Advanced Data Analysis I

Panel Data Part 1

PA 541 Week 13

Michael D. Siciliano

Department of Public Administration

College of Urban Planning and Public Affairs

Remaining coursework

- **Week 13/14** – Panel Data
- **Week 15** – Intro to DAGs & Review for Final
- **Week 16** – Final Exam (similar format to midterm)

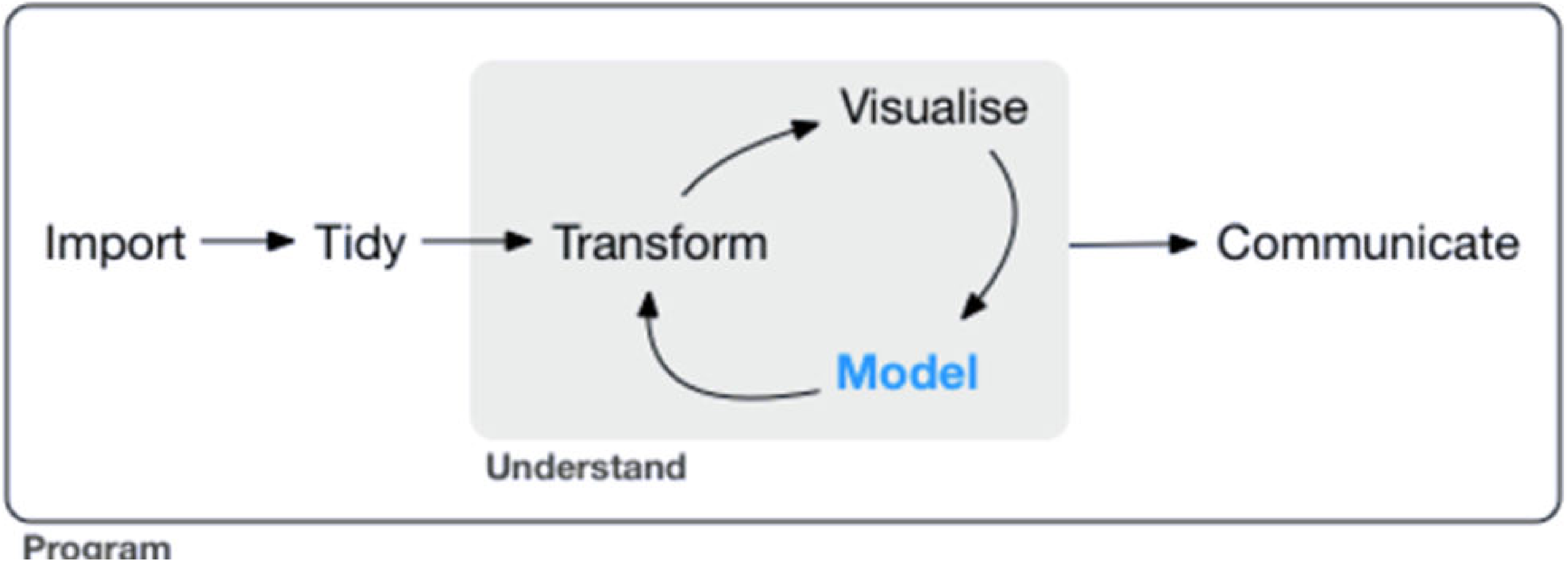
- **Homework 3 is due April 12th:** Covers non-linear relationships, logistic regression, and first part start of today's lecture.

- **Final Papers are due May 5th (for those who have chosen to submit)**

Today’s lecture

- Review logistic regression; go over in-class exercise from last week
- Overview of MLE
- Pooled Cross-Sectional Analysis
- Basic Difference in Difference Models

A look back


Import
Tidy
Transform
Visualise
Model
Communicate
Understand
Program

Odds Ratio

Link Function

Let's

Endogeneity

Logit

REVIEW

GLMs

Week 12

Linear Probability Models

Exponentiated coefficients

REVIEW LOGISTIC REGRESSION

Interpreting the Coefficients – Logged Odds

```
Call:
glm(formula = reject ~ pubrec + black + hispan + loanprc, family = binomial,
    data = loan)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2972  -0.4544  -0.4090  -0.3287   2.7762

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.1480     0.3661 -11.332  < 2e-16 ***
pubrec        1.7297     0.1991   8.687  < 2e-16 ***
black         1.2444     0.1860   6.691 2.21e-11 ***
hispan        0.8436     0.2540   3.321 0.000895 ***
loanprc       2.1399     0.4375   4.892 9.99e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Logged Odds: Again, these coefficients have the exact same interpretation as in OLS regression except that the units of the DV are now in logged odds.
- Note that these Betas can be negative – but in our example all predicators are positively related with the DV, loan rejection.

Interpreting the Coefficients - Odds

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  -4.1480     0.3661 -11.332  < 2e-16 ***
pubrec        1.7297     0.1991   8.687  < 2e-16 ***
black         1.2444     0.1860   6.691 2.21e-11 ***
hispan        0.8436     0.2540   3.321 0.000895 ***
loanprc       2.1399     0.4375   4.892 9.99e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> #to get the odds ratio's for the estimates we can say
> exp(coef(loan2))
(Intercept)      pubrec       black      hispan     loanprc 
 0.01579567  5.63876579  3.47101232  2.32467818  8.49892805 
```

- Odds: As stated before, a coefficient of 1 leaves the odds unchanged (i.e. it has no effect). A coefficient greater than 1 increases the odds of occurrence and a coefficient less then 1 decreases the odds of occurrence. The greater the distance from one in either direction, the greater the impact of the predictor variable.

- So for pubrec, compared to an individual who never filed for bankruptcy, an individual with at least one filing has an increase in odds of loan rejection by 5.6 times.

Interpreting the Coefficients – Odds cont…

```
> #to get the odds ratio's for the estimates we can say
> exp(coef(loan2))
(Intercept)      pubrec       black      hispan     loanprc 
 0.01579567  5.63876579  3.47101232  2.32467818  8.49892805 
```

- It is important to remember that the odds have a multiplicative effect. Lets assume a white person’s odds of rejection based on a set of predictors is 3:1. Thus, if we took those same predictors for a black person, the odds of rejection would be 3*3.471 = 10.413:1.

- Based on this, when we divide the odds of someone who is white by someone who is black (as long as the other predictors are the same) then the result is just Exp(B). More specifically, 10.413/3 = 3.471. Thus, the coefficient shows the ratio of odds for a one unit increase in the independent variable.

- So, if you wanted to calculate the change in odds for increasing loanprc by one and going from 0 to 1 on pubrec, you need to multiply 8.499*5.639. So the odds increase by 47.9.

Interpreting the Coefficients – Odds cont…

```
> #to get the odds ratio's for the estimates we can say
> exp(coef(loan2))
(Intercept)      pubrec       black      hispan    loanprc 
 0.01579567  5.63876579  3.47101232  2.32467818 8.49892805 
```

- Let’s look at this one other way.
- Assume we have two people:
  - Person A: No public record, white, and asking for a loan of 75%.
  - Person B: Public record, white, and asking for a loan of 75%.

$$\frac{Odds_{x1=1,x2=0,x3=0,x4=.75}}{Odds_{x1=0,x2=0,x3=0,x4=.75}} = \frac{\exp(\beta_0 + \ \beta_1 + \ \beta_4 * .75)}{\exp(\beta_0 + \ \beta_4 * .75)} = \exp(\beta_1)$$

Why effects (with regard to odds) are multiplicative in logistic regression

$$\ln\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 x_1 + \varepsilon \longrightarrow \frac{P_i}{1-P_i} = e^{\beta_0+\beta_1 x_1+\varepsilon}$$

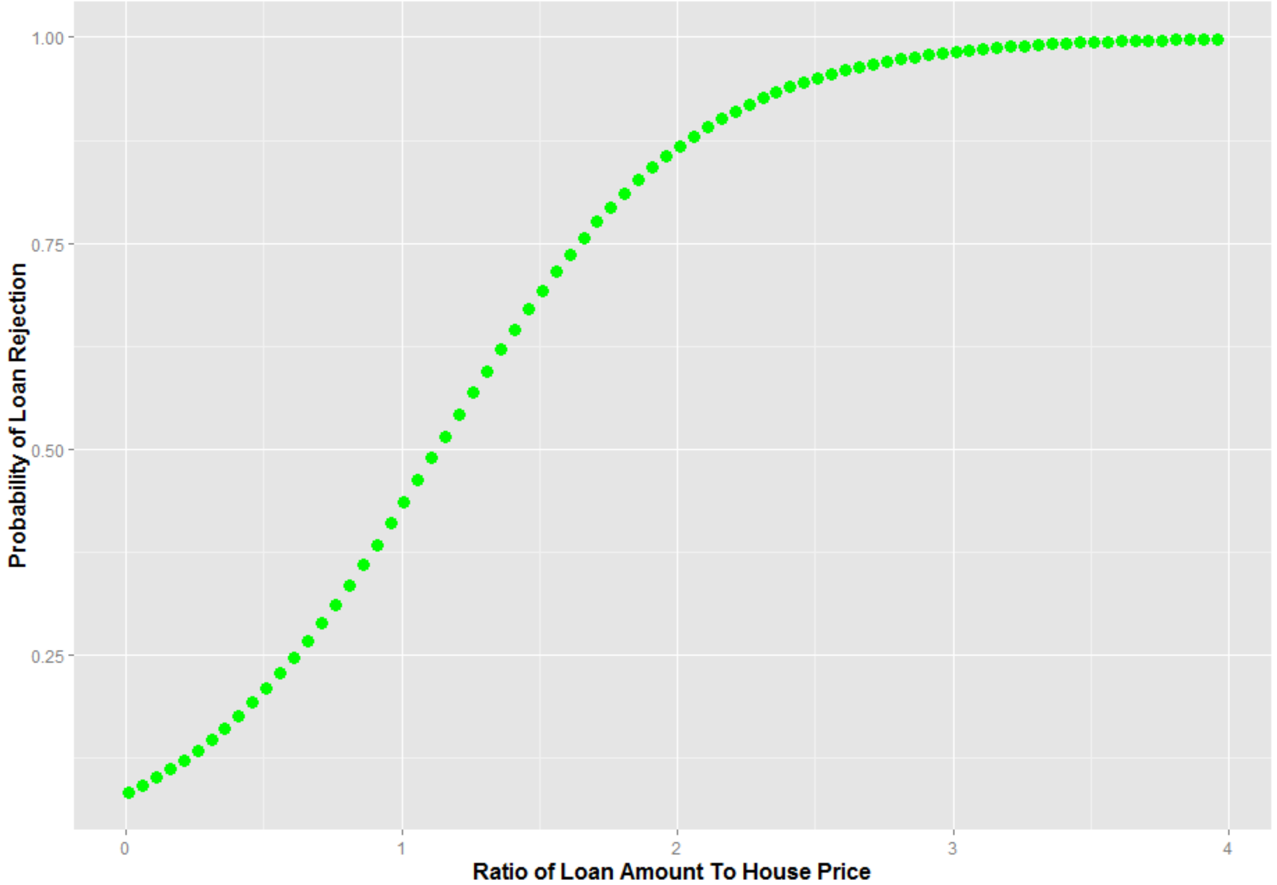
- Note that exp(2+3) = exp(2) * exp(3)
- So if the coefficient on x1 is 1.2. Then, we can say a 1 unit increase in x1 increases the logit by 1.2.
- We can also say the a 1 unit increases multiplies the odds by 3.3. As exp(1.2) = 3.3.
- So, if the odds of success were 10:1 before. The one unit increase results in 33:1 odds. Hence, much more likely to occur.

