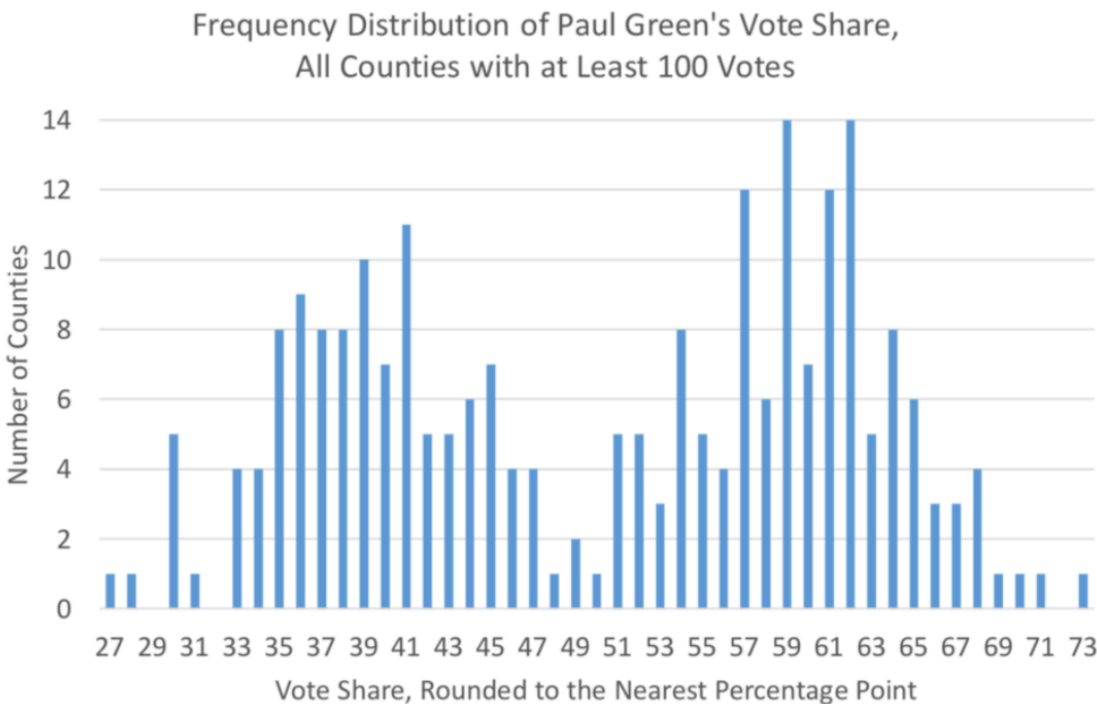


Sample in-class final exam questions for Regression and Other Stories

These questions are roughly in order of the material presented in the book.

1. You are designing a survey to estimate the gender gap: the difference in support for a candidate among men and women. Assuming the survey is a simple random sample of the voting population, how many people do you need to poll so that the standard error is less than 5 percentage points?
2. A teacher gives a midterm exam with possible scores ranging from 0 to 50 and a final exam with possible scores ranging from 0 to 100. A linear regression is fit, yielding the estimate $y = 30 + 1.2 \cdot x$ with residual standard deviation 10. Sketch (by hand, not using the computer) hypothetical data that could yield this fit.
3. In two sentences, give an example of a problem in social science where there has been insufficient attention to the problem of generalizing from observed measurements to the underlying constructs of interest.
4. The following graph was presented along with this explanation: “The histogram has two peaks, at 40% and 60% of the vote. These correspond to second/first ballot position, and imply a ballot order effect of roughly 20 percentage points . . .”



Give two ways this graph can be improved to display the relevant information more effectively.

5. A player takes 10 basketball shots, with a 40% probability of making each shot. Assume the outcomes of the shots are independent. Write a line of R code to compute the probability that the player makes exactly 3 of the 10 shots.

6. Out of a random sample of 50 Americans, zero report having ever held political office. From this information, give a 95% confidence interval for the proportion of Americans who have ever held political office.

