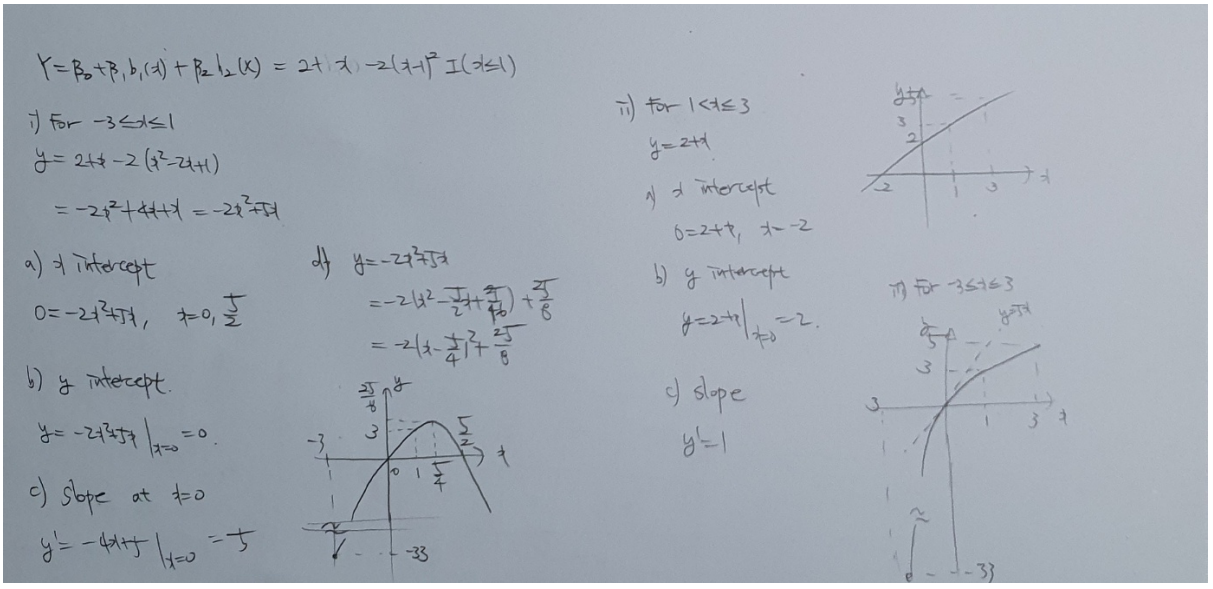
MLDL assignment 4

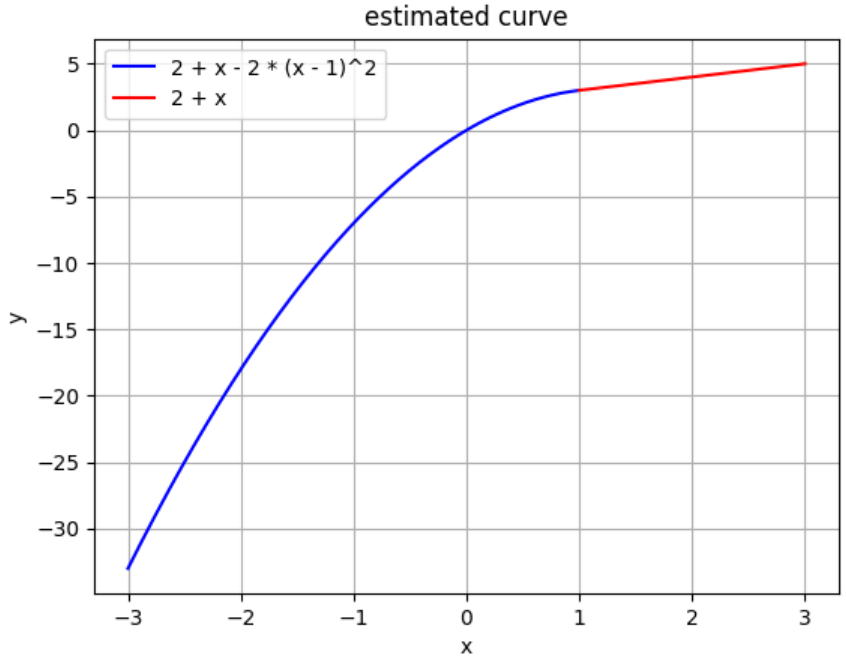
Q1. Basis function [6 pts]