Probability Interpretations

- Logistic coefficients are most often interpreted in terms of odds (as we have been doing).
- However, it is possible to convert logits back to probabilities. We can calculate the predicted probability for any observation using the output.
- To do so, think back to the equations used to transform probabilities into logged odds. We now need to take the inverse to get probabilities again. [recall our discussion on GLMs]

- **Take a look at the to excel file on Blackboard!!**

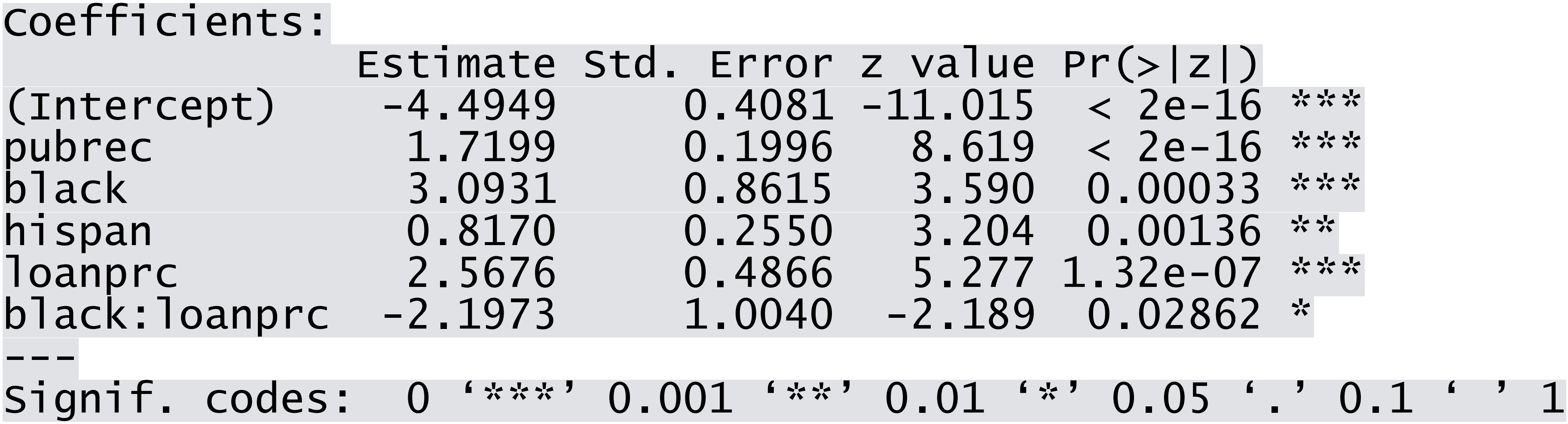
A helpful tool for interpretation/ presentation of results



```
sampdat2 = expand.grid(pubrec = 1,
                       black = 0,
                       hispan=0,
                       loanprc=seq(from=.01, to=4, by=.05))

predsamp2=(predict(loan2, new=sampdat2, type="response"))

#ggplot
ggplot(data=sampdat2, aes(x=loanprc, y=predsamp2)) +
geom_point(colour="green", size=4) +
  xlab("Ratio of Loan Amount To House Price") +
  ylab("Probability of Loan Rejection") +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=16,face="bold"))
```

- Create a new dataset and only vary one of the variables of interest, say loan amount. Use that dataset to produce new predicted values and then plot those predicted values against the predictor you varied.

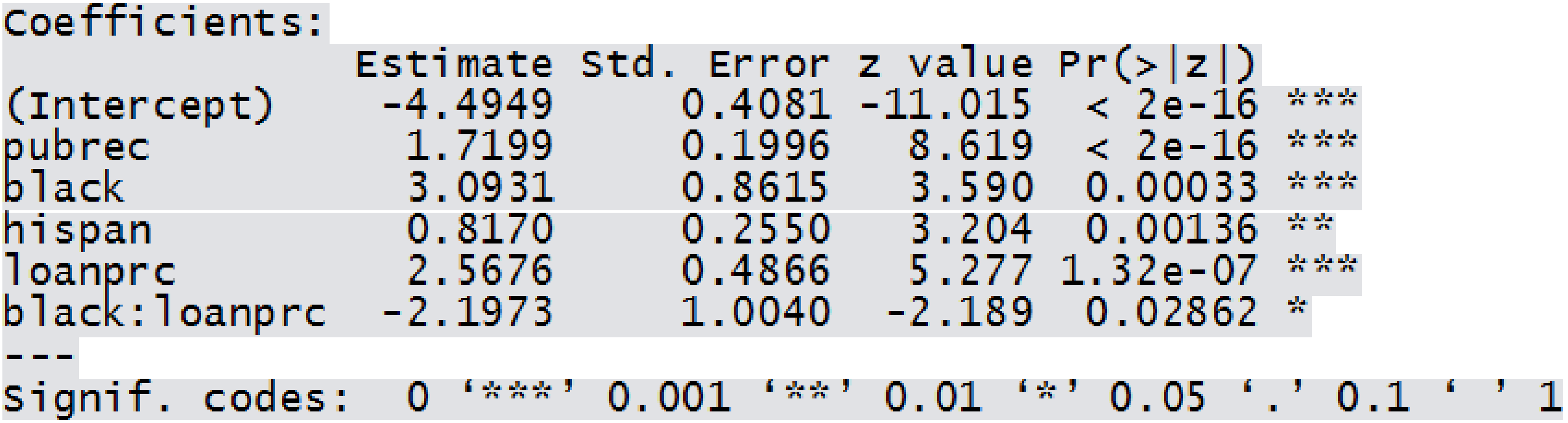
Starter Question: Interpreting Odds with Interaction Terms



```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept)    -4.4949     0.4081 -11.015  < 2e-16 ***
pubrec          1.7199     0.1996   8.619  < 2e-16 ***
black           3.0931     0.8615   3.590  0.00033 ***
hispan          0.8170     0.2550   3.204  0.00136 ** 
loanprc         2.5676     0.4866   5.277 1.32e-07 ***
black:loanprc  -2.1973     1.0040  -2.189  0.02862 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

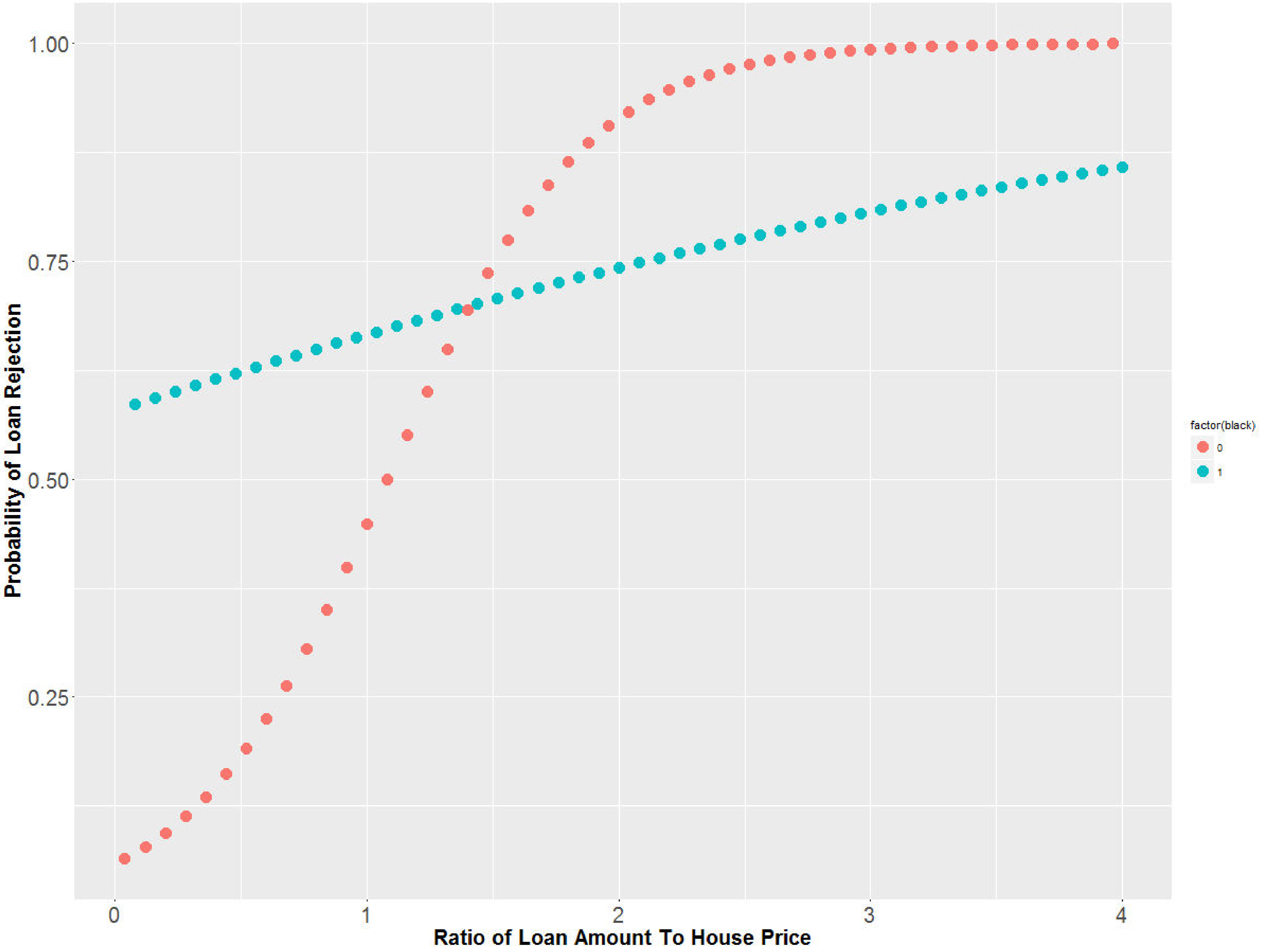
- Using the same loan dataset, I created an interaction between black and loanprc. How do we interpret this interaction term?

Interpreting Odds with Interaction Terms



```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept)    -4.4949     0.4081 -11.015  < 2e-16 ***
pubrec          1.7199     0.1996   8.619  < 2e-16 ***
black           3.0931     0.8615   3.590  0.00033 ***
hispan          0.8170     0.2550   3.204  0.00136 ** 
loanprc         2.5676     0.4866   5.277 1.32e-07 ***
black:loanprc  -2.1973     1.0040  -2.189  0.02862 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- How do we interpret this interaction term?
- When we include interactions, the Beta coefficient must be adjusted to include the interaction term. So for black individuals loanprc = 2.568 + -2.197 and for all others loanprc = 2.568. To calculate the odds ratio we need to exponentiate these values. Exp (2.568 – 2.197) = 1.449 and the Exp (2.568) = 13.040.
- What do these results tell us?


1.00
0.75
0.50
0.25
Probability of Loan Rejection
0
1
2
3
4
Ratio of Loan Amount To House Price
factor(black)
0
1

In class exercise

- Load the ski data from blackboard
- Run a logistic regression predicting falling based on difficulty and season. (Note, consider difficulty as a continuous variable).
- Answer the following:
  - Calculate the increase in **odds** for falling on a slope in winter of difficulty 1 versus difficulty 2.
  - Calculate the increase in **odds** for falling on a slope in winter of difficulty 1 versus difficulty 3. Calculate the increase in odds for falling in a season other than winter of difficulty 1 versus difficulty 3.
  - Calculate the **predicted probability** for falling in winter on a slope of difficulty 2 and difficulty 5.

A QUICK LOOK AT MAXIMUM LIKELIHOOD ESTIMATION

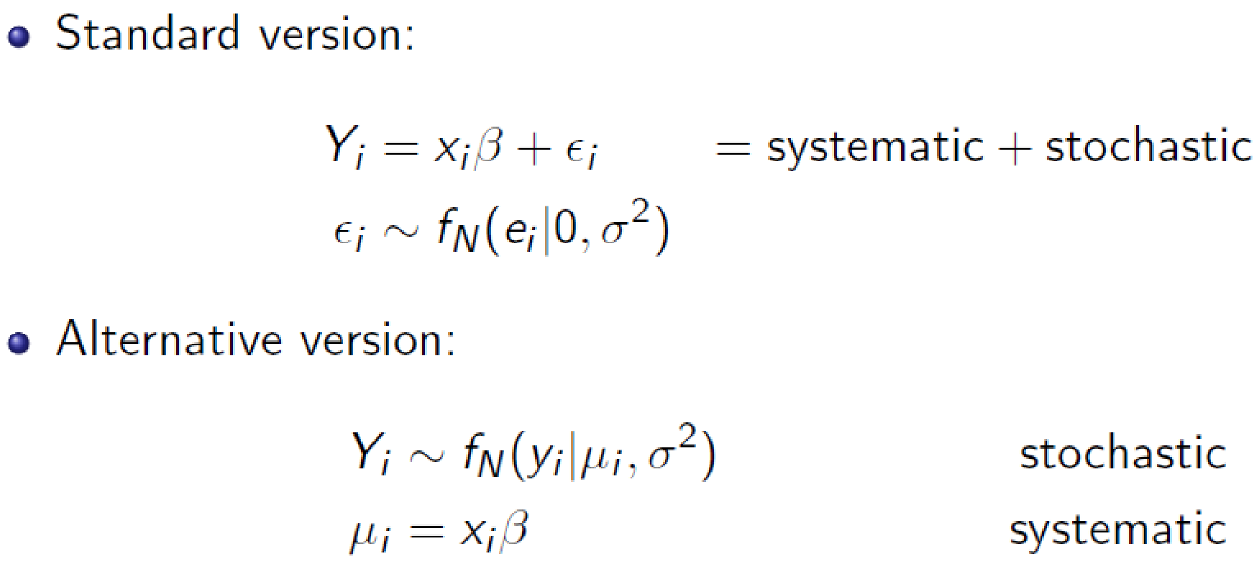
Probability and Statistics

- In probability the parameters are known and they control the behavior of a random variable via a model.
  - We use the known parameters to estimate the probability of certain future events occurring.

