36

## Inference for Cross-Tabs

We want to know whether the percentages in the rows are *close enough* to each other that the difference could be due to sampling or are they sufficiently different for us to infer that there is a difference in the population.

We can go through our five steps:

1. Assume that we have a random sample and sufficiently many observations that the CLT works. Further, assume there is a sufficient number of observations in each cross-classified group (e.g., 5).
2. Generate Hypotheses

$$H_0 : \quad \text{Statistical Independence}$$

$$H_A : \quad \text{Not Statistical Independence}$$

## More Inference for Cross-Tabs

3. Test Statistic: Let's label the parts of the cross-tab as follows:

	var 2		
var 1	0	1	
0	a	b	$f_{0+}$
1	c	d	$f_{1+}$
	$f_{+0}$	$f_{+1}$	$f_{++}$

Where  $f_{0+} = a + b$ ,  $f_{1+} = c + d$ ,  $f_{+0} = a + c$ ,  $f_{+1} = b + d$  and  $f_{++} = a + b + c + d$ .

We can define the *expected* frequency in a cell as

$$f_e^a = \frac{f_{+0} \times f_{0+}}{f_{++}} \quad ; \quad f_e^b = \frac{f_{+1} \times f_{0+}}{f_{++}}$$

$$f_e^c = \frac{f_{+0} \times f_{1+}}{f_{++}} \quad ; \quad f_e^d = \frac{f_{+1} \times f_{1+}}{f_{++}}$$

## More Inference for Cross-Tabs

3. Define:

$$\chi^2 = \sum_{j \in \{a,b,c,d\}} \frac{(j - f_e^j)^2}{(f_e^j)^2}$$

4. We then compare  $\chi^2$ , our statistic to a  $\chi^2$  distribution with degrees of freedom equal to the number of rows in the table minus 1 times number of columns in the table minus 1:  $(n_r - 1) \times (n_c - 1)$ . This produces a p-value, which is the probability of observing a  $\chi^2$  statistic as big or bigger than the one we find if the two variables were statistically independent.

5. Conclusion:

- If the resulting p-value is less than or equal to 0.05, then we reject  $H_0$  that the two variables are statistically independent.
- If the resulting p-value is greater than 0.05, then we do not reject  $H_0$  and we infer that the two variables are statistically independent.

## Properties of the $\chi^2$ Statistic

- If  $j$  (the observed frequency) is equal to  $f_e^j$ , the expected frequency, for each cell, then the  $\chi^2$  statistic will be zero, so this is the lower bound of  $\chi^2$ .
- The upper bound requires a bit more thought.  $\chi^2$  will be the biggest when, for example,  $b = c = 0$ . That is two of the cells (either those on the diagonal or those on the off-diagonal) equal zero. Then, the upper bound, if I've done my math(s) right, is:

$$\frac{a^2(a-1)^2 + 2a^2d^2 + d^2(d-1)^2}{(a+d)^2}$$

- $\chi^2$  does not measure the *strength* of the association, only whether it exists or not. Here, the strength is more of a substantive concern - that is to say, we might want to know not just whether there are statistically "big" relative differences, but also how big these relative differences are likely to be.

For this task, we will need *logistic regression*.

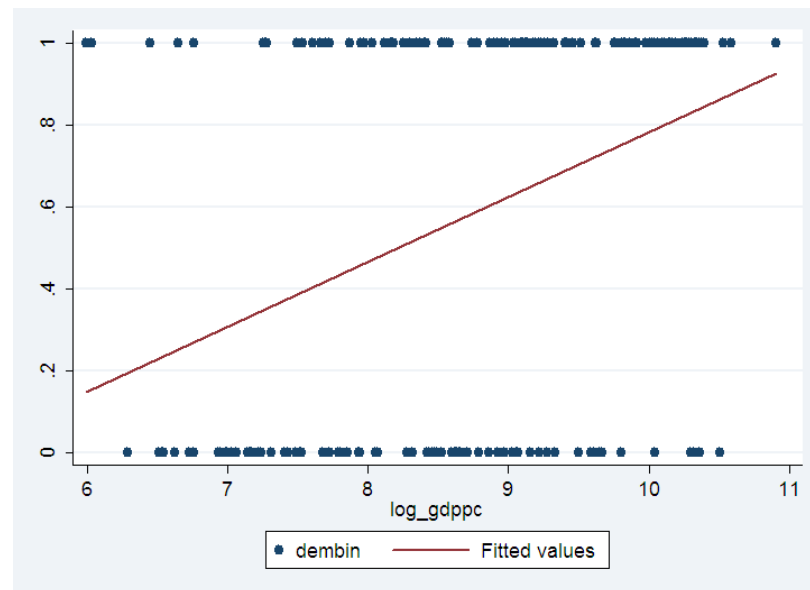
## Why Can't We Use Linear Regression on a Binary Dependent Variable