Suppose we fit a curve with basis functions $b_1(X) = X, b_2(X) = (X-1)^2 I(X \leq 1)$. (Note that $I(X \leq 1)$ equals 1 for $X \leq 1$ and 0 otherwise.) We fit the linear regression model $Y = \beta_0 + \beta_1 b_1(X) + \beta_2 b_2(X) + \epsilon$, and obtain coefficient estimates $\hat{\beta}_0 = 2, \hat{\beta}_1 = 1, \hat{\beta}_2 = -2$. Sketch the estimated curve between $X = -3$ and $X = 3$. Note the intercepts, slopes, and other relevant information.

Note the intercepts, slopes, and other relevant information



Sketch the estimated curve



Q2. Analyze splines [6 pts]

Consider two curves $\hat{g}_1$ and $\hat{g}_2$, defined by

$$\hat{g}_1 = \arg\min_g \left(\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int \left[g^{(3)}(x) \right]^2 dx \right),$$

$$\hat{g}_2 = \arg\min_g \left(\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int \left[g^{(4)}(x) \right]^2 dx \right),$$

where $g^{(m)}$ represents the mth derivative of g.

(a) As λ → ∞, will gˆ1 or gˆ2 have the smaller training RSS? [3 pts]

g_2 will have a smaller training RSS than g_1. Because g_2 penalizes a higher derivative than g_1. So, g_2 is more wiggy (more flexible) than g_1.

(b) As λ → ∞, will gˆ1 or gˆ2 have the smaller test RSS? [3 pts]

Cannot compare test RSS. However, g_2 can be more overfitting than g_1. Thus, if g_2 is overfit then g_1 will have a smaller test RSS than g_2.

Q3. Sketch splines [8 pts]

Suppose that a curve $\hat{g}$ is computed to smoothly fit a set of n points using the following formula:

$$\hat{g} = \arg\min_{g} \left(\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int \left[g^{(m)}(x) \right]^2 dx \right),$$

where $g^{(m)}$ represents the mth derivative of g (and $g^{(0)} = g$). Provide example sketches of $\hat{g}$ in each of the following scenarios.

(a) λ=∞,m=1.[4pts]

m = 1 means that the penalty term considers the first derivative (=slope). λ=∞ means that the penalty term dominates. Thus the g is the line with a slope is 0.

(b) λ=∞,m=2.[4pts]

m = 2 means that the penalty term considers the first derivative (=the change of slope). λ=∞ means that the penalty term dominates. Thus the g is the line with a slope that is constant but not zero.

Q4. Drawing a tree and a predictor space [10 pts]

This question relates to the plots below.

