BST 140.752
Problem Set 4

1 Residuals

1. Consider a linear model $Y = X\beta + \delta\Delta + \epsilon$ where δ is a vector with a 1 at position i_0 and 0 elsewhere. Argue the following.

   A. The i_0 residual is 0 for this model.

   B. The fitted value for β using all of the data and this model is equivalent to that using only the data with the i_0 observation deleted.

   C. Argue that the standardized Press residuals are a test statistic for $\Delta = 0$.

2. Consider the residuals for the ordinary linear model. Derive their mean and variance.

3. Carefully write up the proof that relates the Press residuals to the ordinary residuals. Derive the mean and variance/covariance of the Press residuals.

4. Prove the Sherman/Morrison/Woodburry theorem.

5. Prove that the hat matrix diagonals are between 0 and 1.

6. Why are the studentized residuals not exactly distributed as t statistics?

2 Inference under incorrectly specified models

For all of this section, let Model 1 be $Y = X_1\beta_1 + \epsilon$ and Model 2 be $Y = X_1\beta_1 + X_2\beta_2 + \tilde{\epsilon}$.

1. Suppose that Model 1 is fit while Model 2 represents the actual truth. Give the bias and variance of β_1. Give the expected value of S^2.

2. Suppose that Model 2 is fit while Model 1 is true. Give the bias and variance of the estimted β. Give the expected value of S^2.

3 Confidence ellipsoids

1. Derive a confidence ellipsoid for a linear contrast of coefficients $K\beta$.

2. Derive a confidence ellipsoid for a set predictions at X values X_{new}

3. Derive a prediction ellipsoid for a set of predictions at X values X_{new}.

4 Asymptotics

1. Rigorously prove the consistency of the linear regression estimates (slope and intercept).
2. Rigorously prove an asymptotic confidence interval for the slope of a regression estimate.

5 Coding and data analysis exercises

1. Consider the sleep data from the previous homework.
   A. Consider the model fit from the previous homework. Write a program to grab the hat diagnals as well as use R's lm to obtain them directly. Look at the influence of various data points.
   B. Consider the model fit from the previous homework. Write a program to grab the residuals and Press residuals. Investigate these residuals in the context of this model.
2. Consider the baseball data from the previous exercise.
   A. Consider the model fit from the previous homework. Write a program to grab the hat diagnals as well as use R's lm to obtain them directly. Look at the influence of various data points.
   B. Consider the model fit from the previous homework. Write a program to grab the residuals and Press residuals. Investigate these residuals in the context of this model.
3. Write a function that takes a Y $(n \times 1)$ an X_1 $(n \times 1)$ and an X_2 $(n \times (p-1))$ and produces the partial regression plot of $e_{Y|X_2}$ by $e_{Y|X_2}$.
4. Write an R function that takes in an arbitrary 2 dimensional K matrix and creates and plots the relevant confidence ellipse.
5. Take the `diamond` dataset and create a function that creates and plots confidence ellipsoids for pairs of predictions.
6. Create a Lasso and Ridge regression penalty plot for the sleep data from the last homework where all parameters except the intercept are penalized.
7. Create a Lasso and Ridge regression penalty plot for the baseball data from the last homework where all parameters except the intercept are penalized.