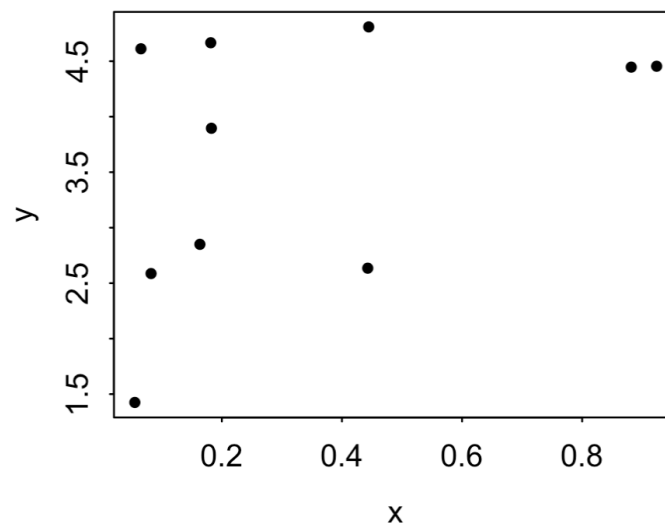
7. A survey is conducted of employees at a large company, with the goal of comparing the satisfaction levels of permanent and temporary employees. Suppose $X\%$ of the employees at the company are permanent and $100 - X\%$ are temporary, and suppose the survey is a simple random sample of N employees. The survey will have one question with a Yes or No response. Write an R function to simulate this process: the function should take the arguments X , N , and the true proportion of Yes responses in each of the two groups, and it should return the estimate and standard error for the difference in proportion Yes comparing the two groups.

8. Write an R function to: (i) simulate N data points from the model, $y = a + bx + \text{error}$, with data points x randomly sampled from the range $(0, 100)$ and with errors drawn independently from the normal distribution with mean 0 and standard deviation σ ; and (ii) fit the regression model to the simulated data. Your function should take as arguments, a , b , N , and σ , and it should return the data and print out the fitted regression.

9. Give R code to regress y on x with the intercept fixed at 0.

10. An experiment is performed comparing two groups of people and it will be summarized by the difference in the average responses for the two groups and a standard error for the difference. How can you compute this difference and standard error using regression?

11. A linear regression is fit to the data below. Which point has the most influence on the slope? Explain.



12. A linear regression is fit to data from high school students, modeling grade point average given household income. Write R code to compute the 90% predictive interval for the difference in grade point average comparing two students, one with household incomes of \$40,000 and one with household income of \$80,000.

13. A new job training program is being tested. Based on the successes and failures of previously proposed innovations, your prior distribution on the effect size on $\log(\text{income})$ is normal with a mean of -0.02 and a standard deviation of 0.05. You then conduct an experiment which gives an unbiased estimate of the treatment effect of 0.16 with a standard deviation of 0.08. What is the posterior mean and standard deviation of the treatment effect?

14. Here is the output from a fitted linear regression of outcome y on treatment indicator z , pre-treatment predictor x , and their interaction:

```
stan_glm
family:      gaussian [identity]
formula:     y ~ x + z + x:z
observations: 100
predictors:   4
```

```
-----
              Median MAD_SD
(Intercept)  1.2      0.2
x             1.6      0.4
z             2.7      0.3
x:z           0.7      0.5
```

```
Auxiliary parameter(s):
              Median MAD_SD
sigma 0.5      0.0
```

Give the estimated regression lines of y on x for the treatment group and the control group.

15. A linear regression is fit on a group of employed adults, predicting their physical flexibility given age. Flexibility is defined on a 0-30 scale based on measurements from a series of stretching tasks. Your model includes age in categories (under 30, 30-44, 45-59, 60+) and also age as a linear predictor. Make a graph of flexibility vs. age, showing what the fitted regression line might look like.

16. Many students and practitioners, when asked about the assumptions of linear regression, mention normality and equal variance of residuals. Why do we say that these are the *least* important assumptions of the model?