Remember back to the Assumptions we made:

1. Random Sampling
2. Linearity
3. Mean of errors is zero
4. Variance of errors is constant,  $\sigma^2$ .
5. Distribution of errors is normal.
6. Errors are uncorrelated with the  $x$  variable(s).

17/36

### Plot of Democracy (binary) with log(GDP/capita)



18/36

## The Goal

Remember, we said that OLS regression is just a sophisticated difference of means test. That is to say, we were trying to get the best guess of what the mean of  $y$  would be given all of the information we had on  $x$ .

The goal now is to make our best guess of the  $Pr(y = 1)$  given all of the information we have on  $x$ . So, we need something like regression, but that does for difference of proportions what OLS does for difference of means.

- Realizing that a proportion is simply the mean of a binary variable, we could start with linear regression and see how far we get.
- We will see that OLS falls short on a number of grounds, so we will eventually move on to something else.

19/36

## Linear Probability Model and Predictions

The model above, where we model  $Pr(y = 1) = \alpha + \beta x + \varepsilon$  is called a *linear probability model*.

One of the big problems here is that that the model *can* generate predictions that are outside the range  $[0, 1]$ . If we think about the prediction as the probability that  $y=1$ , then predictions outside the range  $[0, 1]$  do not make sense theoretically.

We need to make sure that whatever model we settle on only produces theoretically reasonable predictions.

20/36

## Assumptions

Let's think back to the graph of the relationship between binary democracy and  $\log(\text{GDP/capita})$  and consider the OLS Assumptions.

1. There is no particular problem with the random sampling assumption here or correlation between  $e$  and  $x$ .
2. The linearity assumption is arguably not appropriate here, as the relationship does not look especially linear. However, as it turns out, this is not our biggest problem.
3. Is the mean of the errors everywhere zero? Probably not, especially at the extremes of the plot. Since the errors will be skewed, the mean of the residuals is probably not everywhere zero.
4. Is the variance of the errors constant? Probably not, the variance of the errors might be quite small at the extremes and bigger in the middle.
5. Is the distribution of the errors normal? Probably not, the errors in the middle of the plot will be bimodal, with modes around  $-.5$  and  $.5$ . At the extremes, the distribution will likely be skewed.

21 / 36

## What Should the Shape of the Relationship Be?

Think about income per month and the probability of owning a home.

- Let's think about a person's income going from \$0/month to \$1000/month. His overall income is still quite low and this would probably have very little effect on his ability to buy a house (i.e., the probability of owning a home shouldn't change much).
- Now, what about someone who makes \$10000/month. She is probably quite likely to own a home already. An extra \$1000/month is probably not going to make a big difference for her (i.e., she might get a bit more likely to own a home, but not much).
- Now, what about someone going from, let's say \$3000/month to \$4000/month. This could make a relatively big difference in the probability of owning a home.

What we are suggesting here is that the effect of income depends on where you start.

22 / 36

## Modeling $Pr(y = 1)$

Let's think about how we could model  $Pr(y = 1)$  as a function of  $x$ . Remember we want to make sure that we can make only theoretically reasonable predictions.

