Problem set 3

S681

Upload your typed answers to Canvas as a PDF by 11:59 pm, Sunday 24th February. Include R code where applicable.

1. (5 points.) Using the data set `BGSgirls` in package `alr4`:

   (a) Fit a linear regression model to predict girls' weight at age 18 (variable `WT18`) using weight at age 2 (`WT2`) and weight at age 9 (`WT9`) as the regressors. Display the resulting model in a format that someone who has not used R can understand.

   (b) The model with both `WT2` and `WT9` as the regressors has a negative coefficient for `WT2`. A friend sees this and says, "The negative sign means that girls who are heavier than average at age 2 will usually be heavier than average at age 18." Patiently explain why your friend is mistaken, and give a correct interpretation of the negative sign.

   (c) The model with both `WT2` and `WT9` as the regressors has a coefficient of 1.2 for `WT9`. A friend sees this and says, "If two girls have a one pound difference in weight at age 9, the model predicts they'll have a 1.2 difference in weight at age 18." Is your friend correct? Why or why not?

2. (10 points.) The data set `MinnLand` in package `alr4` contains data on "nearly every farm sale" in six economic regions in Minnesota from 2002 to 2011. Suppose we wish to model how sale price per acre (`acrePrice`) depends on `year`. Since sales price per acre is strongly right-skewed, we'll take `log(acrePrice)` as the response in our regressions.

   (a) Fit a linear regression model to predict `log(acrePrice)` from `year` alone, taking year as a continuous variable. Write down the regression equation you obtain.

   (b) Fit a regression model to predict `log(acrePrice)` from `year` alone, taking year as a *factor*. State the coefficient for the year 2008, and explain what this coefficient means.

   (c) Each of these two models can be used to (retrospectively) predict the expected log of sale price per acre from 2002 to 2011. *Plot* these predictions for the two models, and describe the differences.

   (d) Which of these two models fits the data better? Support your answer using graphs or otherwise.

3. (10 points.) The data set `Moore` in the package `carData` contains data from an experiment to see how conformity with someone else's opinion was related to the other person's status. Subjects were paired with a partner of either high or low status; the partners were secretly collaborators of the investigators. On 40 key questions, the partners were told to disagree with the subjects. The experimenters counted the number of times each subject "conformed" by

changing their opinion to agree with their partner. Each subject was also (presumably before the experiment) given a questionnaire to measure their authoritarianism, as authoritarianism could potentially affect how the subject reacted to disagreement.

The variables in the data set are:

- `conformity`: number of conforming responses—could potentially be 0 to 40; observed values ranged from 4 to 24
- `partner.status`: a factor: `high` or `low`
- `fscore`: authoritarianism score—observed values ranged from 15 to 68

The data frame also includes `fcategory`, a categorized version of `fscore`; ignore this.

(a) Show that there's evidence that `partner.status` affects `conformity`. (This might not require any regression...)

(b) Does the effect of `partner.status` differ for people with different `fscore`s? One way to look at this is to fit a linear model with `conformity` as the response and `fscore`, `partner.status`, and their interaction as regressors. Fit this model, give the P-value, and explain what the P-value means and what it tells you about whether the effect of `partner.status` differs for people with different `fscore`s.

(c) A broader question is *how* the effect of `partner.status` differs for people with different `fscore`. This is perhaps easiest to study graphically. Using your model in (b), make predictions for `conformity` for people with `fscore`s ranging from 15 to 68, for both the high status and low status treatments. Plot these predictions on the same graph, clearly distinguishing between the lines for the high and low status groups (e.g. by color.) Assuming your model is close to right, what does this graph tell you about how the effect of `partner.status` differs for people with different `fscore`s?

4. (10 points.) The data set `cakes` contains data from a baking experiment using packaged cake mix. The response, `Y`, is a "palatability score" (higher is tastier.) The explanatory variables are `X1`, baking time in minutes, and `X2`, baking temperature in degrees Fahrenheit. (Ignore the `block` variable.)

