BST 140.752
Problem Set 3
Due: November 30, 2017

1 Variable Selection

1. The general formula for the Akaike Information Criterion (AIC) for some model of the data, denoted by the pdf $p(\mathbf{y}|\theta)$, is given by

$$AIC = -2 \log p(\mathbf{y}|\hat{\theta}) + 2d,$$

where $\hat{\theta}$ is the MLE of this model, for the vector of observations, $\mathbf{y}$. Use this result to compute the AIC for a standard linear model with normal error.

2 Ridge Regression

1. Let $\hat{\boldsymbol{\beta}}_{LS}$ be the least squares estimator and $\hat{\boldsymbol{\beta}}_{ridge}$ be the ridge regression estimator. Prove that $\mathrm{var}(\hat{\boldsymbol{\beta}}_{LS}) \geq \mathrm{var}(\hat{\boldsymbol{\beta}}_{ridge})$.
2. Let $\mathbf{H}_\lambda$ be the hat matrix for the ridge regression model. Derive an expression for $\mathbf{H}_\lambda$ and find its trace.

3 Principal Components

1. Formally prove that the population principal components explain the maximum variability subject to being linear combinations of the data and orthogonal to the others.
2. Assume that the columns of $\mathbf{X}$ have been mean centered so that $\mathbf{J}_n'\mathbf{X} = (0 \ldots 0)'$. Suppose further that $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}'$ where $\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = I$ and $\mathbf{D}$ is a diagonal matrix of singular values. Argue that the matrix $\mathbf{U} = \mathbf{X}\mathbf{V}'\mathbf{D}^{-1}$ results in an orthonormal basis for the same space as the column space of $\mathbf{X}$. Thus, the $\hat{\mathbf{y}}$ matrix treating $\mathbf{X}$ as the outcome and treating $\mathbf{U}$ as the outcome are the same and further $\hat{\mathbf{y}} = \sum_{j=1}^{p} \mathbf{u}_i < \mathbf{u}_i, \mathbf{y} >$ where $\mathbf{u}_i$ are the columns of $\mathbf{U}$.

4 Time Series Analysis

1. Compute the autocorrelation function for an MA(1) process.
2. Consider a general AR(p) model. Find the Yule-Walker estimators for $\phi_1, \ldots \phi_p$ and σ^2.

5 Coding and data analysis exercises

1. Write an `R` function that performs ridge regression. It should be able to take a list of λ values as input and plot a ridge trace as an output. Use the `mtcars` dataset in `R` to try out your function. Let `mpg` be the response variable, and `cyl`, `disp`, `hp`, `drat`, and `wt` be the explanatory variables.

2. Write an `R` function that performs principal components regression. Use the `mtcars` dataset in `R` to try out your function. Let `mpg` be the response variable, and `cyl`, `disp`, `hp`, `drat`, and `wt` be the explanatory variables.