DSC 241 – Homework 1

Problem 1. Write a function confBand(x, y, conf=0.95) taking in a predictor vector $(x_1, \ldots, x_n)$ and a response vector $y = (y_1, \ldots, y_n)$ and return a plot with the points $(x_1, y_1), \ldots, (x_n, y_n)$, the least squares line, and the confidence band at level conf. Apply your function to hp and mpg from the 04cars dataset.

Problem 2. Let $n = 100$ and draw $x_1, \ldots, x_n \overset{\text{iid}}{\sim} \text{Unif}(0, 1)$, which stay fixed in what follows. Repeat the following experiment $N = 1000$ times.

- Generate $y_i = 1 + x_i + \varepsilon_i$, with ε_i i.i.d. $\mathcal{N}(0, 0.2)$.
- Compute the 99% confidence band and record whether it contains the true line, or not.

Summarize the result of this numerical experiment by returning the proportion of times (out of N) that the confidence band contained the true line.

DSC 241 – Homework 2

Problem 1.

a. Perform some simulations to check that the least squares coefficients are indeed normally distributed under the standard assumptions. To simplify things, assume a situation with only one predictor, x, say uniformly distributed in $[-1, 1]$, and that given x the response, y, is generated as $y \sim \mathcal{N}(1+2x, \sigma^2)$, with $\sigma = 0.5$. Generate a sample of size $n \in \{50, 100, 200, 500\}$ from it, and then use that sample to fit the model by least squares; then repeat this procedure $N = 1000$ times. (There are 4 settings, one for each n.) Each time record the intercept and slope. Then produce a plot or two providing (visual) evidence that the least squares intercept and slope are marginally normal. Produce another plot or two providing evidence that they are in fact jointly normal.

b. Repeat, now with errors that have the t-distribution with $k \in \{2, 5, 10, 20, 50\}$ degrees of freedom. Do that for $n \in \{50, 100, 200, 500\}$. (There are $5 \times 4 = 20$ settings now.) Offer some brief comments.

[To visualize the distribution of the least squares coefficient vector, you can compute a 2D histogram and plot it in 3D; or you can compute a kernel density estimate and either plot it in 3D (as a function of 2 variables) or plot its level lines.]

Problem 2. This problem is meant as a practice for performing diagnostics. Consider the `Boston` dataset in the package `MASS`. Fit a linear model with `medv` as response, omitting all the discrete predictor variables.

a. Check the standard assumptions one at a time, except for independence. Offer some brief comments on what you observe.

b. Check for outliers in predictor. Comment on the most significant one if there is any. (What makes that observation unusual?)

c. Same with outliers in response.

d. Same with influential observations.

e. Finally, check for multicollinearity.

Each time, apply all the methods that were introduced in lecture. (This is for you to get familiar with these various tools. In practice, people often have a workflow (routine) they stick to, and only use their preferred methods.)

DSC 241 – Homework 3

Problem 1. (Is polynomial regression well-conditioned?) Consider the canonical design matrix for fitting a polynomial of degree p based on $x_1, \ldots, x_n$, meaning, $\mathbf{X} = (x_i^j)$ for $i = 1, \ldots, n$ and $j = 0, \ldots, p$. Suppose the x_i's are evenly distributed in $(0, 1)$, for example, $x_i = i/(n+1)$ for $i = 1, \ldots, n$. For each $p \in \{1, \ldots, 20\}$ and each $n \in \{30, 50, 100, 200, 500, 1000\}$, compute the condition number of the design matrix. Produce a useful plot for visualizing the result of these computations. Offer brief comments.

Problem 2. (Piecewise constant fit.)

a. Write a function piecewiseConstant(x, y, L, plot = TRUE) taking in a one dimensional predictor variable x with values in $[0, 1]$ and a response y, and fits a piecewise constant model (by least squares) on 2^L intervals of equal length partitioning the unit interval (L is a nonnegative integer) in the form of a numerical vector of length 2^L, with the option of producing a scatterplot with the fit overlaid.

b. Apply your function to explaining City Mpg as a piecewise constant function of Horsepower in the 04cars dataset. Produce a single scatterplot, with lines corresponding to the fit with $L = 2$ (blue), $L = 3$ (green), and $L = 4$ (red). Add a legend, etc, so it looks 'nice'.