17. Anna takes continuous data x_1 and binary data x_2 and creates fake data y from the model, $y = a + b_1x_1 + b_2x_2 + b_3x_1x_2 + \text{error}$, and gives these data to Barb, who, not knowing how the data were constructed, fits a linear regression predicting y from x_1 and x_2 but without the interaction. In these data, Barb makes a residual plot of y vs. x_1 , using dots and circles to display points with $x_2 = 0$ and $x_2 = 1$, respectively. The residual plot indicates to Barb that she should fit the interaction model. Sketch the residual plot that Barb saw when she fit the regression without interaction.

18. A regression was fit to county x year data, predicting the rate of civil conflicts given a set of geographic and political predictors. Here are the estimated coefficients and their z -scores:

(Intercept)	−3.814*** (−20.178)
Pre-2000 Conflict	0.020† (1.861)
Border Distance	0.000 (0.450)
Capital Distance	0.000 (1.629)
Population	0.000* (2.482)
Pct Mountainous	1.641*** (8.518)
Pct Irrigation	−0.027† (−1.663)
GDP pc	−0.000*** (−3.589)

Why are the coefficients for border distance, capital distance, population, and per-capita GDP so small?

19. The following logistic regression has been fit:

```
family:      binomial [logit]
formula:     y ~ x + z
observations: 100
predictors:  3
-----
              Median MAD_SD
(Intercept) -1.9      0.6
x            0.7      0.8
z            0.7      0.5
```

Here, x is a continuous predictor ranging from 0 to 10, and z is a binary predictor taking on the values 0 and 1. Display the fitted model as two curves on a graph of $E(y)$ vs. x.

20. Recall the regression predicting incumbent party's two-party vote share from economic growth: $y = 46.2 + 3.1 \cdot \text{growth} + \text{error}$, where growth is in percentage terms and ranges from -0.5 to 4.5 in the data, and errors are approximately normally distributed with mean 0 and standard deviation 3.8. Suppose instead we were to fit a logistic regression, $\Pr(y=1) = \text{logit}^{-1}(a + b \cdot \text{growth})$. What would be the approximate estimates of a and b?

- a = -0.8, b = 1.1
- a = -1.6, b = 1.1
- a = -0.8, b = 2.2
- a = -1.6, b = 2.2

21. An educational intervention is hoped to increase scores by 5 points on a certain standardized test. An experiment is performed on N students, where half get this intervention and half get the control. Suppose that the standard deviation of test scores in the population is 20 points. Further suppose that a pre-test is available which has a correlation of 0.8 with the standardized test. What will be the standard error of the estimated treatment effect based on a fitted regression, assuming that the treatment effect is constant and independent of the value of the pre-test?

22. Consider the model used to predict yield of mesquite bushes:

```
fit <- stan_glm(log(weight) ~ log(canopy_volume) + log(canopy_shape) + group,
  data=mesquite)
```

We wish to use this model to make inferences about the average mesquite yield in a large population of trees which is summarized by a data frame called `new_trees`. Give R code to obtain an estimate and standard error for this population average.

23. The table below describes a hypothetical experiment on 8 people. Each row of the table gives a participant and her pre-treatment predictor x , treatment indicator z , and potential outcomes y_0 and y_1 .

Person	x	z	y_0	y_1
Anna	3	0	5	5
Beth	5	0	8	10
Cari	2	1	5	3
Dora	8	0	12	13
Edna	5	0	4	2
Fala	10	1	8	9
Geri	2	1	4	1
Hana	11	1	9	13

Give the average treatment effect in the population, the average treatment effect among the treated, and the estimated treatment effect based on a simple comparison of treatment and control.

24. You have fit a model,

```
fit <- stan_glm(y ~ z + age + z:age, family=binomial(link="logit"))
```

with binary outcome y , treatment indicator z , and age measured in years. Give R code to produce an estimate and standard error of average treatment effect in the population, given a vector `n_pop` of length 82 that has the number of people in the population at each age from 18 through 99, ignoring anyone in the population over that age.

25. In an observational study, it can be helpful to perform matching to find a subset of the data where treated and control units are similar on key variables. Which of the following statements is typically correct?

