A06 Answers

David Housman

2022-10-04

**Instructions.** Change the author name on line 3 to be your name. Insert all requested answers. Upload in Moodle a zip file containing this file, a knitted file (html, doc, or pdf), and the various image files that came with this assignment. You may also include the modification of the *Chapter06\_Hand\_Plots.pdf* file containing your answers, or you may turn that in on paper.

### Exercise 1 (72 points).

Answer from *Statistical Modeling: End-of-Chapter Exercises* by Daniel Kaplan the following. I provide enough of each problem (Prob) or reading exercise (Read) so that it should be relatively easy to type your answers (text or code) in this file. I also provide some clarifications and hints. Note that Kaplan does not always include the correct value in his multiple choice questions!

**Prob 6.01 (2 points).** In McClesky vs Georgia, lawyers presented data showing that for convicted murderers, a death sentence was more likely if the victim was white than if the victim was black. For each case, they tabulated the race of the victim and the sentence (death or life in prison). Which of the following best describe the variables their models?

**Prob 6.02 (8 points–1 point for each answer).** In studies of employment discrimination, several attributes of employees are often relevant: age, sex, race, years of experience, salary, whether promoted, whether laid off. For each of the following questions, indicate which is the response variable and which is the explanatory variable.

1. Are men paid more than women?

Response Variable:

Explanatory Variable:

1. On average, how much extra salary is a year of experience worth?

Response Variable:

Explanatory Variable:

1. Are whites more likely than blacks to be promoted?

Response Variable:

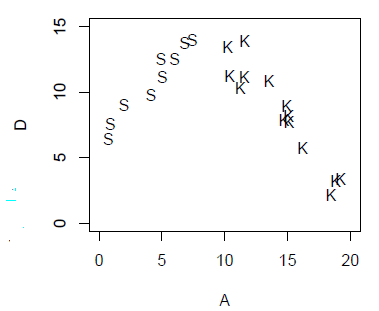
Explanatory Variable:

1. Are older employees more likely to be laid off than younger ones?

Response Variable:

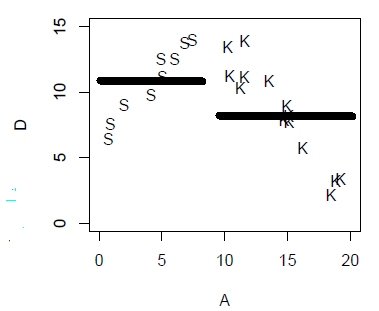
Explanatory Variable:

**Prob 6.03 (18 points–3 points for each answer).** The scatter plot shows some data involving three variables



* D is a quantitative variable
* A is a quantitative variable
* G is a categorical variable with two levels: S & K

Here is a sketch of the function for the model D ~ G:



Sketch similar functions for each of the following models:

1. D ~ A+G
2. D ~ A-1
3. D ~ A
4. D ~ A\*G
5. D ~ 1
6. D ~ poly(A,2)

The *Chapter06\_Hand\_Plots.pdf* file contains six copies of the scatter plot. You may print this file onto paper, sketch your answers, and turn in the piece of paper. Alternatively, you may draw on the document electronically and upload the electronic file with your other work.

**Prob 6.04 (12 points–3 points for each answer).** Using your general knowledge about the world, think about the relationship between these variables:

* speed of a bicyclist.
* steepness of the road, a quantitative variable measured by the grade (rise over run). 0 means flat, + means uphill, - means downhill.
* fitness of the rider, a categorical variable with three levels: unfit, average, athletic.

On a piece of paper (you could use the *Chapter06\_Hand\_Plots.pdf* file), sketch out a graph of speed versus steepness for reasonable models of each of these forms:

1. Model 1: speed ~ 1 + steepness
2. Model 2: speed ~ 1 + fitness
3. Model 3: speed ~ 1 + steepness + fitness
4. Model 4: speed ~ 1 + steepness + fitness + steepness:fitness

**Prob 7.03 (10 points–1 point for each answer).** For some simple models, the cofficients can be interpreted as grand means, group-wise means, or differences between group-wise means. In each of the following, A, B, and C are quantitative variables and color is a categorical variable with levels red, blue, and green.

