BST 760: Advanced Regression
Breheny

Assignment 4
Due: Thursday, February 17

1. Show that $\left[\hat{\beta}_j - t_{\alpha/2,n-p}\widehat{\text{SE}}, \hat{\beta}_j + t_{\alpha/2,n-p}\widehat{\text{SE}}\right]$ is a $(1-\alpha)\times 100\%$ confidence interval for β_j (*i.e.*, show that the probability that the interval contains β_j is $1-\alpha$).

2. For the alcohol metabolism data set that we looked at in lab, create a variable called `Gastric2` that is highly correlated ($r > .99$) with `Gastric`. For interpretation's sake, `Gastric2` could represent a different way of measuring alcohol dehydrogenase activity than was used in `Gastric`.

   (a) Fit a model with `Sex`, `Alcohol`, and `Gastric` as the explanatory variables (no interactions). Give a 95% confidence interval for the effect (on alcohol metabolism) of increasing alcohol dehydrogenase activity by 1 unit.

   (b) Add `Gastric2` to the model. Give a 95% confidence interval for the effect of increasing `Gastric` by 1 unit while keeping `Gastric2` constant.

   (c) Using the same model, give a 95% confidence interval for the effect of increasing both `Gastric` and `Gastric2` by 1 unit.

   (d) Briefly (one or two sentences), explain in a qualitative way how your answers for (a), (b) and (c) relate to each other and why they make sense.

3. For the alcohol metabolism data set that we looked at in lab, fit a model that allows `Gastric` to have a different effect on alcohol metabolism for alcoholics as it does for non-alcoholics (leave `Sex` out of the model). Base your answers to the following questions on this model, and support your answers with statistics; for example, a hypothesis test or confidence interval.

   (a) Create a trellis plot of alcohol metabolism versus alcohol dehydrogenase, with separate panels for alcoholics and non-alcoholics.

   (b) Does alcohol dehydrogenase have an effect upon alcohol metabolism for alcoholics?

   (c) Does alcohol dehydrogenase have an effect upon alcohol metabolism for non-alcoholics?

   (d) Is there a difference in the effect of alcohol dehydrogenase upon alcohol metabolism between alcoholics and non-alcoholics?

   (e) Do non-alcoholics have higher alcohol metabolism than alcoholics, assuming both groups have an alcohol dehydrogenase activity of 2?

   (f) For the comparison in (e), quantify the difference and provide a confidence interval.

   (g) Repeat (e) for an alcohol dehydrogenase level of 4.

   (h) For the comparison in (g), quantify the difference and provide a confidence interval.

4. Recall that $\hat{\boldsymbol{\mu}} = \mathbf{H}\mathbf{y}$, and therefore, each $\hat{\mu}_i$ is a linear combination of the $\{y_i\}$.

   (a) Show that if the design matrix has an intercept, then all the rows and columns of $\mathbf{H}$ add up to 1 (Hint: recall that $\mathbf{H}\mathbf{X} = \mathbf{X}$).

(b) Given that the rows of $\mathbf{H}$ add up to 1, a way of thinking about $\hat{\mu}_i$ is that it represents a weighted average of the $\{y_i\}$ – instead of the (unweighted) average $\bar{y} = \sum_i \frac{1}{n} y_i$, $\hat{\mu}_i = \sum w_j y_j$, where the weights $\{w_j\}$ come from the projection matrix $\mathbf{H}$. These weights add up to 1, just like the $\frac{1}{n}$'s do, but don't place equal weight on each y_j. Fit a regression model to one of the data sets on the course website. Pick an observation i from the data set. Which three observations have the highest weights $\{w_j\}$ in determining $\hat{\mu}_i$? Why did these observations get the highest weights? Which three have the lowest weights? Why?