Homework 7

Advanced Methods for Data Analysis (36-402)

Due Friday March 22, 2019, at 3:00 PM

See the syllabus for general instructions regarding homework. Note that you should **always show all your work** and submit **exactly two files** (Rmd or R file as *code*, and knitted or scanned/merged pdf or html as *writeup*). Make sure that everything you submit is readable.

Problem 1: Cat Data Again

To load the data run:

```
library(MASS)
data(cats)
```

Part a

Fit a linear regression model for `Hwt`, in which `Bwt` interacts with `Sex` and the intercept is forced to be zero. Why is the forcing the intercept to be zero a reasonable thing to do?

Hint: There are some subtleties regarding how to force the intercept to be zero, make sure you are not accidentally introducing intercept terms by using the wrong `R` syntax.

Part b

Under the assumption that the residuals are normally distributed, conduct a statistical test to find out whether there is evidence that the slope coefficient for `Bwt` differs between female and male cats. Make sure you state the hypotheses (null and alternative), the null distribution, the value of the test statistic and the p-value.

Part c

Answer the same hypothesis as in Part b, but this time perform a permutation test. You might find previous homeworks and lecture notes on permutation tests useful.

Part d

Recall that in HW 6, Q3-Part c, the test using the boostrap rejected the null hypothesis at $\alpha = 0.05$. Compare that result to the one you obtained in Part c of this homework. If in Part c of this homework, you did not reject the null hypothesis at $\alpha = 0.05$, propose an explanation for why you reached a different conclusion than in HW 6.

Hint: Think about what the null hypothesis was in HW 6 and what the null is here.

Part e

Look back at HW 6, Q3-Part c and check what the residual terms were in the null model. Is there a difference between the null models from HW 6, Q3-Part c and the null model you fitted in Part a of this homework?

Problem 2: Problem 3 from HW 5 Revisited

For this problem, we use the `abalone` data set, which contains nine variables measured on 4173 abalones. Take a look at HW 5 for a description of the variables. Recall that, in HW 5, we fitted the following models:

Model 1: A linear regression of log(`Shucked.weight`), on the logarithms of all three predictors.

Model 2: A kernel regression of `Shucked.weight`, on all three predictors: `Diameter`, `Length`, and `Height`.

Part a

Fit a smoothing spline to predict `Shucked.weight` from `Diameter` times `Length` times `Height`. Divide the product by 10^6 so that it is measured in dm^3. Just like in the previous HW, we will compute 95% pivotal confidence intervals for the regression function $r(x)$ at each x from 0 to 140 in steps of 2 (67 different values of x). Call this vector of x values `x0`. Note that the `smooth.spline` command in R will internally choose the λ parameter by GCV, but suppose we don't want to vary λ in each bootstrap sample. In later lectures, we will discuss how to properly tune λ, for now use $\lambda = 0.00004$. For each bootstrap sample b, fit a smoothing spline, and predict the response `Shucked.weight` at each of the 67 values in `x0`. (We can call these $\hat{r}_b^*(x)$.) The command

```
pred=predict(ssmod,x0)$y
```

will produce the predictions requested, if the original fit is stored as `ssmod`. For each $x \in$ `x0`, compute a 95% pivotal bootstrap confidence interval for $r(x)$. Draw a plot with the original data and the estimated regression function $\hat{r}(x)$ for each x in `x0` added as a line. Finally, add to the plot the upper and lower endpoints of all of the pivotal bootstrap confidence intervals (using a different line type than used for $\hat{r}(x)$.)

Part b

Now let's consider Model 1 again. Recall that one issue with transforming the response in a regression is the following. We defined $r(x)$ to be $\mathbb{E}[Y \mid X = x]$, the conditional mean of Y given $X = x$. In Model 1, we modeled

$$\log(Y) = \tilde{r}(x) + \varepsilon, \tag{1}$$

where ε is independent of X and has a distribution centered at 0. However, if the mean of ε is 0, $\exp(\tilde{r}(x))$ is *not* the conditional mean of Y given $X = x$. We can still use $\exp\left(\widehat{\tilde{r}(x)}\right)$ as an estimator of $r(x) = \mathbb{E}[Y|X = x]$, but it will be biased. In the model of equation (1) assume only that the cases, i.e. (X, Y) pairs, are iid. Use the appropriate version of bootstrap to get an estimate of the bias of $\exp\left(\widehat{\tilde{r}(x)}\right)$ as an estimator of $r(x)$ for each three-dimensional vector x of predictors as defined by the rows of the following table:

Length	Diameter	Height
600	500	150
400	350	125
250	200	140

Say what assumption you are making about how the distribution of $\exp\left(\widehat{\tilde{r}(x)}\right)$ relates to the distribution of the bootstrap calculations $\exp\left(\widehat{\tilde{r}(x)}_b^*\right)$. Offer a reason for why the biases seem so different.

Note: In HW5 Part 3g (the extra credit question), we approximated $e^{\widehat{\tilde{r}(x)}}$ via a Taylor expansion so that it was easier to compute its variance, and hence a confidence interval for $e^{\tilde{r}(x)}$. Here the bootstrap procedure allows us to see the bias from centering the CI for $r(x)$ at $e^{\widehat{\tilde{r}(x)}}$.

Have a great spring break!