1. The model A ~ color gave these coefficients:

| term | coefficient |
| --- | --- |
| Intercept | 10 |
| colorBlue | 5 |
| colorGreen | 12 |

* What is the mean of A for those cases that are Blue?
* What is the mean of A for those cases that are Green?
* What is the mean of A for those cases that are Red?
* What is the grand mean of A for all cases?

1. The model B ~ color - 1 gave these coefficients:

| term | coefficient |
| --- | --- |
| colorRed | 100 |
| colorBlue | -40 |
| colorGreen | 35 |

* What is the group mean of B for those cases that are Blue?
* What is the group mean of B for those cases that are Red?
* What is the group mean of B for those cases that are Green?
* What is the grand mean of B for all cases?

1. The model C ~ 1 gave this coefficient:

| term | coefficient |
| --- | --- |
| Intercept | 4.7 |

* What is the group mean of C for those cases that are Blue?
* What is the grand mean of C for all cases?

**Prob 7.04 (12 points–3 points for each answer).** Rather than stating the modeling statement, include a code chunk that executes the appropriate lm command.

1. From the CPS85 data, what is the mean age of single people?
2. From the CPS85 data, what is the difference between the mean ages of married and single people?
3. From the SwimRecords data, what is the mean swimming time for women?
4. From the Utilities data, what is the mean CCF for November?

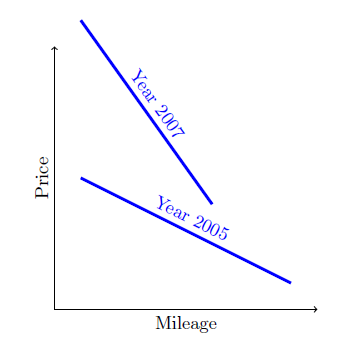
**Prob 7.07 (6 points–2 points for each answer).** In the SAT data, the variables have these units:

* sat has units of “points.”
* expend has units of “dollars.”
* ratio has units of “students.”
* frac has units of “percentage points.”

Consider the model formula sat = 994 + 12.29 expend - 2.85 frac.

1. What are the units of the coefficient 994?
2. What are the units of the coefficient 12.29?
3. What are the units of the coefficient 2.85?

**Prob 7.08 (6 points–2 points for each answer).** The graph shows schematically a possible relationship between used car price, mileage, and the car model year.