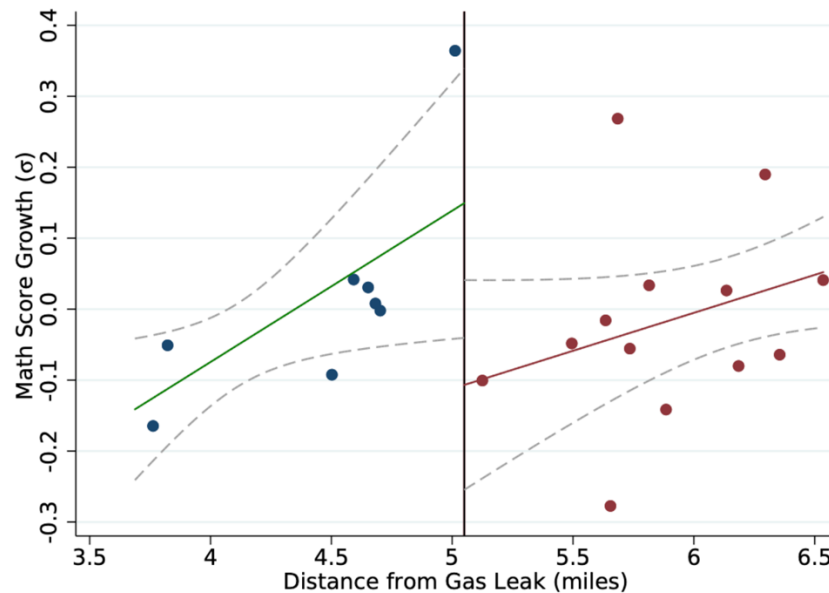
- When performing the analysis on the matched data, you should not just compare treated and control averages; you should fit a regression adjusting for key pre-treatment variables.
- It is a good idea for the matching to be proportional, so that the ratio of treatment to control units in the matched subset is approximately equal to the ratio of treatment to control units in the population.
- The set of variables used in the matching should not include any variables that could be affected by the treatment.
- The resulting analysis relies on the assumption of ignorability conditional on the variables included in the matching and regression.

26. The following study is performed at a university. Students are sent emails encouraging them to click on a university website. Each student is randomly assigned to one of two sites: a site with encouragement to vote in the upcoming student government election, and a neutral site with study tips. The students are then followed up to see if they voted. Define $y = 1$ if the student voted and 0 otherwise; define $u = 1$ if the student was assigned to the encouragement site and 0 if he or she was assigned to the neutral site; define $v = 1$ if the student actually accessed the site (which can be checked

using unique id's) and 0 if he or she never clicked on the link. From which of these regressions or pair of regressions can we compute the instrumental variables estimate of the effect of accessing the site on voting?

- Regression of y on u .
- Regression of y on v .
- Regression of y on u and v .
- Regression of y on u , and the regression of v on u .
- Regression of y on v , and the regression of v on u .

27. The following graph shows the result of a regression discontinuity model fit to data on schools located near a suspected gas leak. Schools less than 5 miles from the leak were given air filters, and the analysis purports to show that the filters had a large positive effect on test scores. Each dot on the graph shows the average change in test scores in a school.



Which of the following is a good explanation of what went wrong with this analysis?

- It was inappropriate to include all the data: the analyst should only have included the schools that are very close to the discontinuity.
- Fitting straight lines was a mistake: the analyst should have fitted nonparametric curves or a higher-degree polynomial.
- There is overfitting: the result cannot be trusted because it depends too strongly on the particular choices used in the model fit.
- Clustered data: the tests are of individual students and so it is inappropriate to use school averages.

28. Suppose you fit a regression model with a large number of predictors, and you want to regularize so that the estimates are more stable. Which of these is *not* a way to do this?

- Give the coefficients independent prior distributions with mean 0 and standard deviations that are near 0.
- Give the coefficients independent prior distributions with mean 0 and standard deviations that are large but finite.
- Reparameterize the model so that the coefficients can be modeled hierarchically.
- Use cross-validation to choose a model.