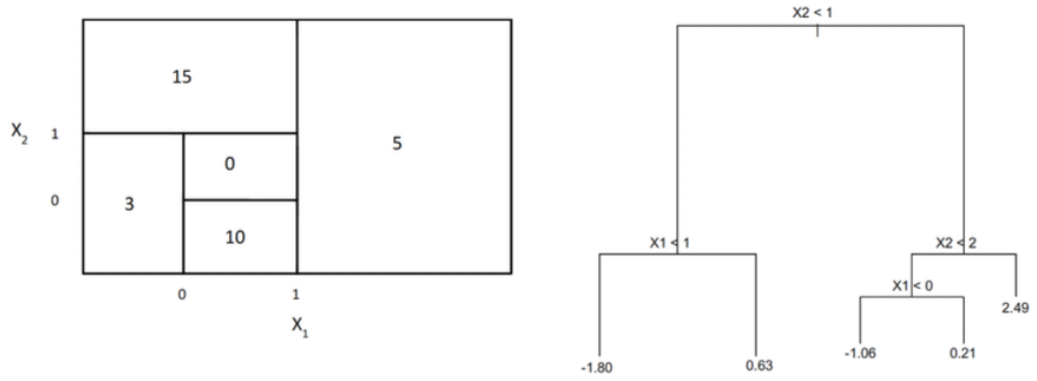
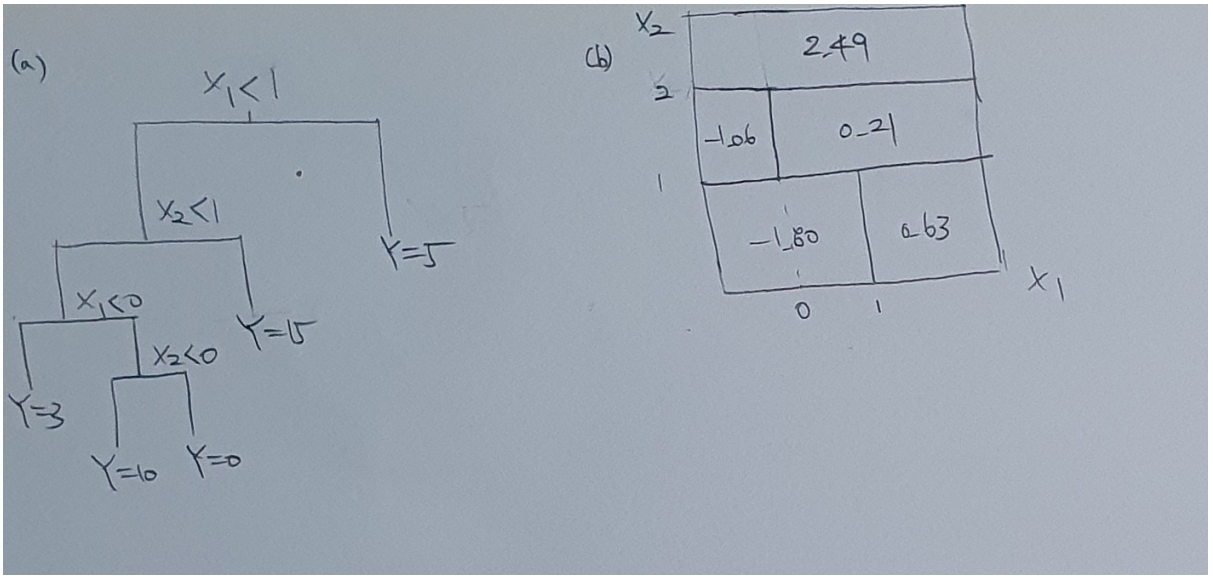


Figure 1: Left: A partition of the predictor space corresponding to problem (a). Right: A tree corresponding to problem (b).

(a) Sketch the tree corresponding to the partition of the predictor space illustrated in the left-hand panel of Figure 1. The numbers inside the boxes indicate the mean of Y within each region. [5 pts]

(b) Create a diagram similar to the left-hand panel of Figure 1, using the tree illustrated in the right-hand panel of the same figure. You should divide up the

