Homework 10
Statistics 200B
Due April 25, 2019

1. Derive the expression for the variance of $\hat{\beta}$ in simple linear regression, given in equation (13.11) in Wasserman, using the multivariate normal distribution for $\hat{\beta}$ we found in class (page 163 of the notes).

2. Assume a multiple linear regression model with normal errors. Take σ to be known. Show that the model with the highest AIC is the model with the lowest Mallows C_p statistic.

3. Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

with $E[\epsilon_i] = 0$ and $V[\epsilon_i] = \sigma^2$. Under the following two cases, find what happens to the estimates, standard errors, and Wald test statistics for β_0 and β_1.

   (a) Instead of X, we regress Y on a new variable $Q = aX + b$.

   (b) Instead of Y, we construct a new variable $R = aY + b$ and regress this on X.

4. Consider the multiple regression model with k possible predictors. For a particular model, let $S \subseteq \{1, \ldots, k\}$ denote the indices of the included regressors. Prove that

$$R^2(S) = \frac{\sum_{i=1}^n (\hat{Y}_i(S) - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

can not decrease by adding an additional term to S.

5. Use `load("hitters.RData")` to load the data frame `hitters` into `R`. This data frame contains information about baseball player salaries in the 1987 season, as well as a number of possible predictors of salary, described below.

   (a) Using the `stepAIC` function in `R` (you will first need to call `library(MASS` to load the package with this function), use backward stepwise selection to choose two models for predicting the log of the 1987 salary, one based

on AIC and one based on BIC. (*Hint: look at the help page for* `stepAIC` *to see how to "trick" stepAIC into using BIC instead.*) What variables are included in the chosen models for each? How do they compare?

(b) Write a function in `R` to compute the leave-one-out cross-validation risk estimator for a fitted model `mod` obtained using `lm`. *Hint:* `lm.influence(mod)$hat` *returns the diagonal elements of the matrix U in (13.30) of Wasserman.* Use your function to compute the risk estimates for the models selected via AIC and BIC from part (a). Which model is preferred by this criterion?

(c) Look at the p-values for the individual t-tests being computed for the full model (with all possible regressors) to the model selected by BIC. You should see that many of the individual t-tests are significant relative to the smaller model, but are not significant in the larger model. Why do you think this occurs?

Description of the baseball data: The Statistical Graphics and Statistical Computing Sections of the American Statistical Association sponsor a bi-annual "Data Exposition" (`http://stat-computing.org/dataexpo`), for which this data appeared in 1988. The data as I present it here incorporates some corrections and variable transformations as described in "Applied Regression Analysis and Generalized Linear Models" by John Fox.

In addition to salary (`salary1987`), the data set contains the following variables for the 1986 season: at bats (`AB86`), hits (`H86`), home runs (`HR86`), runs scored (`R86`), runs batted in (`RBI86`), walks (`W86`), put-outs (`PO86`), assists (`A86`), errors (`E86`), batting average (`AVG86`), and on-base percentage (`OBP86`). The next eight variables in the data frame are equivalent to the first eight 1986 variables, but over the course of the player's professional career. Next we have number of years in the major league (`years`), and the same eight variables but on a per-year basis (each career variable divided by (`years`). The next four variables are indicators for position: middle infielders (`MI`), catchers ((`C`), center fielders (`CF`), and designated hitters (`DH`). (Note that this dataset does not contain pitchers.) After 3 years in the major leagues, players are eligible for salary arbitration, and after 6 years they are eligible for free agency (can negotiate a contract with any team). The last two variables (`Y35` and `YG6`) are dummy variables for these two categories.