- In statistics (probability in reverse) the random variables (or the data) are known, and they are used to estimate the unknown parameters that gave rise to them via a model.

What are statistical models?

- A statistical model is a formal representation of the process by which a social system produces output.
- Equivalent notation (King 1998)



- Standard version:

$$\begin{aligned} Y_i &= x_i\beta + \epsilon_i && = \text{systematic} + \text{stochastic} \\ \epsilon_i &\sim f_N(e_i|0, \sigma^2) \end{aligned}$$

- Alternative version:

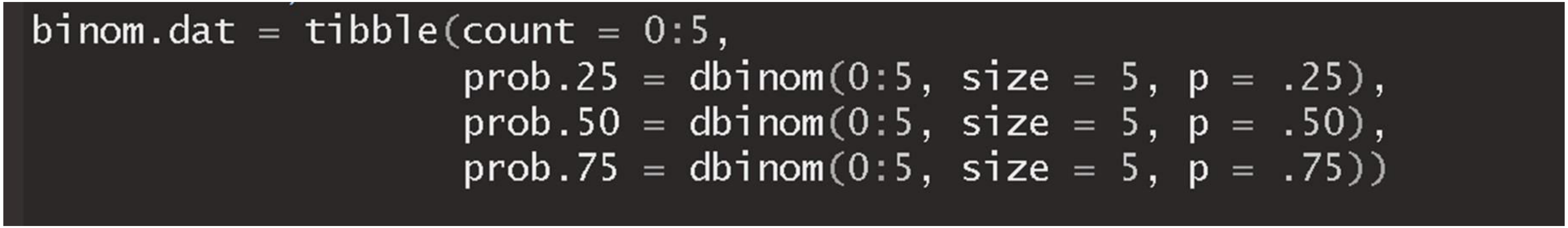
$$\begin{aligned} Y_i &\sim f_N(y_i|\mu_i, \sigma^2) && \text{stochastic} \\ \mu_i &= x_i\beta && \text{systematic} \end{aligned}$$

Maximum Likelihood Estimation

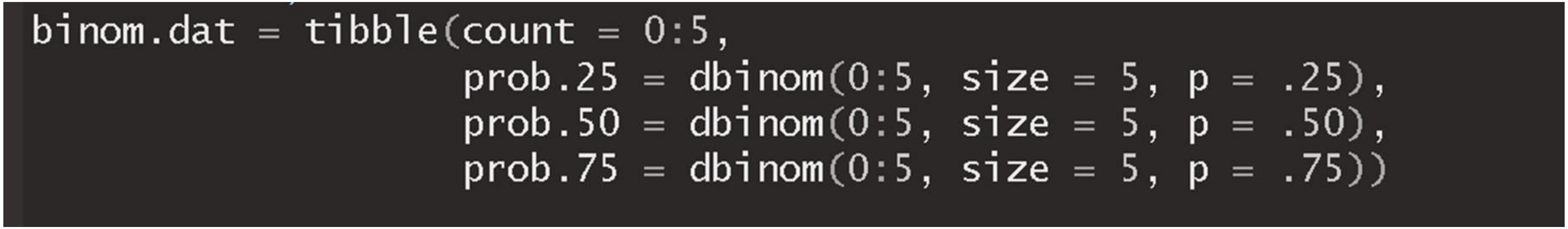
- Maximum likelihood estimation (MLE) finds estimates of model parameters that are most likely to give rise to the pattern of observations in the sample data.
- We will look at the mechanics of how this is done. This is simply to help you gain a more intuitive sense of what is happening when you run generalized linear models in R and where the MLE comes from.

Looking at possible distributions

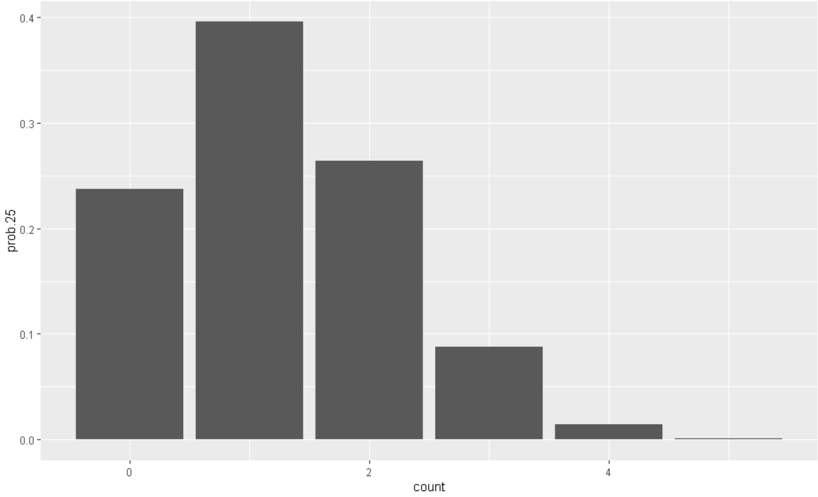
- Say we are interested in the prevalence of bullying across different sixth grade classrooms. Because any given classroom can have bullying or not, we can consider a binomial probability distribution.



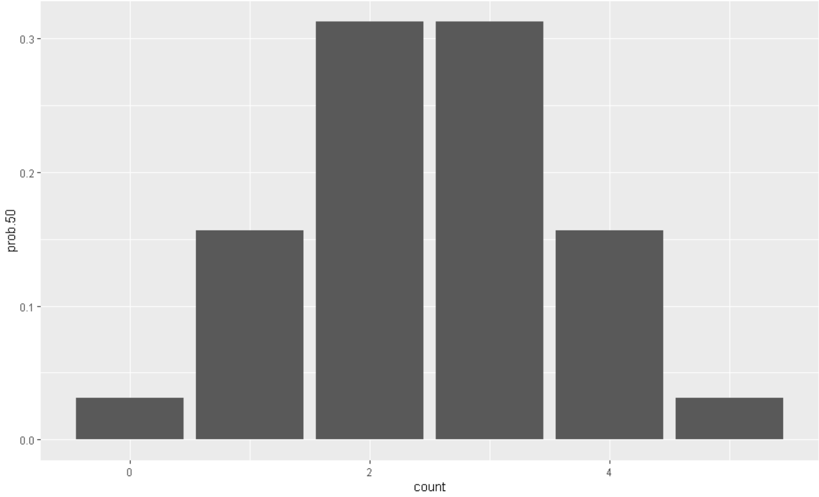
```
binom.dat = tibble(count = 0:5,
                   prob.25 = dbinom(0:5, size = 5, p = .25),
                   prob.50 = dbinom(0:5, size = 5, p = .50),
                   prob.75 = dbinom(0:5, size = 5, p = .75))
```



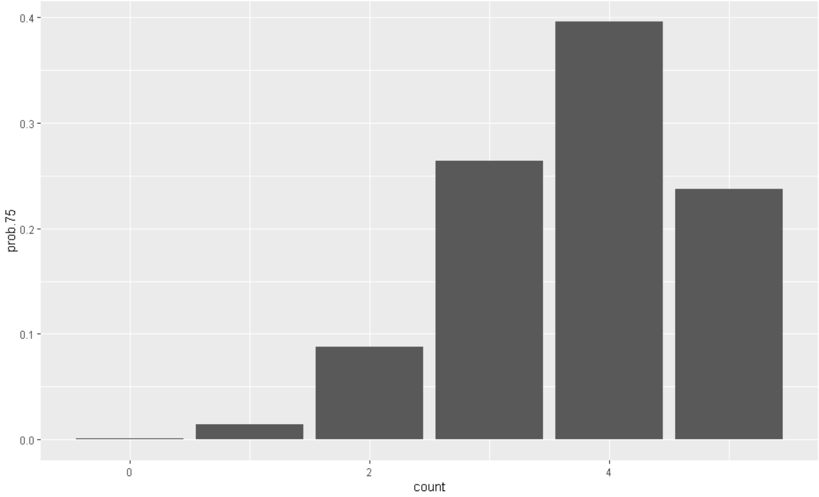
```
binom.dat = tibble(count = 0:5,
                   prob.25 = dbinom(0:5, size = 5, p = .25),
                   prob.50 = dbinom(0:5, size = 5, p = .50),
                   prob.75 = dbinom(0:5, size = 5, p = .75))
```



p = .25



p = .50

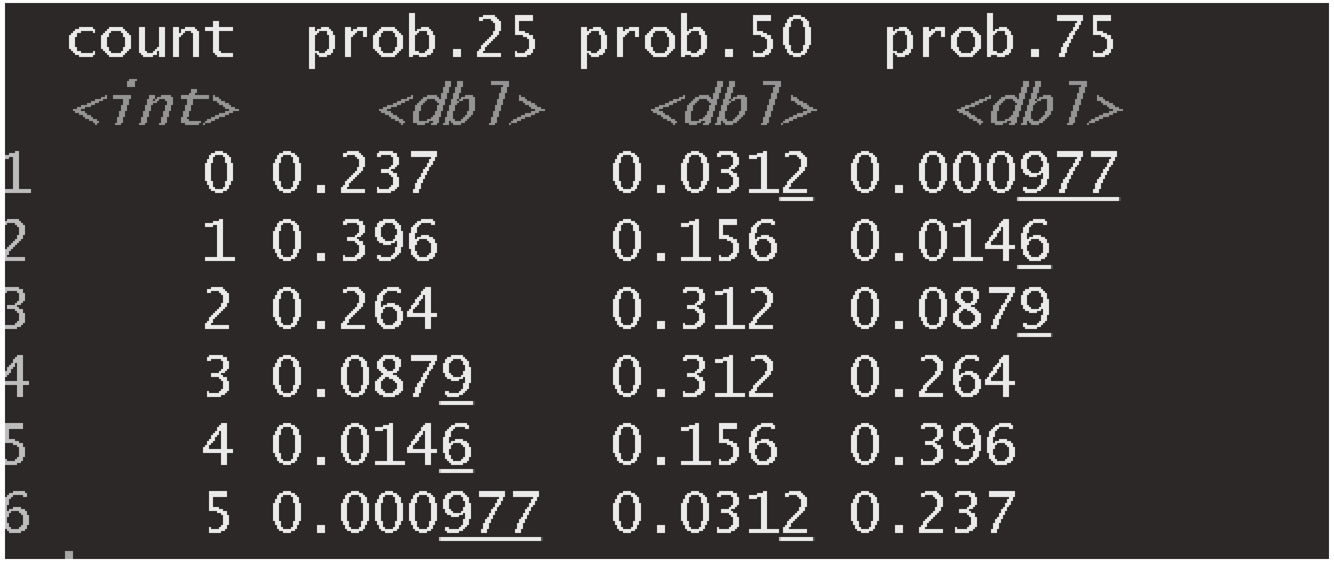


p = .75

The plots tell us the probability of observing any given count of bullying for each of three probability distributions.