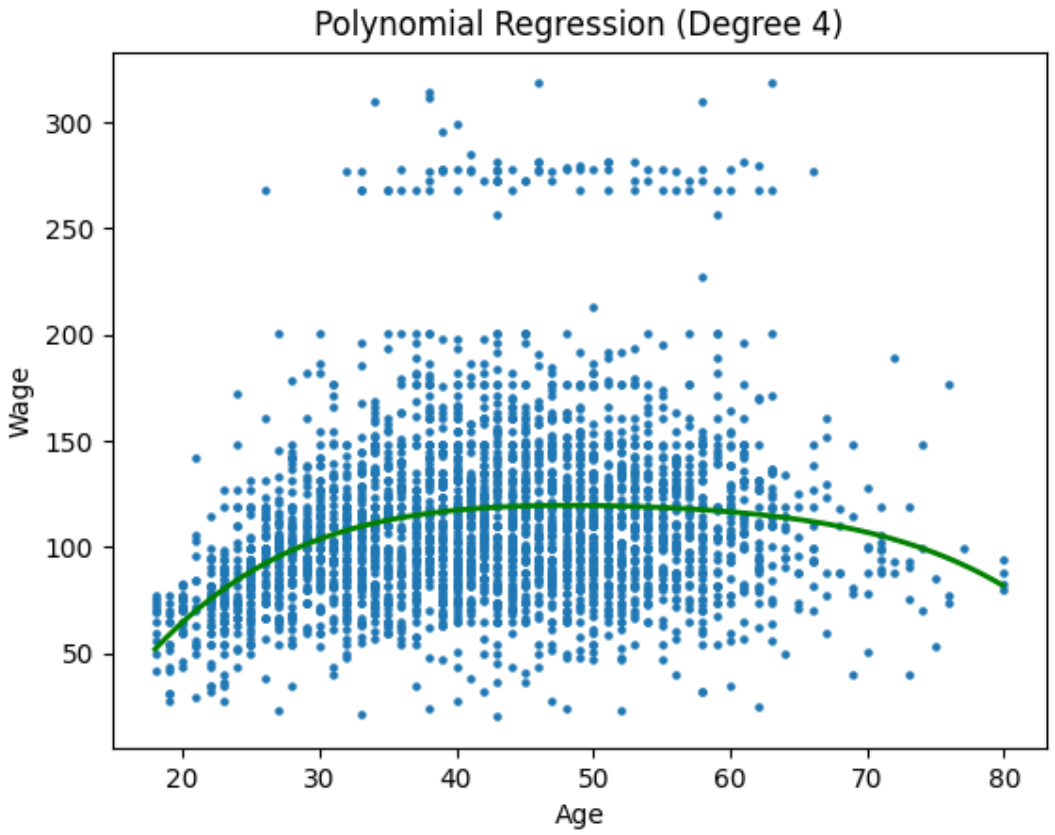
predictor space into the correct regions and indicate the mean for each region. [5 pts]



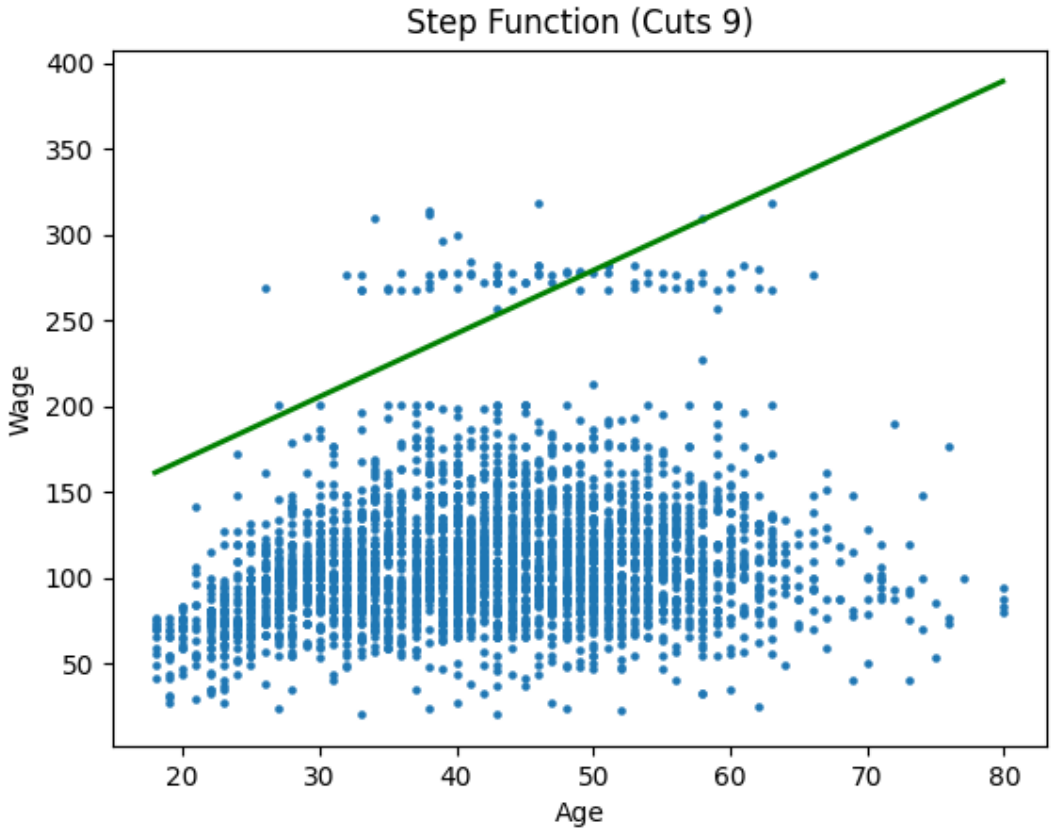
Q5. Finding the best fit - Polynomial and Step function [10 pts]