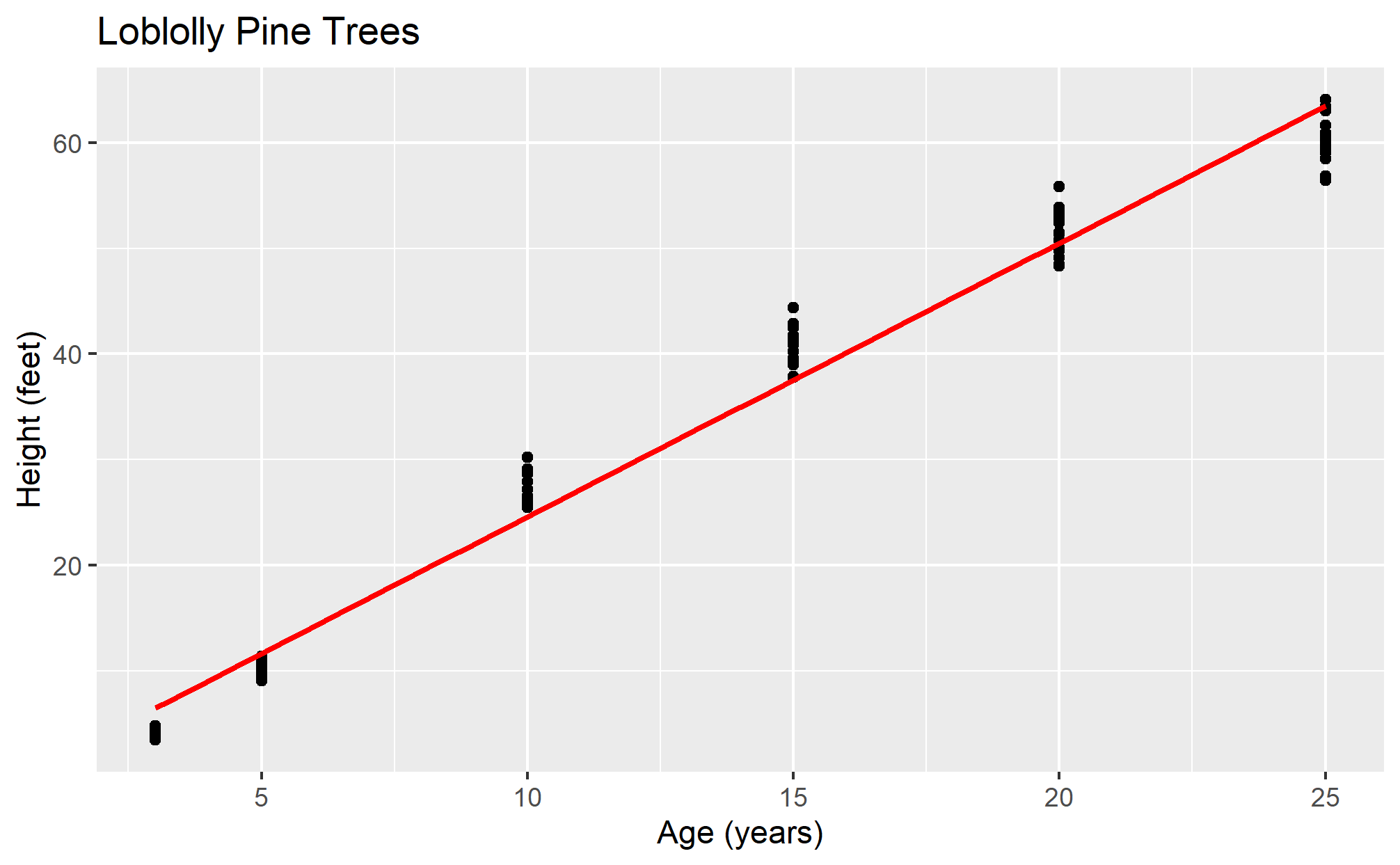
Consider the model price ~ mileage\*year.

In your answers, treat year as a simple categorical variable, and use year 2005 as the reference group when thinking about coefficients.

1. What will be the sign of the coefficient on mileage?
2. What will be the sign of the coefficient on model year?
3. What will be the sign of the interaction coefficient?

### Exercise 2 (20 points).

Consider the scatter plot below with the overlaid best fit straight line model. Interpret the lowest point (6 points), estimate an equation for the straight line (6 points), provide an interpretation of the coefficients in your equation (6 points), and estimate the standard deviation of the residuals (2 points).



### Exercise 3 (51 points).

With the KidsFeet data, do the following.

1. (10 points) Obtain a scatter plot of foot width vs. foot length with the points different colors for boys and girls. Overlay on the plot straight line models for each sex. This will be similar to the plot provided by Kaplan in Prob 7.05 but with straight lines rather than points for the two models.
2. (15 points) Obtain the best fit formula for the model width ~ length (5 points). Provide a meaningful interpretation of the formula (6 points), residual standard error (2 points), and R-squared (2 points).
3. (21 points) Obtain the best fit formula for the model width ~ sex \* length(5 points). Provide a meaningful interpretation of the formula (12 points), residual standard error (2 points), and R-squared (2 points).
4. (5 points) Discuss which of the two previous models is a better representation of the data.