- We could model  $\log(Pr(y = 1)) = \alpha + \beta x$ . Since  $Pr(y = 1) = [0, 1] \rightarrow \log(Pr(y = 1)) = [-\infty, 0]$ . This means that predictions above 0, would suggest  $Pr(y = 1) > 1$ , which doesn't make much sense.
- We could model what is called the *odds* of  $y = 1$ :  $\frac{Pr(y=1)}{1-Pr(y=1)} = \alpha + \beta x$ . This suffers from the same problem, but on the other end. The odds range from  $[0, \infty]$ . A bit of math shows that this implies:  $Pr(y = 1) = \frac{\alpha + \beta x}{1 + (\alpha + \beta x)}$ , so when  $\alpha + \beta x = -1$ ,  $Pr(y = 1)$  would equal infinity.
- What about the log of the odds? If  $\frac{Pr(y=1)}{1-Pr(y=1)} = [0, \infty] \rightarrow \log\left(\frac{Pr(y=1)}{1-Pr(y=1)}\right) = [-\infty, \infty]$ .

23 / 36

## The Logistic Regression Model

The logistic regression model implements the strategy suggested in the last point above:

$$\log\left(\frac{Pr(y = 1)}{1 - Pr(y = 1)}\right) = \alpha + \beta x$$

With a bit of math, we get the following equation for the probability:

$$Pr(y = 1) = \frac{1}{1 + \exp(-(\alpha + \beta x))}$$

Remember how we interpreted the effect of  $\log\_gdppc$  in the last class? This is the same strategy we will use here.

24 / 36

## Interpreting the Effect of Variables

We want to know how the  $Pr(y = 1)$  changes when we change  $x$ . What we can do is pick a few different values of  $x$  and then calculate the change in predicted probabilities. An example will help illustrate:

```
. logit dembin log_gdppc, nolog
-----+-----
Logistic regression              Number of obs   =        176
                                LR chi2(1)      =        25.67
                                Prob > chi2     =         0.0000
Log likelihood = -106.91983      Pseudo R2      =         0.1072
-----+-----

```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
log_gdppc	.7118738	.1516607	4.69	0.000	.4146243 1.009123
_cons	-5.844657	1.317828	-4.44	0.000	-8.427552 -3.261761

25/36

## Predictions

$$Pr(y = 1 | \log(G)) = \frac{1}{1 + \exp(-(-5.8 + .71 \log(G)))}$$

Let's consider the three sets of predictions we looked at last time, changing GDP/capita from 1000  $\rightarrow$  2000, 9000  $\rightarrow$  10000 and 50000  $\rightarrow$  51000.

$$\begin{aligned} Pr(Y = 1 | G = 1000) &= \frac{1}{1 + \exp(-(-5.8 + .71 \log(1000)))} \\ &= \frac{1}{1 + \exp(-(-5.8 + .71(6.91)))} \\ &= \frac{1}{1 + \exp(0.89)} \\ &= \frac{1}{1 + 2.43} \\ &= 0.29 \\ Pr(Y = 1 | G = 2000) &= \frac{1}{1 + \exp(-(-5.8 + .71 \log(2000)))} \\ &= \frac{1}{1 + \exp(-(-5.8 + .71(7.6)))} \\ &= \frac{1}{1 + \exp(0.40)} \\ &= 0.40 \end{aligned}$$

26/36

## More Predictions

$$\begin{aligned} Pr(Y = 1 | G = 9000) &= \frac{1}{1 + \exp(-(-5.8 + .71 \log(9000)))} \\ &= \frac{1}{1 + \exp(-(-0.66))} \\ &= 0.66 \\ Pr(Y = 1 | G = 10000) &= \frac{1}{1 + \exp(-(-5.8 + .71 \log(10000)))} \\ &= \frac{1}{1 + \exp(-(-0.73))} \\ &= 0.67 \\ Pr(Y = 1 | G = 50000) &= \frac{1}{1 + \exp(-(-5.8 + .71 \log(50000)))} \\ &= \frac{1}{1 + \exp(-(-1.88))} \\ &= 0.867 \\ Pr(Y = 1 | G = 51000) &= \frac{1}{1 + \exp(-(-5.8 + .71 \log(51000)))} \\ &= \frac{1}{1 + \exp(-(-1.90))} \\ &= 0.869 \end{aligned}$$