In this exercise, you will further analyze the Wage data set considered throughout this chapter.

(a) Perform polynomial regression to predict wage using age. Use cross-validation to select the optimal degree d for the polynomial. What degree was chosen, and how does this compare to the results of hypothesis testing using ANOVA? Make a plot of the resulting polynomial fit to the data. [6 pts]



(b) Fit a step function to predict wage using age, and perform cross-validation to choose the optimal number of cuts. Make a plot of the fit obtained. [4 pts]



Q6. Finding the best fit - Regression Spline [7 pts]

This question uses the variables dis (the weighted mean of distances to five Boston employment centers) and nox (nitrogen oxides concentration in parts per 10 million) from the Boston data. We will treat dis as the predictor and nox as the response.

(a) Use the bs() function from the ISLP.models to fit a regression spline to predict nox using dis. Report the resulting RSS by changing a range of degrees of freedom. Describe the results obtained. [4 pts]

Degree of Freedom: 3, RSS: 1.9341067071790703
Degree of Freedom: 4, RSS: 1.922774992811925
Degree of Freedom: 5, RSS: 1.8401728014885235
Degree of Freedom: 6, RSS: 1.8339659031602091
Degree of Freedom: 7, RSS: 1.829884445923284
Degree of Freedom: 8, RSS: 1.8169950567252335

Degree of Freedom: 9, RSS: 1.8256525103870564
Degree of Freedom: 10, RSS: 1.7925348895561337

(b) Perform cross-validation or another approach in order to select the best degrees of freedom for a regression spline on this data, and explain your results. [3 pts]