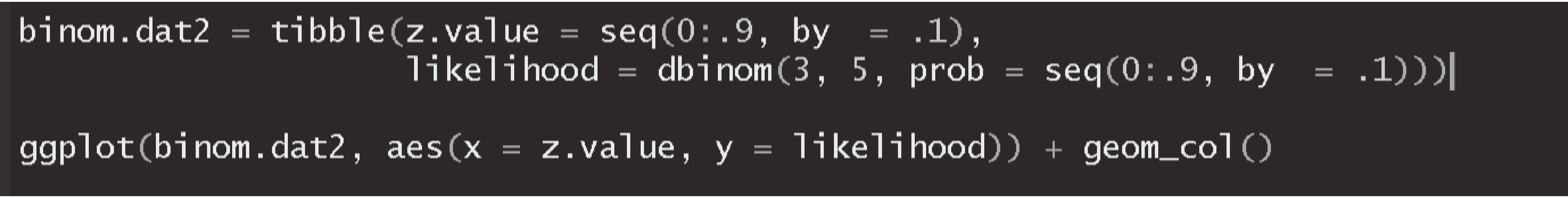
Go get some data

- Let’s go to five classrooms and observe if bullying is present or not.
- Say we observe the following - Y: {1,0,0,1,1}
- Which distribution (out of the three we looked at) gives us the highest probability of observing this data?



```
   count  prob.25 prob.50  prob.75
   <int>    <dbl>   <dbl>    <dbl>
1      0 0.237     0.0312 0.000977
2      1 0.396     0.156  0.0146
3      2 0.264     0.312  0.0879
4      3 0.0879    0.312  0.264
5      4 0.0146    0.156  0.396
6      5 0.000977  0.0312 0.237
```

- We know the binomial distribution with a parameter of .5 is better than the other two, but what about other possible values for the parameter?

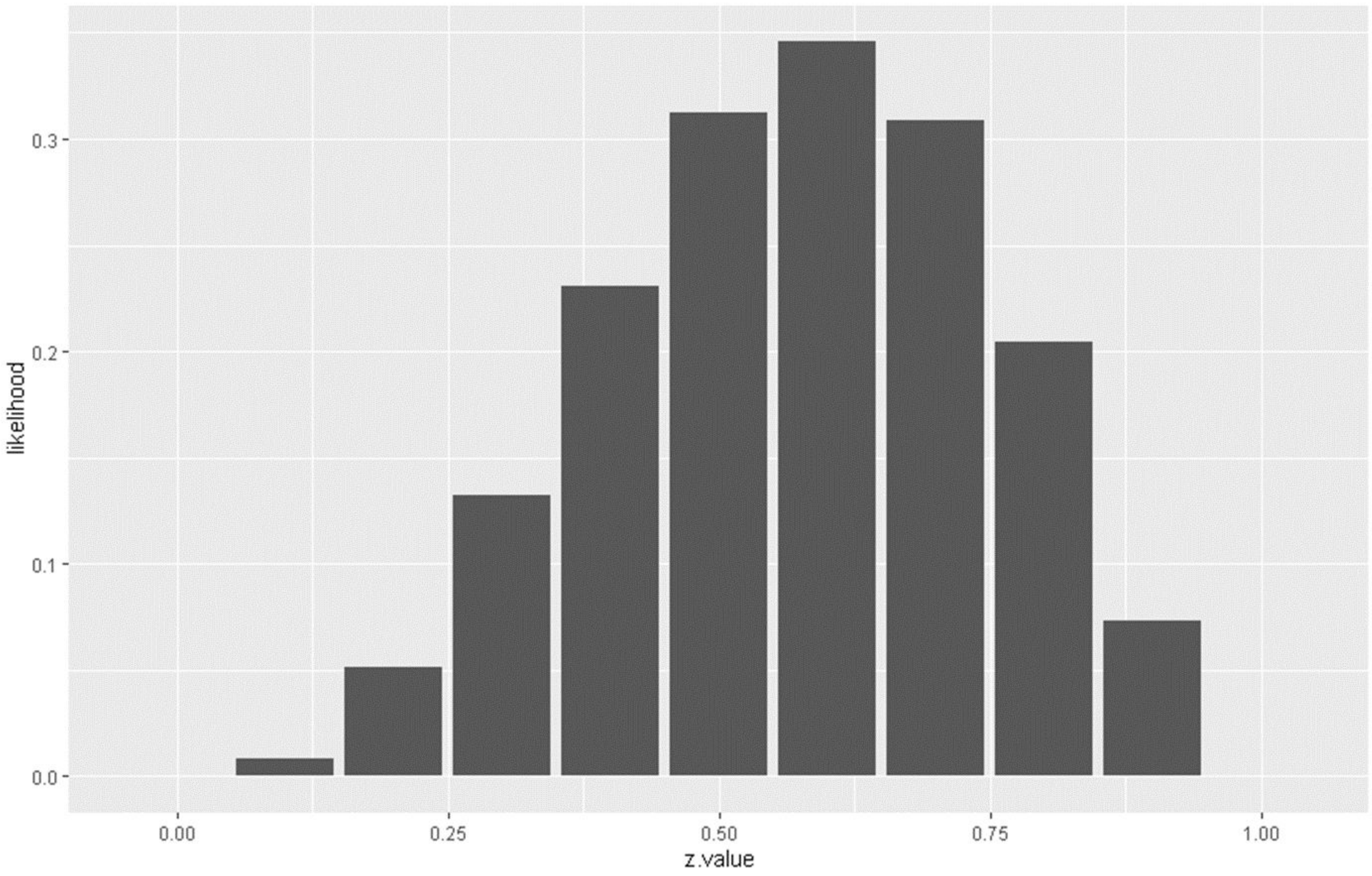


```r
binom.dat2 = tibble(z.value = seq(0:.9, by  = .1),
                    likelihood = dbinom(3, 5, prob = seq(0:.9, by  = .1)))

ggplot(binom.dat2, aes(x = z.value, y = likelihood)) + geom_col()
```

- This plot now shows us the probability of finding 3 classrooms with bullying out of 5 observations for 10 different values of the parameter.

- This is the essence of maximum likelihood.

- **Likelihood** is simply the extent to which a sample provides support for a given parameter value.



The Likelihood Function

- Let’s suppose again that we have the following data on a binary outcome Y: {1,0,0,1,1}
- We assume that Y is distributed Bernoulli with a constant probability across our observations.
- The model we propose is:

$$Y_i \sim f_{bern}\,(y_i|\pi_i)$$

- We know that:

$$Y_i = \begin{cases} 1\ with\ probability\ \pi \\ 0\ with\ probability\ 1-\ \pi \end{cases}$$

- So, the joint distribution of our data Y: {1,0,0,1,1} is:

$$\begin{aligned} Pr(\mathbf{y}|\pi) &= Pr(Y_1 = 1, Y_2 = 0, ..., Y_5 = 1|\pi) \\ &= Pr(Y_1 = 1|\pi)Pr(Y_2 = 0|\pi)...Pr(Y_5 = 1|\pi) \\ &= \pi \cdot (1-\pi) \cdot (1-\pi) \cdot \pi \cdot \pi \\ &= \pi^3(1-\pi)^2 \end{aligned}$$

- According to the theory of likelihood:

$$L(\pi|\boldsymbol{y})\ is\ proportional\ to\ p(\boldsymbol{y}|\pi)$$

$$L(\pi|\boldsymbol{y})\ \ is\ proportional\ to\ \ \pi^3(1-\ \pi)^2$$

- We take the log of the equation for computational purposes and get:

$$\ln[\pi^3(1-\ \pi)^2]\ =\ \ln(\pi^3)+\ln((1-\ \pi)^2))$$

$$\ln[\pi^3(1-\ \pi)^2]\ =3\ \ln\pi+2\ln(1-\ \pi)$$

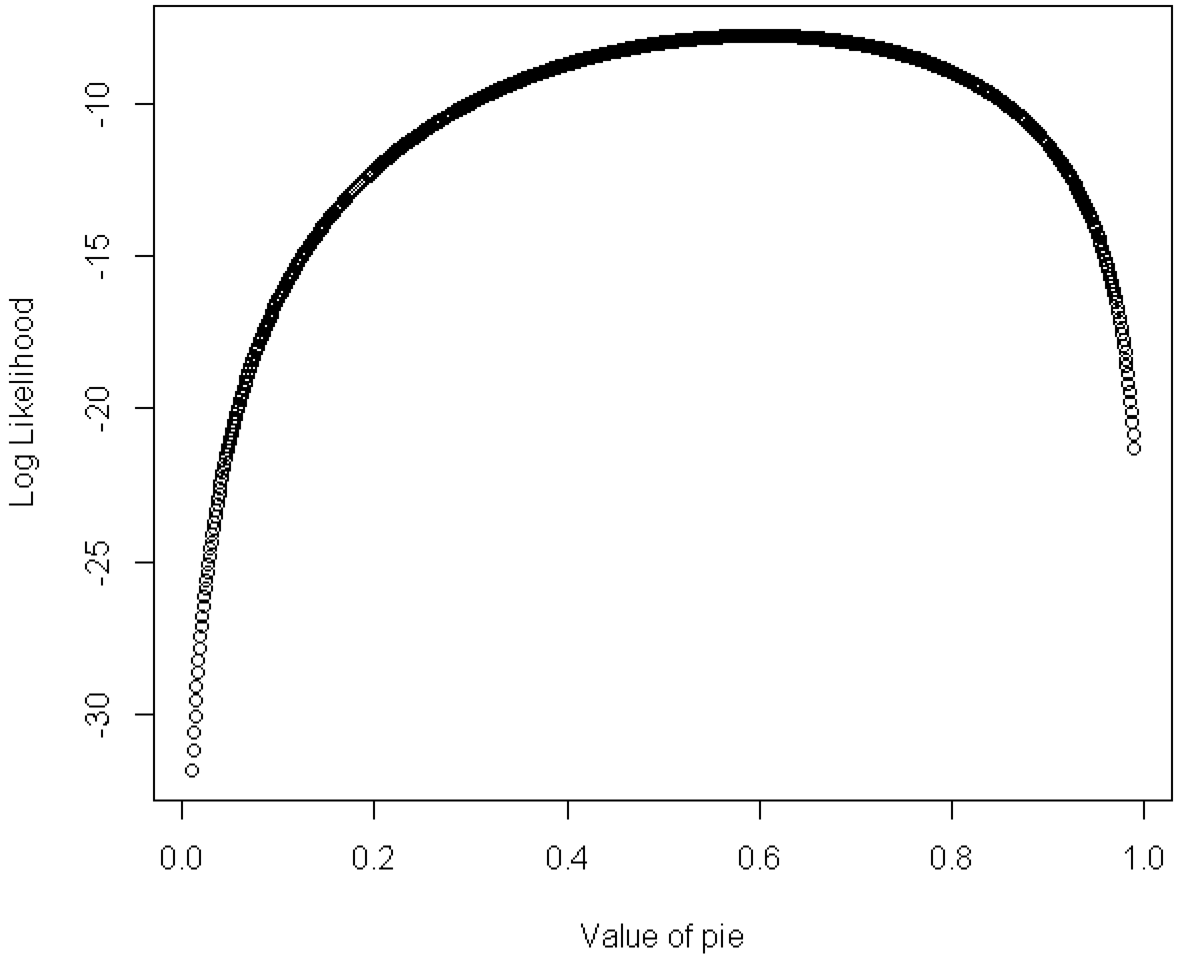
Maximizing the Function

$$\ln[\pi^3(1-\pi)^2] = 3\ln\pi + 2\ln(1-\pi)$$

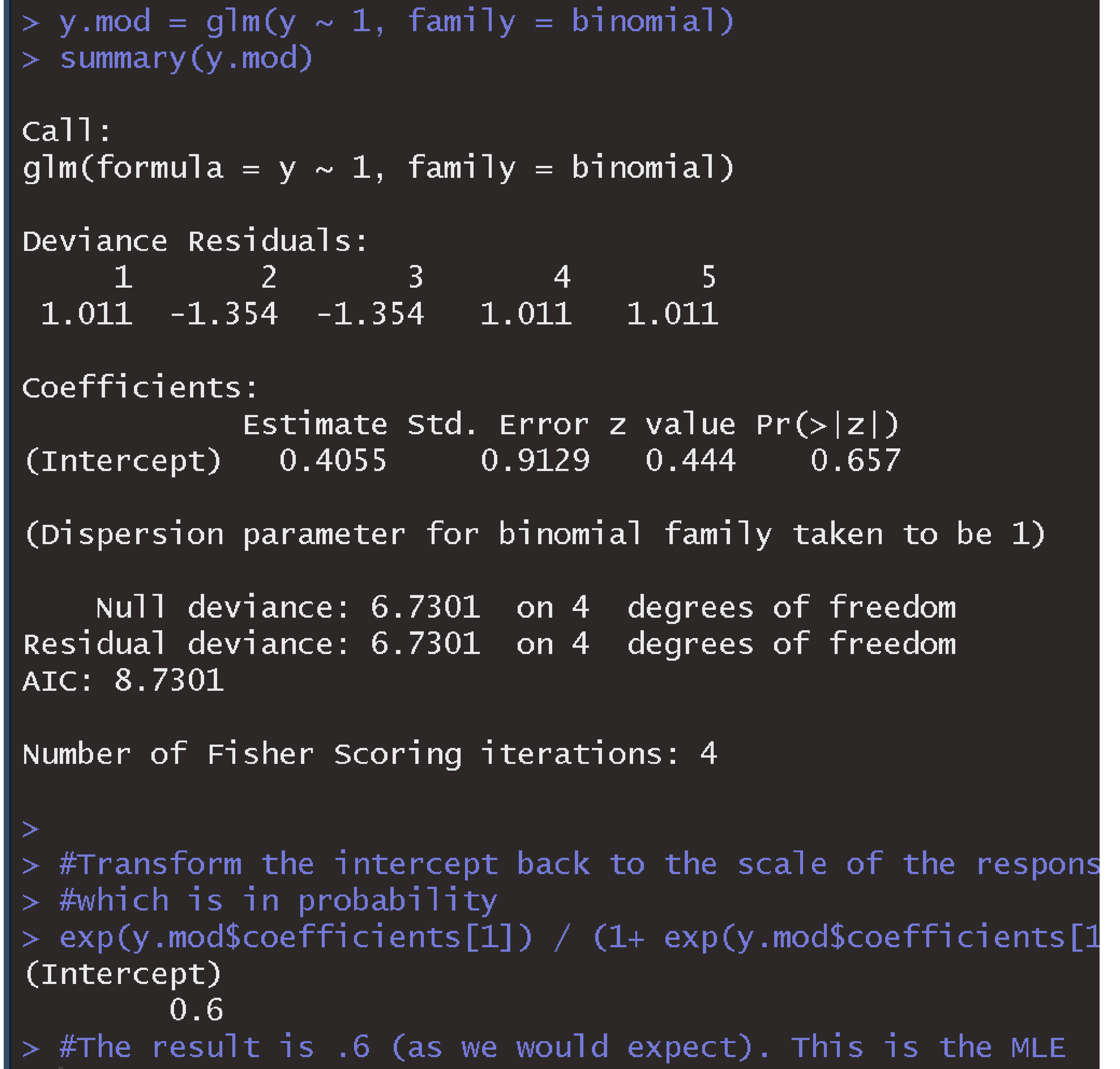
- We want to find the point when the above function is maximized.

$$Y_i \sim f_{bern}(y_i|\pi_i)$$

$$\pi_i = g(x_i B)$$



Run a logistic regression model



```
> y.mod = glm(y ~ 1, family = binomial)
> summary(y.mod)

Call:
glm(formula = y ~ 1, family = binomial)

Deviance Residuals: 
     1       2       3       4       5  
 1.011  -1.354  -1.354   1.011   1.011  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.4055     0.9129   0.444    0.657

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6.7301  on 4  degrees of freedom
Residual deviance: 6.7301  on 4  degrees of freedom
AIC: 8.7301

Number of Fisher Scoring iterations: 4

> 
> #Transform the intercept back to the scale of the respons
> #which is in probability
> exp(y.mod$coefficients[1]) / (1+ exp(y.mod$coefficients[1
(Intercept) 
        0.6 
> #The result is .6 (as we would expect). This is the MLE
```

We will examine this is below, but note the residual deviance is 6.73. This value is simply -2 times the loglikelihood.

-2 * (3*log(.6) + 2*log(1-.6)) = 6.37

- Let’s look at an example using our Ski data from last week.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -4.4990     1.5300  -2.941 0.003276 **
Difficulty     1.5688     0.4761   3.295 0.000984 ***
Seasonwinter   0.4773     1.0141   0.471 0.637861
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 48.263  on 34  degrees of freedom
Residual deviance: 26.470  on 32  degrees of freedom
AIC: 32.47

Number of Fisher Scoring iterations: 5
```

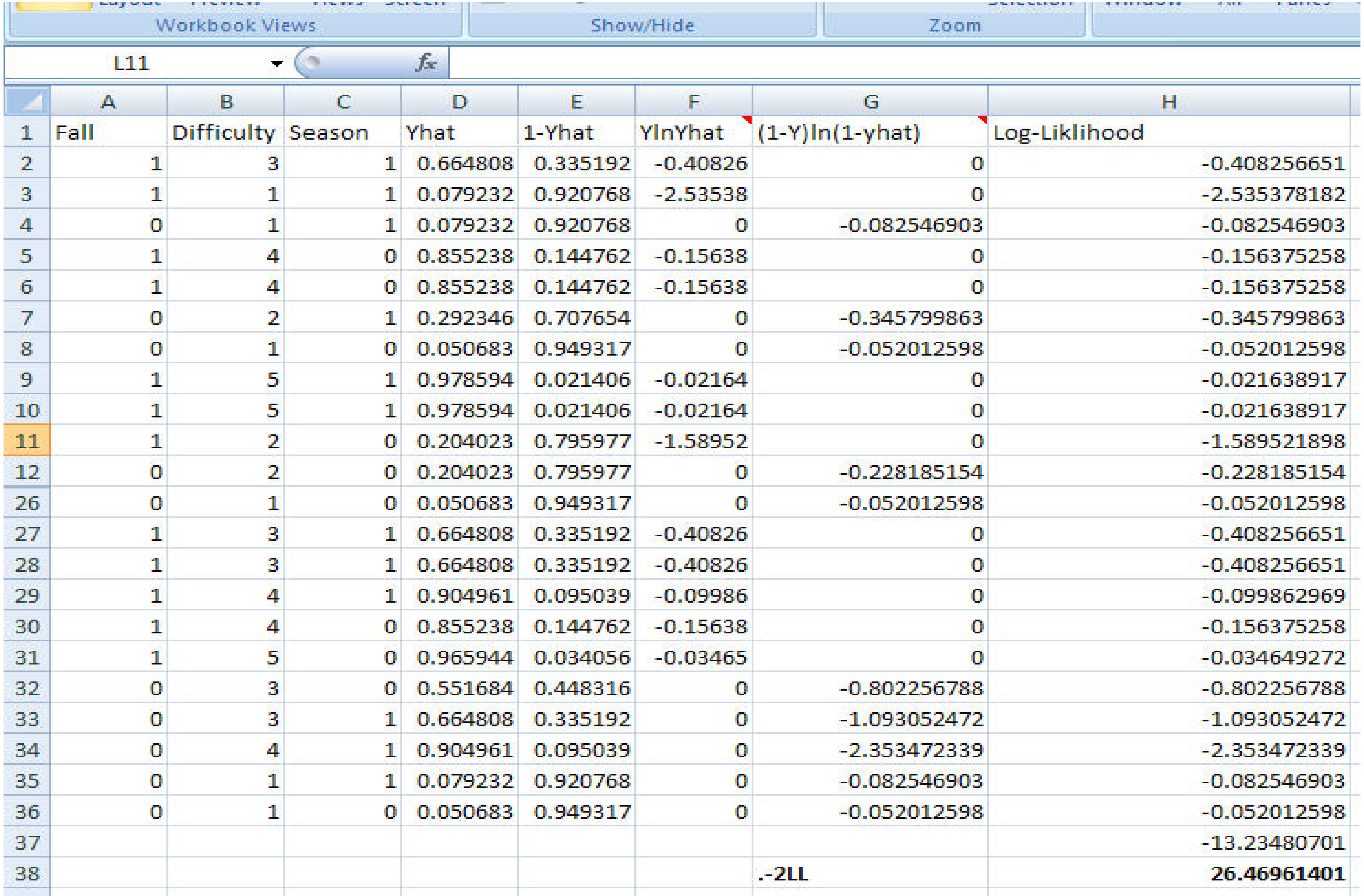
Log-Likelihood – Assessing the Model

$$log - likelihood = \sum_{i=1}^{N} \{Y_i \ln\left(P(Y_i)\right) + \ (1 - \ Y_i)\ln\left[1 - \mathrm{P}(\mathrm{Y_i})\right]\}$$

- The log-likelihood is therefore based on summing the likelihood associated with the observed outcomes. It is akin to the residual sum of squares in multiple regression as it is an indicator of how much unexplained information there is after the model has been fitted.
- Large values of the log-likelihood indicate poorly fitted models.
- Recall that the ln(1) = 0
- Log-likelihood allows us to compare two models by computing the difference in their log-likelihoods.
- In the model summary output for some software you will see -2 log likelihood or deviance, thus it is simply the formula above multiplied by -2.   This is done because it allows us to compare models based on the chi-square statistic. In R it is just called the deviance.

Calculating -2LL by Hand

$$log - likelihood = \sum_{i=1}^{N}\{Y_i \ln\left(P(Y_i)\right) + (1 - Y_i)\ln[1 - P(Y_i)]\}$$



	A	B	C	D	E	F	G	H
1	Fall	Difficulty	Season	Yhat	1-Yhat	YlnYhat	(1-Y)ln(1-yhat)	Log-Liklihood
2	1	3	1	0.664808	0.335192	-0.40826	0	-0.408256651
3	1	1	1	0.079232	0.920768	-2.53538	0	-2.535378182
4	0	1	1	0.079232	0.920768	0	-0.082546903	-0.082546903
5	1	4	0	0.855238	0.144762	-0.15638	0	-0.156375258
6	1	4	0	0.855238	0.144762	-0.15638	0	-0.156375258
7	0	2	1	0.292346	0.707654	0	-0.345799863	-0.345799863
8	0	1	0	0.050683	0.949317	0	-0.052012598	-0.052012598
9	1	5	1	0.978594	0.021406	-0.02164	0	-0.021638917
10	1	5	1	0.978594	0.021406	-0.02164	0	-0.021638917
11	1	2	0	0.204023	0.795977	-1.58952	0	-1.589521898
12	0	2	0	0.204023	0.795977	0	-0.228185154	-0.228185154
26	0	1	0	0.050683	0.949317	0	-0.052012598	-0.052012598
27	1	3	1	0.664808	0.335192	-0.40826	0	-0.408256651
28	1	3	1	0.664808	0.335192	-0.40826	0	-0.408256651
29	1	4	1	0.904961	0.095039	-0.09986	0	-0.099862969
30	1	4	0	0.855238	0.144762	-0.15638	0	-0.156375258
31	1	5	0	0.965944	0.034056	-0.03465	0	-0.034649272
32	0	3	0	0.551684	0.448316	0	-0.802256788	-0.802256788
33	0	3	1	0.664808	0.335192	0	-1.093052472	-1.093052472
34	0	4	1	0.904961	0.095039	0	-2.353472339	-2.353472339
35	0	1	1	0.079232	0.920768	0	-0.082546903	-0.082546903
36	0	1	0	0.050683	0.949317	0	-0.052012598	-0.052012598
37								-13.23480701
38							**.-2LL**	**26.46961401**

Comparing Models Using Log-Likelihood (likelihood ratio test)

- Models must be nested to be compared. This means that all of the predictors in the smaller (restricted) model must also be in the bigger (unrestricted) model.
- Taking the difference in the -2LL (or deviance) for the models (smaller model (restricted model) – bigger model(unrestricted model)) produces a test statistic that is distributed as a chi-square with df equal to the difference in the number of predictors.

$$\chi^2 = [-2LL(smaller\,model) \; - \; -2LL(biggermodel)]$$

- For example, in our dataset for falling when skiing if we ran a model with just the constant and difficulty as a predictor we would get a -2LL of 26.692 then:

$$\chi^2 = [(26.692) - (26.470)] = .222$$

Comparing Models Using Log-Likelihood

$$F = \frac{(SSR_r - SSR_{ur}) / q}{SSR_{ur}/(n - k - 1)}$$

$$\chi^2 = [(26.692) - (26.470)] = .222$$

- So basically, we are testing whether or not season added to our ability to predict falling. (While not the most interesting example, it can be easily extending to include multiple predictors as we did in the baseball example using the F-test and the residual sum of squares between the restricted and unrestricted models. Thus, you can think of this chi-square test as comparing the reduction in the sum of the residuals when moving from a restricted model (smaller model) to an unrestricted model (larger model).
- **Degrees of freedom for the χ^2 critical value is determined by the difference in in degrees of freedom between the big and small model.** Thus, in our case the degrees of freedom for the big model is 3 (1 for each predictor and the constant) and the small model has 2 df (the constant and only 1 predictor).
- The critical value at α = .05, with 1 df is 3.84 and so we fail to reject the null and conclude that the model will all of the independent variables does not predict better than the one with only difficulty as the independent variable. An expected result given the lack of statistical significance for season.

Let’s look at another example of the likelihood ratio test

- We can use the likelihood ratio test to compare nested models and also single predictors. When sample sizes are small, the likelihood ratio test is often preferred to the Wald statistic or the z-test. Let’s take a look at how the likelihood ratio test works.
- We will examine the significance of the interaction term from our model predicting loan rejections.

- Thus we need to run two models:
  - One model with the interaction (larger, or unrestricted model)
  - Another model without the interaction (smaller, or restricted model)