   (a) Show graphically that it is *not* appropriate to model expected palatability score as a linear function of `X1` and `X2`. Explain why we should have known this even before we looked at the data.

   (b) Fit a model to predict palatability score as the sum of quadratic functions of baking time and baking temperature. (For simplicity, we recommend you do not fit any interaction.) Display the fitted model graphically, e.g. through colored or faceted plots.

   (c) For how long and at what temperature should you bake a cake using this mix to maximize the predicted palatability? (Hint: Recall from Calc I that a quadratic $ax^2 + bx + c$ is maximized at $-b/(2a)$ if a is negative.)

5. (10 points.) Returning to the `MinnLand` data set, one subject the data was collected to answer was the relationship between sale price per acre and `crpPct`, the percentage of the land enrolled in the Conservation Reserve Program. However, there are many potential

confounding variables associated with `crpPct` that could affect sale prices. For example, land in the Conservation Reserve Program is disproportionately in northwest Minnesota, and sale prices in northwest Minnesota tend to be lower than in the rest of the state for reasons that may have less to do with the program than with negative temperatures in the winter.

One way to study this would be to fit models that include both `crpPct` and `region` as predictors. However, it is not clear a priori whether an interaction between `crpPct` and `region` will help.

(a) Fit a linear regression model to predict `log(acrePrice)` from `crpPct` and `region`, with no interaction. State the coefficient of `crpPct` in this model, and explain what this coefficient tells you about the relationship between `crpPct` and `log(acrePrice)`.

(b) Fit a regression model to predict `log(acrePrice)` from `crpPct` and `region` with an interaction. Explain what this model tells you about the relationship between `crpPct` and `log(acrePrice)`.

(c) Perform an ANOVA to compare your models from parts (a) and (b). State the P-value that you get, and explain what, if anything, this P-value tells you.

(d) Your ANOVA in part (c) made certain assumptions. Check the residuals of your model from (b) to see if these assumptions are close to satisfied.

6. (10 points.) The data set `BigMac2003` in `alr4` gives the price of a Big Mac in 2003 (`BigMac`), measured in minutes of labor required to buy one, in 69 cities. Of the many potential explanatory variables in the data set, `FoodIndex`, a measure of food prices (relative to a baseline where Zurich is 100), both logically makes sense as a predictor and has a fairly strong correlation with `BigMac`. We thus wish to first try a model to predict `BigMac` from `FoodIndex`, but these variables may require transformation.

   (a) Choose interpretable transformations to apply to `FoodIndex` and `BigMac`, such that the relationship between the transformed variables is approximately linear. (Note that you may choose "no transformation" for either variable.) Justify your choice using graphs or otherwise.

   (b) The model can straightforwardly be improved by adding another predictor. Fit a better model that predicts the transformed `BigMac` variable from `FoodIndex` and *one* other variable. Convince the grader that your model is an improvement. (Your model may include complex regressors such as interactions if you wish.)

   (c) Using numbers, graphs, and words, explain what your improved model tells you about how the price of Big Macs relates to the Food Index and your other explanatory variable.

7. (5 points.) The data set `BGSall` in `alr4` gives measurements on all subjects of the Berkeley Guidance Study, both male and female. Our goal is to find the model that best predicts height at age 18 (`HT18` gives this height in cm) using measurements available at age 9: `Sex`, `WT2`, `HT2`, `WT9`, `HT9`, `LG9`, and `ST9`. See `?BGSall` for definitions of all these variables.

   Find the best predictive model you can. You may *not* transform `HT18` but you may transform any of the predictors. You may also consider interactions and higher-order terms. AIC isn't the be-all and end-all, but I managed to get a model with an AIC of 708.1 without looking too hard, so your model's AIC should get close to that. In addition, you should give some measure of how large you would expect the prediction errors if your model was applied to individuals similar to those in the data set (children born in California in 1928–29.)