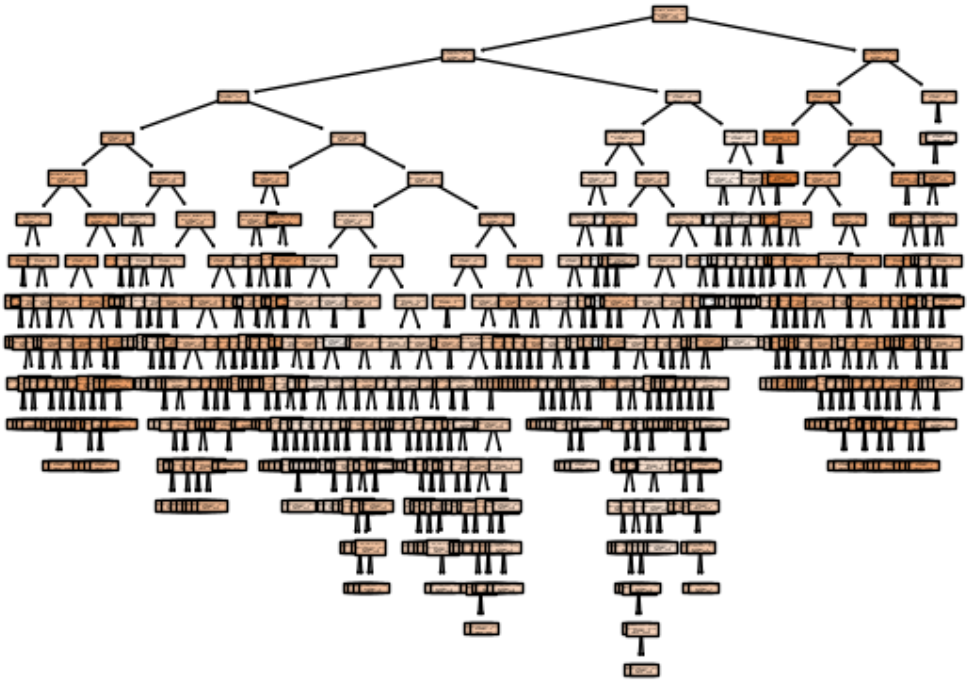
Best degree of freedom: 3

Q7. Regression trees [13 pts]

In the lab, a classification tree was applied to the Carseats data set after converting Sales into a qualitative response variable. Now we will seek to predict Sales using regression trees and related approaches, treating the response as a quantitative variable.

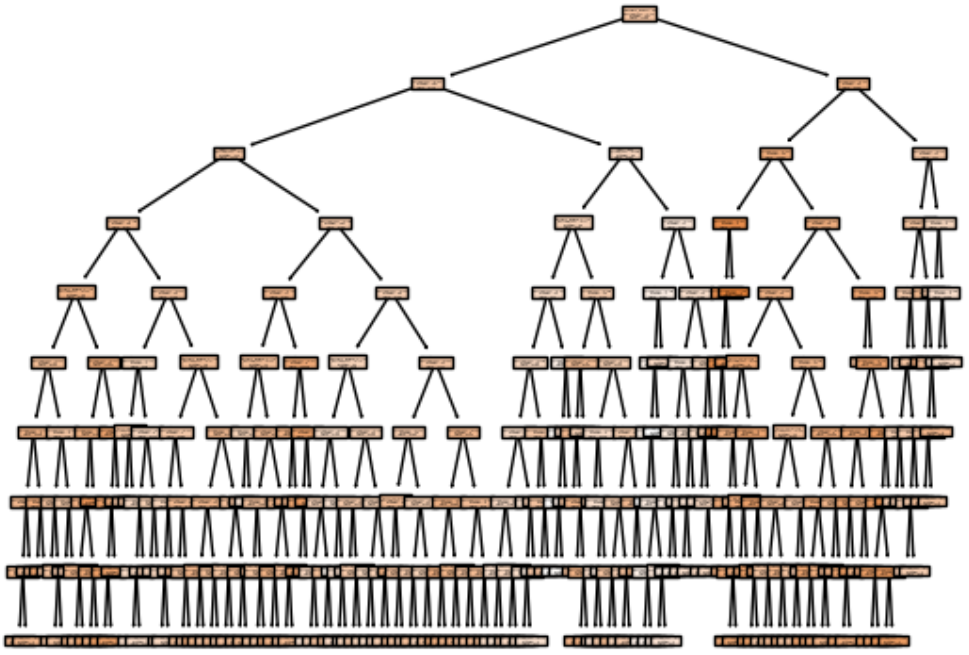
(a) Split the data into a training and a test set, and fit a regression tree to the training set. Plot the tree, and interpret the results. What test MSE do you obtain? [3 pts]

Regression Tree Test MSE: 5.657312499999999



(b) Use cross-validation in order to determine the optimal level of tree complexity. Does pruning the tree improve the test MSE? [3 pts]

Optimal Max Depth: 9
Optimal Regression Tree Test MSE: 4.932234696642867



(c) Use the bagging approach in order to analyze this data. What test MSE do you obtain? Use the feature importance values to determine which variables are most important. [3 pts]

Bagging Test MSE: 3.2506822564999993

(d) Use random forests to analyze this data. What test MSE do you obtain? Use the feature importance function to determine which variables are most important. Describe the effect of m, the number of variables considered at each split, on the error rate obtained. [4 pts]