27/36

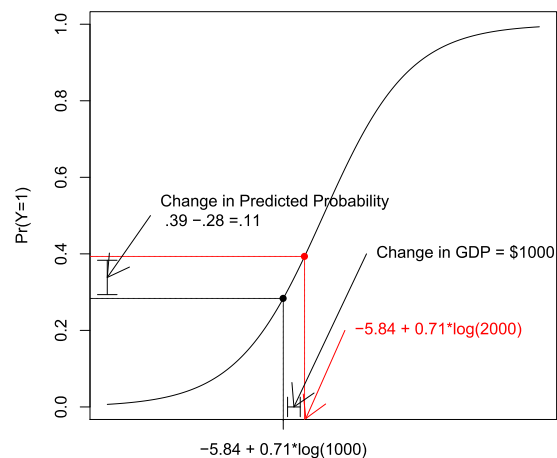
## Effect of GDP/capita

$x_0 \rightarrow x_1$	$Pr(y = 1   x_0)$	$Pr(y = 1   x_1)$	$\Delta$
1000 $\rightarrow$ 2000	0.29	0.40	0.11
9000 $\rightarrow$ 10000	0.66	0.67	0.01
50000 $\rightarrow$ 51000	0.867	0.869	0.002

Just as in last week's example, the effect of GDP/capita changes depending on where you start. Now, since this is a non-linear model, this is true of *all*  $x$  variables, regardless of whether or not they are themselves logged.

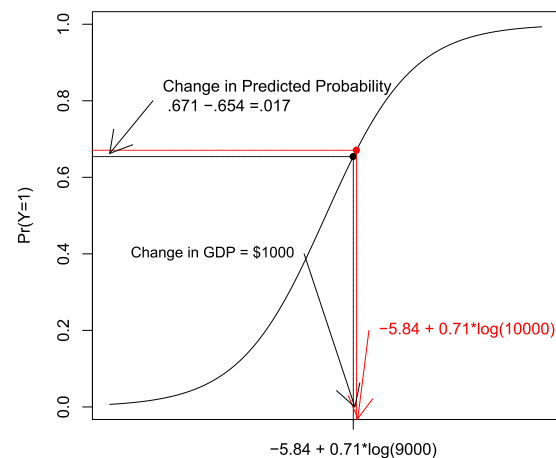
28/36

Let's see how this plays out visually: GDP 1000 → 2000



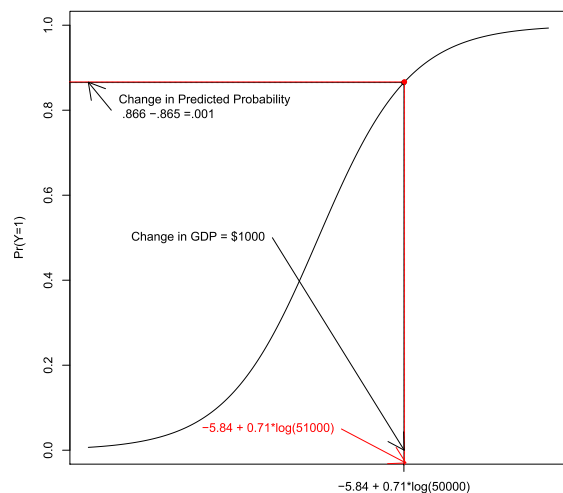
29/36

Let's see how this plays out visually: GDP 9000 → 10000



30/36

Let's see how this plays out visually: GDP 50000 → 51000



31/36

### Multiple Logistic Regression

Just as with linear regression, we can (and usually should) have more than one variable in our regression models. Let's think about including population in the model above:

```
. logit dembin log_gdppc log_pop
```

```
Iteration 0: log likelihood = -119.75714
Iteration 1: log likelihood = -105.28212
Iteration 2: log likelihood = -105.06641
Iteration 3: log likelihood = -105.06591
```

```
Logistic regression                Number of obs =      176
LR chi2(2) =                        29.38
Prob > chi2 =                       0.0000
Pseudo R2 =                         0.1227
```

```
Log likelihood = -105.06591
```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
log_gdppc	.6656546	.1535341	4.34	0.000	.3647333 .9665759
log_pop	-.1620674	.0862036	-1.88	0.060	-.3310234 .0068885
_cons	-4.032258	1.601727	-2.52	0.012	-7.171586 -.8929298