```
> loan3=glm(reject ~ pubrec + black + hispan + loanprc + loanprc*black,
+           family=binomial, data=loan)
> summary(loan3)

Call:
glm(formula = reject ~ pubrec + black + hispan + loanprc + loanprc * 
    black, family = binomial, data = loan)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept)    -4.4949     0.4081 -11.015  < 2e-16 ***
pubrec          1.7199     0.1996   8.619  < 2e-16 ***
black           3.0931     0.8615   3.590  0.00033 ***
hispan          0.8170     0.2550   3.204  0.00136 ** 
loanprc         2.5676     0.4866   5.277 1.32e-07 ***
black:loanprc  -2.1973     1.0040  -2.189  0.02862 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1480.7  on 1988  degrees of freedom
Residual deviance: 1293.0  on 1983  degrees of freedom
AIC: 1305

Number of Fisher Scoring iterations: 5
```

Unrestricted model with the interaction

```
> loan3r = glm(reject ~ pubrec + black + hispan + loanprc,
+              family=binomial, data=loan)
> summary(loan3r)

Call:
glm(formula = reject ~ pubrec + black + hispan + loanprc, family = binomial, 
    data = loan)


Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  -4.1480     0.3661 -11.332  < 2e-16 ***
pubrec        1.7297     0.1991   8.687  < 2e-16 ***
black         1.2444     0.1860   6.691 2.21e-11 ***
hispan        0.8436     0.2540   3.321 0.000895 ***
loanprc       2.1399     0.4375   4.892 9.99e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1480.7  on 1988  degrees of freedom
Residual deviance: 1297.5  on 1984  degrees of freedom
AIC: 1307.5

Number of Fisher Scoring iterations: 5
```

Restricted model with no interaction

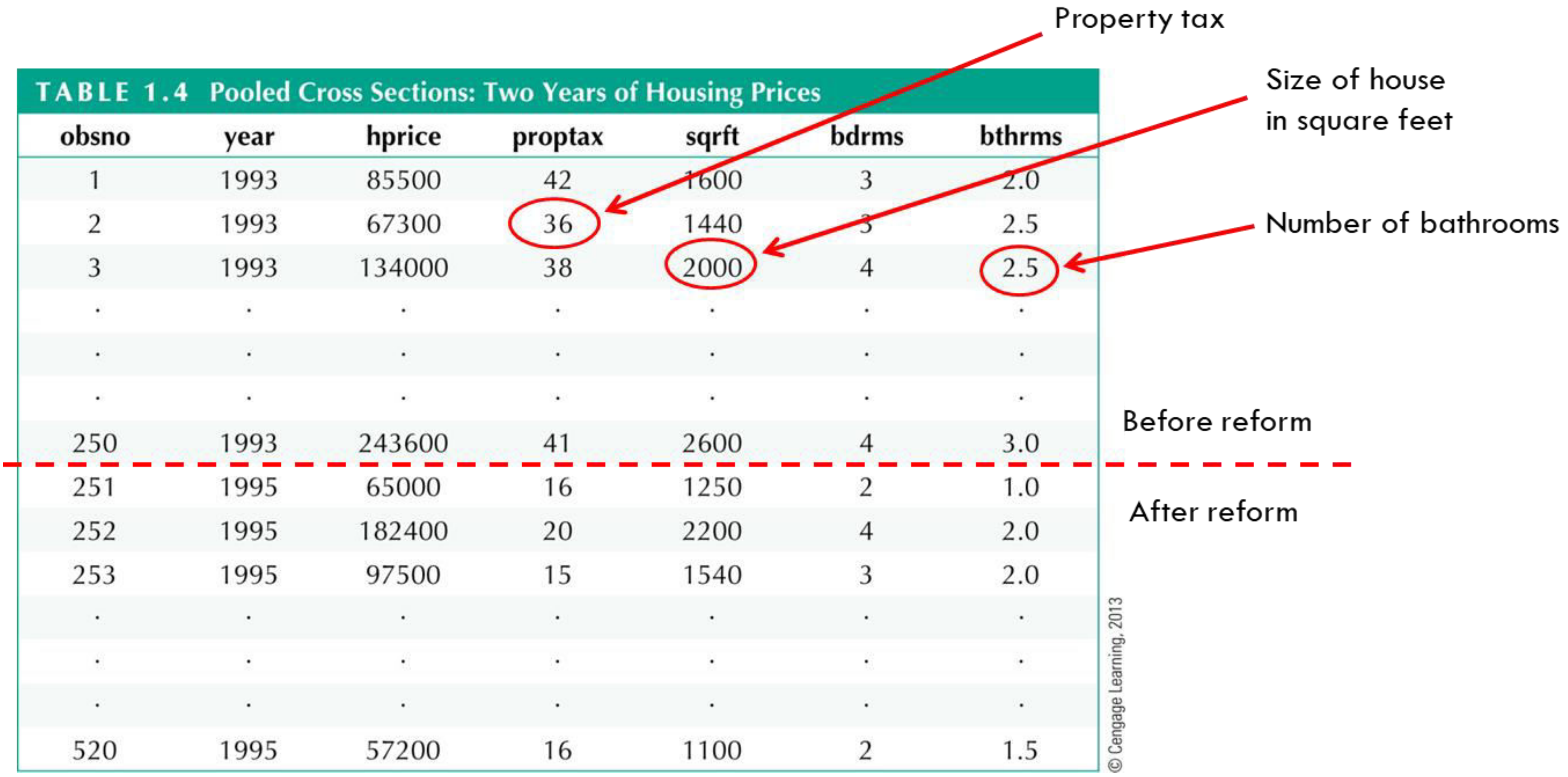
```
> loan3=glm(reject ~ pubrec + black + hispan + loanprc + loanprc*black,
+           family=binomial, data=loan)
> loan3r = glm(reject ~ pubrec + black + hispan + loanprc,
+              family=binomial, data=loan)
> anova(loan3r, loan3, test="Chisq")

Analysis of Deviance Table

Model 1: reject ~ pubrec + black + hispan + loanprc
Model 2: reject ~ pubrec + black + hispan + loanprc + loanprc * black
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      1984     1297.5
2      1983     1293.0  1   4.5001  0.03389 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


> #another way
> library(lmtest)
> lrtest(loan3r, loan3)
Likelihood ratio test

Model 1: reject ~ pubrec + black + hispan + loanprc
Model 2: reject ~ pubrec + black + hispan + loanprc + loanprc * black
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   5 -648.74
2   6 -646.49  1 4.5001    0.03389 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

POOLED CROSS SECTIONAL DATA

Independently pooled cross section

- Surveys like the General Social Survey (GSS) or the Current Population Survey (CPS) are repeated at regular intervals with a random sample of individuals each year.
- Pooling multiple years together gives us an independently pooled cross section.

Pooled cross sections on housing prices



Property tax

Size of house in square feet

Number of bathrooms

Before reform

After reform

TABLE 1.4 Pooled Cross Sections: Two Years of Housing Prices

obsno	year	hprice	proptax	sqrft	bdrms	bthrms
1	1993	85500	42	1600	3	2.0
2	1993	67300	36	1440	3	2.5
3	1993	134000	38	2000	4	2.5
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
250	1993	243600	41	2600	4	3.0
251	1995	65000	16	1250	2	1.0
252	1995	182400	20	2200	4	2.0
253	1995	97500	15	1540	3	2.0
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
520	1995	57200	16	1100	2	1.5

Advantages of pooling cross-sectional data

- Researchers can generate more precise parameter estimates by including more observations.
- Pooling is beneficial:
  - When the sample size in any given year is small.
    - The effect being examined is too small to discern using a sample from a single year, but additional data will produce a statistically significant result.
  - When researchers want to assess if outcomes change over time or if the coefficients in the model change over time.

Statistical issues with pooling

- Only minor statistical issues are raised:
  - Populations may have different distributions in different years.
    - We can allow the intercept to vary across different time periods by including dummy variables for years [just as we have done for gender or race].
    - These year dummy variables may be of substantive interest to the researcher.
  - Error variance may change over time
    - We already know how to deal with this - use robust standard errors.

An example from Wooldridge

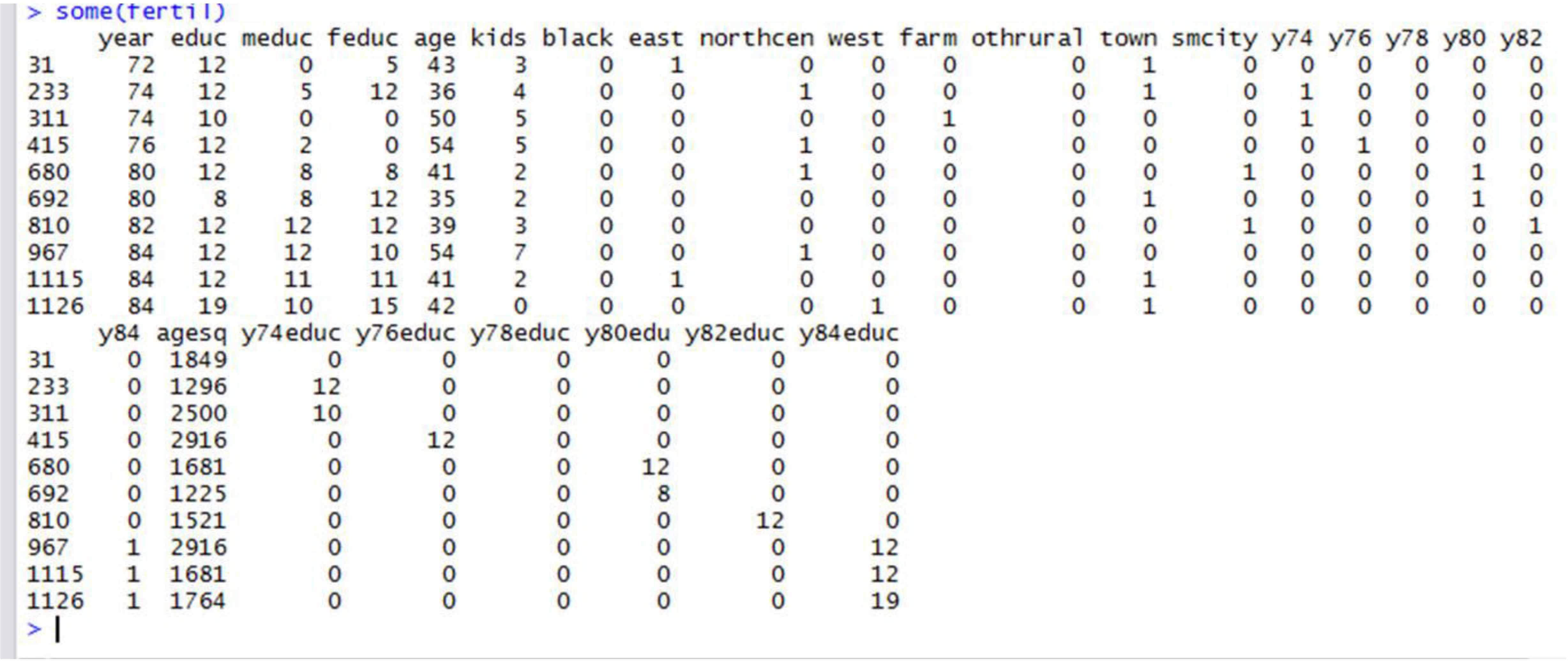
- Using GSS data from 1972-1984 we will estimate the number of kids being born.
- Question of interest: after controlling for other observable factors, what has happened to fertility rates overtime?

