Homework 4. Due Friday Sept 29, 3:00

1. (10pts) Recall the regression which goes through the origin:

$$Y_i = \beta X_i + \epsilon_i$$

where ϵ_i's are independent Gaussian random variables. You have shown that the least squares estimate $\widehat{\beta}$ of β is unbiased. Show that $\widehat{\beta}$ has a Gaussian distribution, and find the mean and variance.

Questions 2–3 use the dataset:

`http://www.stat.cmu.edu/~larry/=stat401/auto-mpg.csv`

which comes from the 1983 American Statistical Association Exposition. The response variable of interest is fuel consumption, measured in miles per gallon. Other attributes of cars like the weight, horsepower, number cylinders, and acceleration time were also recorded for each car. We will study the relationship between `mpg` and `weight` (in lbs).

2. *(20pts) Diagnostics and Transformations*

   (a) (6pts) Someone argues that a linear regression model would be a good fit. Make a residual plot of e_i versus the fitted values. What does the plot suggest about the linearity assumption of the regression model?

   (b) (8pts) Apply the log transformation on `weight`, refit the linear regression model, and produce a new residual plot. What does the plot suggest about the linearity assumption now? What else does it suggest? What assumptions are violated?

   (c) (6pts) Prepare a qq-plot of the residuals in (b). Do the residuals appear to be Normally distributed?

3. *(20pts) Answer with respect to the final model you chose in the previous question.*

   (a) (5pts) How do you interpret the estimated intercept and slope?

   (b) (10pts) Test whether or not there is a linear association between `mpg` (after the transformation you selected, if any) and `weight` (after the transformation you selected, if any), using $\alpha = 0.05$. State the alternative hypothesis, decision rule, and conclusion. What is the p-value of the test?

   (c) (5pts) Find a 90% confidence interval for β_1. How do you interpret it?

4. **Data Analysis Practice (50pts):** In practice, data analysis involves more than just running a model and turning in the output. You need to be able to describe the problem, choose the right analyses, interpret your results, and explain them to an audience that may or may not know advanced statistics.

   Research Problem: Abalones, also called ear-shells or sea ears, are one type of reef-dwelling marine snails. The flesh of abalones is widely considered to be a desirable food, and is consumed raw or cooked in a variety of cultures [1]. It is difficult to tell the ages of abalones because their shell sizes not only depend on how old they are, but also depend on the availability of food. The study of age is usually by obtaining a stained sample of the shell and looking at the number of rings through a microscope. A research group are interested in using some of abalones' physical measurements, especially the height measurement to predict their ages. The research group believe that a simple linear regression model with normal error assumption is appropriate to describe the relationship between the height of abalones and their ages, and particularly, that a larger height is associated with an older age. Our data set is

   `http://www.stat.cmu.edu/~larry/=stat401/abalone.csv`

   More information on this data set is available from `https://archive.ics.uci.edu/ml/datasets/Abalone`.

   The following points are an example template for analyzing the data set. Your answers should be always given in the context of the problem, rather than abstractions like "the independent variable". Your language/word choices should be clear, concise, and scientifically accurate. Don't use phrases like "I think". Don't claim results that aren't true or for which you lack evidence; in particular remember that association is not the same thing as causation. **Answer the following in a report of around 3 pages.**

   - Write two/three sentences introducing the research problem and describing the research hypothesis. Cite any information sources in parentheses.
   - Examine the two variables individually (univariate). Find summary measures of each (mean, variance, range, etc). Graphically display each; describe your graphs. What is the unit of `height`?
   - Generate a labeled scatterplot of the data. Describe interesting features/trends.
   - Fit a simple linear regression to the data predicting number of rings using height of the abalones

[1] https://en.wikipedia.org/wiki/Abalone

- Generate a labeled scatterplot that displays the data and the estimated regression function line (can add to the previous scatterplot). Describe the line's fit.
- Do diagnostics to assess whether the model assumptions are met; if not, appropriately transform height and/or number of rings and re-fit your model. Justify your decisions (and re-check your diagnostics).
- interpret your final parameter estimates in context. Provide 95% confidence intervals for β_0, β_1. Interpret in context of problem.
- Is there a statistically significant relationship between the height and the number of rings (and hence, the age) of abalones? Explain in context of problem.
- Find the point estimate and the 95% confidence interval for the **average** number of rings for abalones with height at 0.128 (in the same unit as other observations of height). Interpret in context.
- We are interested in predicting the number of rings for an abalone with height at 0.132 (in the same unit as other observations of height). Find the predicted value and a 99% prediction interval.
- What are your conclusions? Identify a key finding and discuss its validity. Can you come up with any reasons for what you see? Do you have any suggestions or recommendations for the researchers? How could this analysis be improved? (6–8 sentences total)