The interpretation is a bit different here. Now, the changes in probability are going to be conditional not only on where we start on GDP/capita, but *also* where we are holding population constant. Let's demonstrate:

32/36

$$Pr(Y = 1) = \frac{1}{1 + \exp(-0.67 * \log(\text{GDP/capita}) - 0.16 * \log(\text{Population}) - 4.03)}$$

Let's now see what going from \$1000 to \$2000 on GDP does with population at 100

$$\begin{aligned} Pr(Y = 1) &= \frac{1}{1 + \exp(-0.67 * \log(1000) - 0.16 * \log(100) - 4.03)} \\ &= \frac{1}{1 + \exp(-0.67 * 6.91 - 0.16 * 4.61 - 4.03)} \\ &= \frac{1}{2.15} \\ &= 0.465 \end{aligned}$$

$$\begin{aligned} Pr(Y = 1) &= \frac{1}{1 + \exp(-0.67 * \log(2000) - 0.16 * \log(100) - 4.03)} \\ &= \frac{1}{1 + \exp(-0.67 * 7.6 - 0.16 * 4.61 - 4.03)} \\ &= \frac{1}{1.73} \\ &= 0.58 \end{aligned}$$

33/36

Let's now see what going from \$1000 to \$2000 on GDP does at with population at 10000

$$\begin{aligned} Pr(Y = 1) &= \frac{1}{1 + \exp(-0.67 * \log(1000) - 0.16 * \log(10000) - 4.03)} \\ &= \frac{1}{1 + \exp(-0.67 * 6.91 - 0.16 * 9.21 - 4.03)} \\ &= \frac{1}{3.40} \\ &= 0.294 \end{aligned}$$

$$\begin{aligned} Pr(Y = 1) &= \frac{1}{1 + \exp(-0.67 * \log(2000) - 0.16 * \log(10000) - 4.03)} \\ &= \frac{1}{1 + \exp(-0.67 * 7.6 - 0.16 * 9.21 - 4.03)} \\ &= \frac{1}{2.51} \\ &= 0.40 \end{aligned}$$

34/36

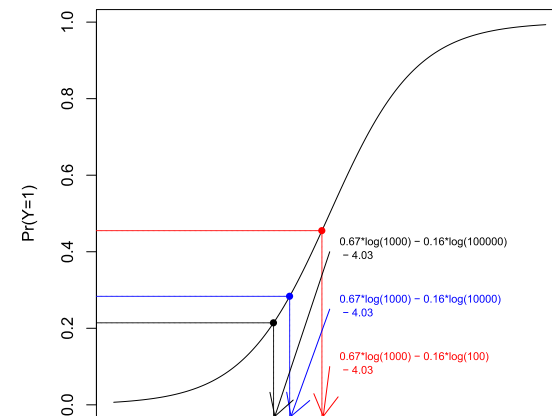
Let's now see what going from \$1000 to \$2000 on GDP does at with population at 100000

$$\begin{aligned} Pr(Y = 1) &= \frac{1}{1 + \exp(-0.67 * \log(1000) - 0.16 * \log(100000) - 4.03)} \\ &= \frac{1}{1 + \exp(-0.67 * 6.91 - 0.16 * 11.51 - 4.03)} \\ &= \frac{1}{4.46} \\ &= 0.22 \end{aligned}$$

$$\begin{aligned} Pr(Y = 1) &= \frac{1}{1 + \exp(-0.67 * \log(2000) - 0.16 * \log(100000) - 4.03)} \\ &= \frac{1}{1 + \exp(-0.67 * 7.6 - 0.16 * 11.51 - 4.03)} \\ &= \frac{1}{3.18} \\ &= 0.31 \end{aligned}$$

35/36

Let's see what is happening visually



36/36