The data on Number of Kids Born

Obs: 1129

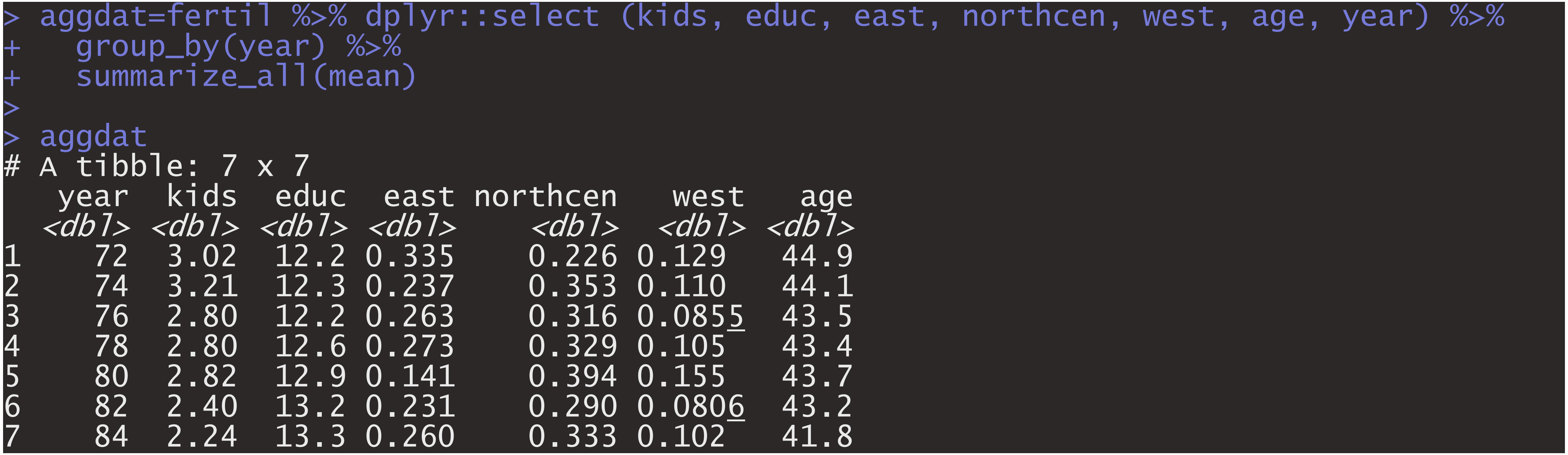
1. year	72 to 84, even
2. educ	years of schooling
3. meduc	mother's education
4. feduc	father's education
5. age	in years
6. kids	# children ever born
7. black	= 1 if black
8. east	= 1 if lived in east at 16
9. northcen	= 1 if lived in nc at 16
10. west	= 1 if lived in west at 16
11. farm	= 1 if on farm at 16
12. othrural	= 1 if other rural at 16
13. town	= 1 if lived in town at 16
14. smcity	= 1 if in small city at 16
15. y74	= 1 if year = 74
16. y76	
17. y78	
18. y80	
19. y82	
20. y84	
21. agesq	age^2

Data structure



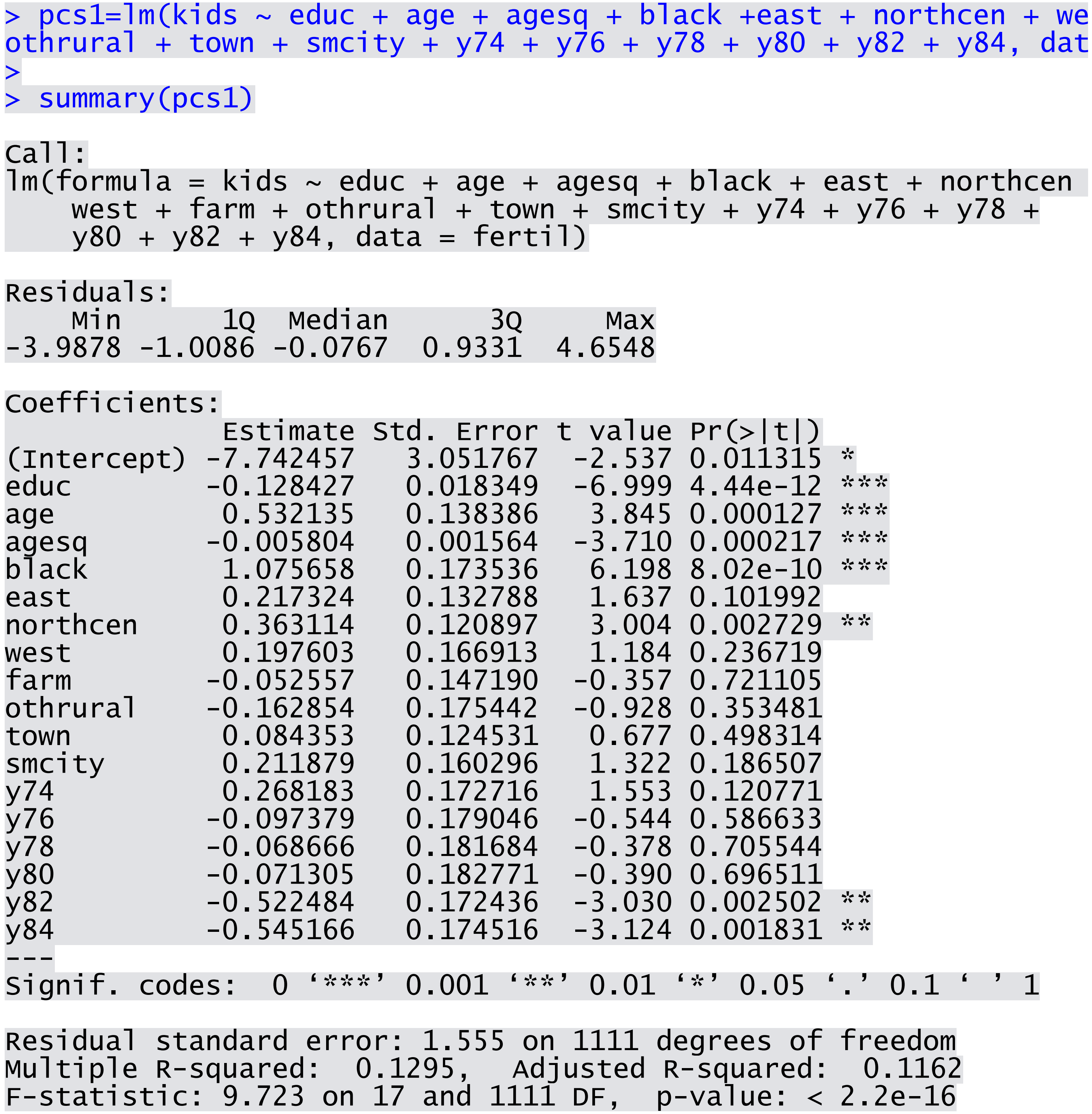
```
> some(fertil)
     year educ meduc feduc age kids black east northcen west farm othrural town smcity y74 y76 y78 y80 y82
31     72   12     0     5  43    3     0    1        0    0    0        0    1      0   0   0   0   0   0
233    74   12     5    12  36    4     0    0        1    0    0        0    1      0   1   0   0   0   0
311    74   10     0     0  50    5     0    0        0    0    1        0    0      0   1   0   0   0   0
415    76   12     2     0  54    5     0    0        1    0    0        0    0      0   0   1   0   0   0
680    80   12     8     8  41    2     0    0        1    0    0        0    0      1   0   0   0   1   0
692    80    8     8    12  35    2     0    0        0    0    0        0    1      0   0   0   0   1   0
810    82   12    12    12  39    3     0    0        0    0    0        0    0      1   0   0   0   0   1
967    84   12    12    10  54    7     0    0        1    0    0        0    0      0   0   0   0   0   0
1115   84   12    11    11  41    2     0    1        0    0    0        0    1      0   0   0   0   0   0
1126   84   19    10    15  42    0     0    0        0    1    0        0    1      0   0   0   0   0   0
     y84 agesq y74educ y76educ y78educ y80edu y82educ y84educ
31     0  1849       0       0       0      0       0       0
233    0  1296      12       0       0      0       0       0
311    0  2500      10       0       0      0       0       0
415    0  2916       0      12       0      0       0       0
680    0  1681       0       0       0     12       0       0
692    0  1225       0       0       0      8       0       0
810    0  1521       0       0       0      0      12       0
967    1  2916       0       0       0      0       0      12
1115   1  1681       0       0       0      0       0      12
1126   1  1764       0       0       0      0       0      19
> |
```

Key variables by year



```
> aggdat=fertil %>% dplyr::select (kids, educ, east, northcen, west, age, year) %>%
+   group_by(year) %>%
+   summarize_all(mean)
>
> aggdat
# A tibble: 7 x 7
   year  kids  educ  east northcen   west   age
  <dbl> <dbl> <dbl> <dbl>    <dbl>  <dbl> <dbl>
1    72  3.02  12.2 0.335    0.226 0.129   44.9
2    74  3.21  12.3 0.237    0.353 0.110   44.1
3    76  2.80  12.2 0.263    0.316 0.0855  43.5
4    78  2.80  12.6 0.273    0.329 0.105   43.4
5    80  2.82  12.9 0.141    0.394 0.155   43.7
6    82  2.40  13.2 0.231    0.290 0.0806  43.2
7    84  2.24  13.3 0.260    0.333 0.102   41.8
```

- Therefore, one reason fertility may have declined may not simply be due to behavioral changes but rather changes in population characteristics that are associated with fertility.



```
> pcs1=lm(kids ~ educ + age + agesq + black +east + northcen + we
othrural + town + smcity + y74 + y76 + y78 + y80 + y82 + y84, dat
>
> summary(pcs1)

Call:
lm(formula = kids ~ educ + age + agesq + black + east + northcen
    west + farm + othrural + town + smcity + y74 + y76 + y78 +
    y80 + y82 + y84, data = fertil)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9878 -1.0086 -0.0767  0.9331  4.6548

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.742457   3.051767  -2.537 0.011315 *
educ        -0.128427   0.018349  -6.999 4.44e-12 ***
age          0.532135   0.138386   3.845 0.000127 ***
agesq       -0.005804   0.001564  -3.710 0.000217 ***
black        1.075658   0.173536   6.198 8.02e-10 ***
east         0.217324   0.132788   1.637 0.101992
northcen     0.363114   0.120897   3.004 0.002729 **
west         0.197603   0.166913   1.184 0.236719
farm        -0.052557   0.147190  -0.357 0.721105
othrural    -0.162854   0.175442  -0.928 0.353481
town         0.084353   0.124531   0.677 0.498314
smcity       0.211879   0.160296   1.322 0.186507
y74          0.268183   0.172716   1.553 0.120771
y76         -0.097379   0.179046  -0.544 0.586633
y78         -0.068666   0.181684  -0.378 0.705544
y80         -0.071305   0.182771  -0.390 0.696511
y82         -0.522484   0.172436  -3.030 0.002502 **
y84         -0.545166   0.174516  -3.124 0.001831 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.555 on 1111 degrees of freedom
Multiple R-squared:  0.1295,	Adjusted R-squared:  0.1162
F-statistic: 9.723 on 17 and 1111 DF,  p-value: < 2.2e-16
```

Model Output

- Here we examine whether fertility decisions have changed over time after controlling for other factors that affect fertility.
- Year dummy variables are included to capture time effects of fertility.

- What can we say about fertility rates overtime?
- What about education – what is the difference in average number of children between a high school educated and a college educated mother?

Some things to consider

- **ONE**: There may be heteroskedasticity in the error term underlying the estimated equation.
  - We can test for this using the Breush-Pagan test.
  - Using robust standard errors to correct for heteroskedasticity is generally a good approach when pooling cross sectional data.
- **TWO**: The model assumes that the effect of the explanatory variable has remained constant. This may not be true and may be of substantive interest to the researcher.
  - This is another reason to pool cross-sectional data. It allows us to determine whether or not the coefficients in the model have changed over time.