DSC 241 – Homework 4

Problem 1. In this problem, you practice working with predictor variables that are discrete. Consider the Boston dataset in the package MASS. Take as response the median property value.

a. Look at side-by-side boxplots for medv where the groups are defined by chas. Comment on what you observe. In particular, compare the different groups visually. Then fit a model explaining medv as a function of chas. Output an ANOVA table. What is the F-test testing? Is the result consistent with the boxplots?

b. Repeat with rad in place of chas.

c. Produce a nice boxplot display of medv where the groups are defined by chas and rad jointly. Comment on what you observe. Then look at an interaction plot. Then fit a model explaining medv as a function of chas and rad with interactions. Output an ANOVA table. What are the different F-tests testing? Compare with the previous F-test as appropriate. Are the results of these tests consistent with the plots you just looked at?

d. It makes sense that median property value decreases with the percentage of lower status population lstat, and this is indeed what is observed here. Does the rate of decrease depend on whether the area borders the Charles River? Produce a plot that helps answer that question. Then formulate that into a hypothesis testing problem and perform an appropriate test.

Problem 2. Consider the same dataset and turn to the problem of fitting a polynomial model explaining medv as a function of lstat.

a. Fit a polynomial model of degree 3 by least squares.

b. Repeat with each robust method covered in the lecture notes/slides.

c. Produce a scatterplot and overlay all these fits with different colors and a legend.

[HINT: use the function predict.]

DSC 241 – Homework 5

Problem 1. Write a function named bootLS(x, y, conf = 0.95, B = 1000) that fits a simple linear model explaining y in terms of x, and returns a studentized bootstrap confidence interval at the desired level based on the specified number of repeats for each coefficient vector. (If conf = 0.95, then each interval will have nominal level 0.95, so the the confidence is individual and not simultaneous, as is the case with the function confint.)

Problem 2. Perform some simulations to compare the length and confidence level of the studentized bootstrap confidence interval (from Problem 1) and of the student confidence interval (the classical one). Compare them at various sample sizes and in settings involving different distributions, for example, the normal distribution and a skewed distribution like the exponential distribution (centered to have mean 0). In the code, first briefly explain in words what you intend to do, and then do it, and at the end offer some brief comments on the results of your simulation study.

DSC 241 - Homework 6

Problem 1. Implement k-fold cross-validation and sequential model selection for linear regression models.

a. Write a function **cv.lm(x, y, k)** which estimates the prediction error of the linear regression model with **y** as response using k-fold cross-validation

b. Write a function **SequentialSelection(x, y, method)** which computes the forward selection path for linear regression from 'intercept only' to 'full model' and chooses the model on that path using different criteria specified by **method**. The function should support these methods:

- **method** = "AdjR2": Sequentially include the columns of **x** and choose the model that gives the largest adjusted R^2.
- **method** = "AIC": Sequentially include the columns of **x** and choose the model that gives the smallest AIC.
- **method** = "CV5": Sequentially include the columns of **x** and choose the model that gives the smallest 5-fold cross-validation prediction error.

Problem 2. Consider a regression setting where the predictor variable is real valued and the goal is to fit a polynomial model. Specifically, we assume that $x_1, ..., x_n$ are iid uniform in $[0, 2\pi]$ and conditional on these, $y_1, ..., y_n$ are independent, with y_i normal with mean $\sin(3x_i) + x_i$ and variance 1. Take $n = 200$ and set the maximum degree at 20. Perform simulations (at least 100 data instances) to compare the choice of degree by the sequential model selection methods in Problem 1. Produce plots of 3 example data instances and their best model fits according to different methods. Produce plots of the distribution of the polynomial degrees chosen by the different methods over all simulated instances. Offer comments on what you observe.

DSC 241 - Homework 7

Problem 1. Find the dataset **Placekick.csv** in the **Datasets** subfolder. Use this dataset to build a logistic regression model to estimate the probability of success for a placekick. Here is the data dictionary:

- **week**: Week of the season
- **distance**: Distance of the placekick in yards
- **change**: Lead-change (1) vs. non-lead-change (0) placekicks
- **elap30**: Number of minutes remaining before the end of the half
- **PAT**: Type of placekick, where a PAT attempt is a 1 and a field goal attempt is a 0
- **type**: Outdoor (1) vs. dome (0) placekicks
- **field**: Grass (1) vs. artificial turf (0) placekicks
- **wind**: Windy conditions (1) vs. nonwindy conditions (0)
- **good**: Successful (1) vs. failed (0) placekicks; this is our response variable

a. Fit a logistic regression model with **good** as response and **distance** as predictor. Interpret the fitted model coefficients and visualize the model fit.

b. Now consider all predictors. Apply the forward selection algorithm to compute the forward selection path from 'intercept only' to 'full model' and chooses the model on that path that minimizes the AIC.

c. Consider the model selected by the forward selection algorithm. Compute the decision boundary when the decision threshold for the probability of success is 0.5.

Problem 2

a. Write a function **bootGLM(x, y, B=1000)** that resamples observations and returns standard errors for each of the predictor variables (when the others are present in the model) in a logistic model.

b. Consider the model selected by the forward selection algorithm from Problem 1(b). Apply your **bootGLM**, and compare with the standard errors returned by the summary function.