Random Forest Test MSE: 4.304821365499997

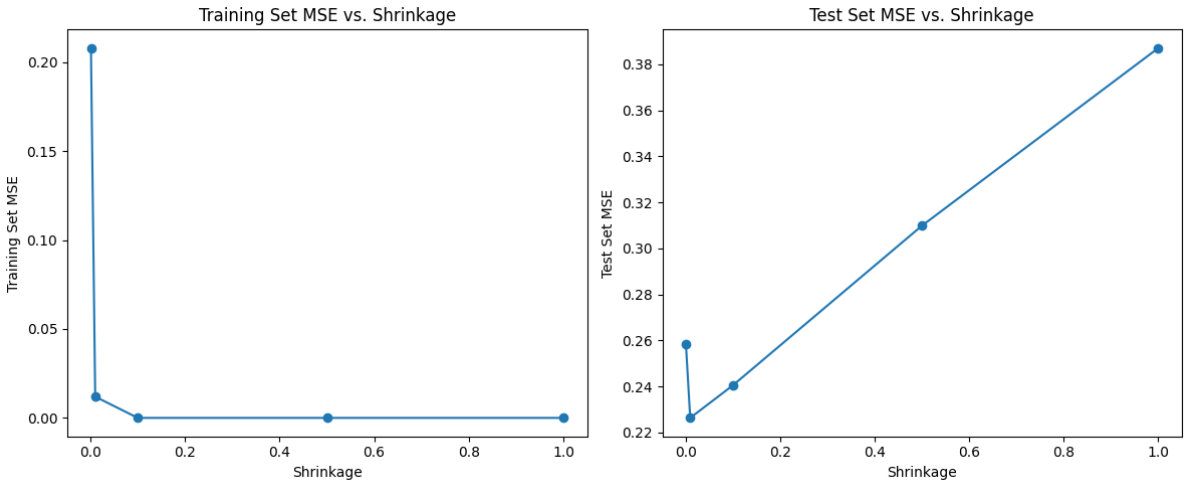
Feature Importance

Price: 0.2320963582000836
ShelveLoc_Good: 0.1414110620601396
Age: 0.11562068937535608
Advertising: 0.09316784236140743
CompPrice: 0.08551435765930225
Unnamed: 0: 0.07744894323302373
Income: 0.07172431714306747
Population: 0.06828032739371646
Education: 0.04865940452813053
ShelveLoc_Medium: 0.04000468120492429
US_Yes: 0.015493859785432945
Urban_Yes: 0.01057815705541554

Q8. Boosted trees and its competitors [10 pts]

We now use boosting to predict Salary in the Hitters data set.

(a) Remove the observations for whom the salary information is unknown, and then log-transform the salaries. Next, create a training set consisting of the first 200 observations, and a test set consisting of the remaining observations. Perform boosting on the training set with 1,000 trees for a range of values of the shrinkage parameter λ. Produce two plots: one with different shrinkage values on the x-axis and the corresponding training set MSE on the y-axis, another with different shrinkage values on the x-axis and the corresponding test set MSE on the y-axis. [5 pts]



(b) Compare the test MSE of boosting to the test MSE that results from applying two of the regression approaches seen in Chapter 3 and 6 of the book. [3 pts]

Test MSE - Boosting: 0.24038261111334228
Test MSE - Linear Regression: 0.5156972276434887
Test MSE - Ridge Regression: 0.5156596724392274

(c) Which variables appear to be the most important predictors in the boosted model? [2 pts]

Feature Importance:
CAtBat: 0.5479178925985807
CHits: 0.07114305977135572
AtBat: 0.05472360274785938
Walks: 0.0524301721353389
CRuns: 0.04656409437099669
CHmRun: 0.036539961410281506
CRBI: 0.03554142629835031
Hits: 0.03237440051351287
Years: 0.025188943000885488
CWalks: 0.024069217171252426
RBI: 0.022467890435856668
PutOuts: 0.01567984994347088
Runs: 0.01443713463933932
Errors: 0.008177707810942545

HmRun: 0.007224412582952788
Assists: 0.005520234569023919

Q9