ONE: Let’s test and deal with heteroskedasticity

```
> bptest(pcs1) #based on the test it appears that heteroskedasticity is a
problem and we should probably use a robust standard error

        studentized Breusch-Pagan test

data:  pcs1
BP = 55.3154, df = 17, p-value = 6.098e-06

> #Robust Standard Errors in R
> pcs1$newse = vcovHC(pcs1) #create a new var-cov matrix which allows
> #us to produce robust standard errors
> coeftest(pcs1,pcs1$newse) #update the table with the robust SEs

t test of coefficients:

              Estimate Std. Error t value  Pr(>|t|)    
(Intercept) -7.7424566  3.1043103 -2.4941 0.0127722 *  
educ        -0.1284268  0.0214273 -5.9936  2.77e-09 ***
age          0.5321346  0.1404363  3.7892 0.0001593 ***
agesq       -0.0058040  0.0015963 -3.6358 0.0002898 ***
black        1.0756575  0.2042940  5.2652  1.68e-07 ***
east         0.2173240  0.1285529  1.6905 0.0912050 .  
northcen     0.3631140  0.1176391  3.0867 0.0020742 ** 
west         0.1976032  0.1646558  1.2001 0.2303570    
farm        -0.0525575  0.1473475 -0.3567 0.7213910    
othrural    -0.1628537  0.1829857 -0.8900 0.3736692    
town         0.0843532  0.1294116  0.6518 0.5146512    
smcity       0.2118791  0.1555361  1.3623 0.1733949    
y74          0.2681825  0.1890246  1.4188 0.1562465    
y76         -0.0973795  0.2016918 -0.4828 0.6293236    
y78         -0.0686665  0.1994630 -0.3443 0.7307184    
y80         -0.0713053  0.1954117 -0.3649 0.7152570    
y82         -0.5224842  0.1895297 -2.7567 0.0059337 ** 
y84         -0.5451661  0.1875224 -2.9072 0.0037192 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

TWO: Let’s deal with non-constant effects

- Education’s impact on the number of children
  - Does the impact of education change overtime?
    - In our previous model, the effect of education was constant across time. Why? – Because we modeled it that way.
  - If we are interested in the changes in education’s impact:
    - What are we asking specifically?
    - How can we test for this?

```
> pcs2=lm(kids ~ educ + age + agesq + black +east + northcen + west + farm +
othrural + town + smcity + y74 + y76 + y78 + y80 + y82 + y84 + educ:y74 + edu
c:y76 + educ:y78 + educ:y80 + educ:y82 + educ:y84, data=fertil)
>
> summary(pcs2)

Call:
lm(formula = kids ~ educ + age + agesq + black + east + northcen +
    west + farm + othrural + town + smcity + y74 + y76 + y78 +
    y80 + y82 + y84 + educ:y74 + educ:y76 + educ:y78 + educ:y80 +
    educ:y82 + educ:y84, data = fertil)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5343 -1.0340 -0.0823  0.9550  4.6006

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.477302   3.126360  -2.712 0.006801 **
educ        -0.022515   0.053618  -0.420 0.674628
age          0.507466   0.138922   3.653 0.000271 ***
agesq       -0.005525   0.001570  -3.519 0.000451 ***
black        1.074055   0.173701   6.183 8.82e-10 ***
east         0.206056   0.133143   1.548 0.121998
northcen     0.348287   0.121099   2.876 0.004104 **
west         0.177122   0.167452   1.058 0.290402
farm        -0.072162   0.147508  -0.489 0.624791
othrural    -0.191154   0.175934  -1.087 0.277491
town         0.088229   0.124536   0.708 0.478804
smcity       0.205358   0.160210   1.282 0.200182
y74          0.946915   0.904159   1.047 0.295196
y76          1.019963   0.882034   1.156 0.247777
y78          1.805985   0.951866   1.897 0.058047 .
y80          1.114183   0.897601   1.241 0.214762
y82          1.199807   0.876289   1.369 0.171218
y84          1.671261   0.899050   1.859 0.063304 .
educ:y74    -0.056425   0.072561  -0.778 0.436958
educ:y76    -0.092100   0.070875  -1.299 0.194053
educ:y78    -0.152387   0.075282  -2.024 0.043187 *
educ:y80    -0.097905   0.070452  -1.390 0.164912
educ:y82    -0.138945   0.068371  -2.032 0.042371 *
educ:y84    -0.176097   0.069915  -2.519 0.011918 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.553 on 1105 degrees of freedom
Multiple R-squared:  0.1365,	Adjusted R-squared:  0.1185
F-statistic: 7.593 on 23 and 1105 DF,  p-value: < 2.2e-16
```

What is the difference in the effect of education between year 1972 and 1984?

POLICY ANALYSIS WITH POOLED CROSS SECTIONS - DIFFERENCE-IN-DIFFERENCE

Policy Analysis with Pooled Cross Sections

- We can use this type of analysis to analyze the impact of policy changes on outcomes of interest.
- An empirical example from Wooldridge
  - – Discussion of building a garbage incinerator in North Andover began in 1978 and construction began in 1981. Residents were concerned for the effect of the incinerator on housing prices.
  - – We will look at data on the price of houses sold in 1978 versus 1981. The hypothesis is that the price of houses located near the incinerator would fall relative to the price of more distant houses.
    - Two main variables: rprice - housing price in 1978 dollars and near – a dummy variable indicating if the house is within three miles of the incinerator.

Begin with a Naïve analysis

- We can simply use the 1981 data and regress the housing price on near.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   101308       3093  32.754  < 2e-16 ***
nearinc       -30688       5828  -5.266 5.14e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- How do we interpret the intercept? The coefficient on nearinc?
- Does this imply that the incinerator is causing lower home prices?
- What else could be going on?

Using the 1978 data

- When we run the same model with just the 1978 data we find a similar effect. Thus, even before there was talk of an incinerator the home values near the site were less.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)    82517       2654  31.094  < 2e-16 ***
nearinc       -18824       4745  -3.968 0.000105 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

So what now?

- How can we tell if the incinerator is actually depressing housing values. The key is to look at how the coefficient on nearinc changed between 1978 and 1981.
- The effect of nearinc was much larger in 1981
  - 30,688 – 18,824 = 11,864
- This value often referred to as the **difference-in-differences** estimator.
- We can test if this is significant by running the following model pooling the data over both years:

$$rprice = \beta_0 + \delta_0 y81 + \beta_1 nearinc + \delta_1 y81 * nearinc + u$$

$$rprice = \beta_0 + \delta_0 y81 + \beta_1 nearinc + \delta_1 y81 * nearinc + u$$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)    82517       2727  30.260  < 2e-16 ***
y81            18790       4050   4.640 5.12e-06 ***
nearinc       -18824       4875  -3.861 0.000137 ***
y81:nearinc   -11864       7457  -1.591 0.112595    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- The intercept, β_0, is the average price of a home not near the incinerator in 1978.
- The parameter, δ_0, captures the changes in all housing values in North Andover between 1978 and 1981.
- The coefficient on *nearinc*, measures the location effect that is not due to the presence of the incinerator , i.e. it is the effect we saw in the 1978 regression(before there was any discussion of the incinerator).
- The parameter of interest is the interaction between *y81* and *nearinc*. It measures the decline in housing values due to the new incinerator (assuming there are not other reasons that could account for the decline during those years).

- We actually do not find that the incinerator was significant, however, if we add controls to the model, such as the age of the house, number of rooms, size of the plot etc… the effect is significant. Again, we always need proper model specification.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.381e+04  1.117e+04   1.237  0.21720    
y81          1.393e+04  2.799e+03   4.977 1.07e-06 ***
nearinc      3.780e+03  4.453e+03   0.849  0.39661    
age         -7.395e+02  1.311e+02  -5.639 3.85e-08 ***
agesq        3.453e+00  8.128e-01   4.248 2.86e-05 ***
intst       -5.386e-01  1.963e-01  -2.743  0.00643 ** 
land         1.414e-01  3.108e-02   4.551 7.69e-06 ***
area         1.809e+01  2.306e+00   7.843 7.16e-14 ***
rooms        3.304e+03  1.661e+03   1.989  0.04758 *  
baths        6.977e+03  2.581e+03   2.703  0.00725 ** 
y81:nearinc -1.418e+04  4.987e+03  -2.843  0.00477 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Why the difference in difference estimator works

- It operates as a natural experiment. An exogenous event that changes the way the entities being studied operate.
- In such experiments, there is a control group that is not affected by the policy change and a treatment group that is.
  - Note, because assignment was not random, we need data at two time points (one prior and one after the policy change) to assess impact.

Difference in Differences cont…

- The regression model for the difference in differences estimator is the same as the categorical interaction model we discussed the other week.

$$y_i = \alpha_0 + \alpha_1 D_{1i} + \alpha_2 D_{2i} + \alpha_3 D_{1i} D_{2i} + \beta x_i + \varepsilon_i$$

- In the above interaction model, we had four separate groups, white men, white women, non-white men and non-white women
- In the D-in-D model we still have four groups: control group before policy change, control group after policy change, experimental group before policy change and experimental group after policy change.

Another Example

- Assume we are interested in whether an AIDS prevention program in Pennsylvania was effective and we had data on a random sample of residents at two time points (one before and one after the program).
- Simply looking at the before and after rates of HIV/AIDS in Pennsylvania will not work if all US states were experiencing decreases in AIDS cases.
- To overcome this issue, suppose you have second state (say New Jersey) with data on AIDS from corresponding time points that is arguably very similar except it was not exposed to the same policy program. But obviously, this state was still exposed to the changes in the economic and social conditions of the northeast US.
- The differences in changes between these two groups, the differences in differences, can be more confidently related back to the policy change.

Difference in Differences cont…

- The regression model can be written as follows:

$$y_i = \alpha_0 + \alpha_T T_i + \alpha_A AFTER_i + \alpha_{DD} T_i * AFTER_i + \varepsilon_i$$

- Where
  - Y_i = outcome of interest
  - T_i = 1 if in the treatment group, 0 otherwise
  - $After_i$ = 1 if after the policy change, 0 otherwise

Difference in Differences cont…

$$y_i = \alpha_0 + \alpha_T T_i + \alpha_A AFTER_i + \alpha_{DD} T_i * AFTER_i + \varepsilon_i$$

- We can look at the conditional expectations to understand how we interpret the regression coefficients.

- Mean outcome of **control group** **before policy**
  - $E(y_i | T=0, AFTER=0) = \alpha_0$
- Mean outcome of **control group** **after policy**
  - $E(y_i | T=0, AFTER=1) = \alpha_0 + \alpha_A$
- Mean outcome of **treatment group** **before policy**
  - $E(y_i | T=1, AFTER=0) = \alpha_0 + \alpha_T$
- Mean outcome of **treatment group** **after policy**
  - $E(y_i | T=1, AFTER=1) = \alpha_0 + \alpha_T + \alpha_A + \alpha_{DD}$

Difference in Differences cont…

$$y_i = \alpha_0 + \alpha_T T_i + \alpha_A AFTER_i + \alpha_{DD} T_i * AFTER_i + \varepsilon_i$$

- How do we interpret the coefficient on $\alpha_{DD?}$
- The change in the expected value of the outcome for the **control group** before and after the policy change is:
  - $E(y_i \mid T=0, After=1) – E(y_i \mid T=0, After = 0)$

    $= (\alpha_0 + \alpha_A) - \alpha_0$

    $= \alpha_A$
- The change in the expected value of the outcomes for the treatment group before and after the policy change is:
  - $E(y_i \mid T=1, After = 1) – E(y_i \mid T=1, After = 0)$

    $= (\alpha_0 + \alpha_T + \alpha_A + \alpha_{DD}) – (\alpha_0 + \alpha_T)$

    $= \alpha_A + \alpha_{DD}$

The difference in difference is simple: $(\alpha_A + \alpha_{DD}) – \alpha_A = \alpha_{DD}$

Note: that the before-after change for the treatment group consists of two parameters, which is why we cannot use just this difference to identify the effect of the policy. There are two things going on, the societal change and the policy change.

IN CLASS EXERCISE

In Class Exercise – Pooled Cross-Sectional Analysis

- Using the cps_inclass dataset, build a single regression model to assess:
  - Whether the gender gap in wages has increased or decreased between 1978 and 1985
  - Whether the return to education has changed between 1978 and 1985.
  - Include in your model the other following variables: y85 + exper + expersq + union

- The variables in the cps dataset are as follow:

1. educ	years of schooling
2. south	=1 if live in south
3. nonwhite	=1 if nonwhite
4. female	=1 if female
5. married	=1 if married
6. exper	age - educ - 6
7. expersq	exper^2
8. union	=1 if belong to union
9. lwage	log hourly wage
10. age	in years
11. year	78 or 85
12. y85	=1 if year == 85