**Problem 1: Concept Review**

1. Lasso can completely ignore some factors, setting those regression coefficients to 0. This is because lasso considers not only the factor’s importance but also its simplicity which is through the l1 penalty. When lasso finds the best-fitting model, it balances the importance of each regression coefficient with the penalty for using that variable. If a predictor doesn’t contribute much to improving the model’s accuracy, the penalty may be stronger than its importance, and lasso will set that coefficient to 0, removing it from the model. On the other hand, ridge regression can’t entirely remove factors, it can reduce the value of the regression coefficient.
2. *Part a*: (iii).

As λ increases, the penalty term becomes more influential, and it encourages the regression coefficients to become smaller. This causes the model to fit the training data less closely, leading to an increase in the training MSE as it becomes less accurate in capturing the training data.

*Part b*: (ii).

As λ increases, initially, the test MSE decreases because the increased penalty on the coefficients reduces overfitting, leading to better generalization and lower test error. But as λ continues to increase, the model becomes constrained, causing it to underfit the data and the test MSE starts to increase.

*Part c*: (iv).

As we increase λ, we are restricting the Bj coefficients more and more, the coefficients will deviate from their least squares estimates. So the model will become less and less flexible which causes the steady decrease in variance.

*Part d*: (iii).

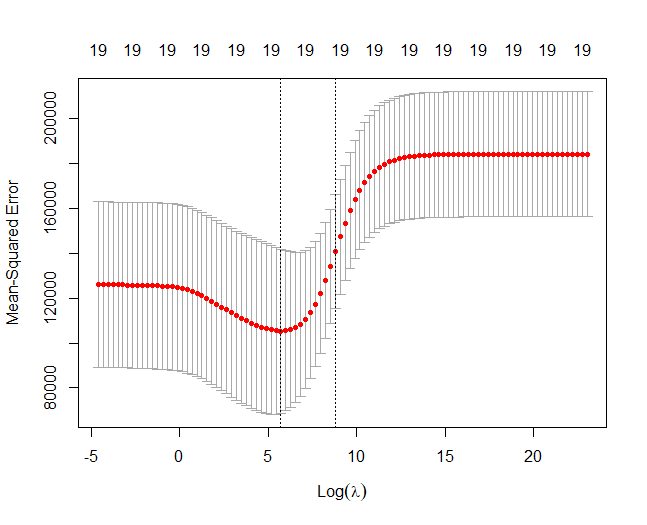
As we increase λ, we are restricting the Bj coefficients more and more, the coefficients will deviate from their least squares estimates. So the model becomes less and less flexible which provokes a steady increase in bias.

*Part e*: (v).

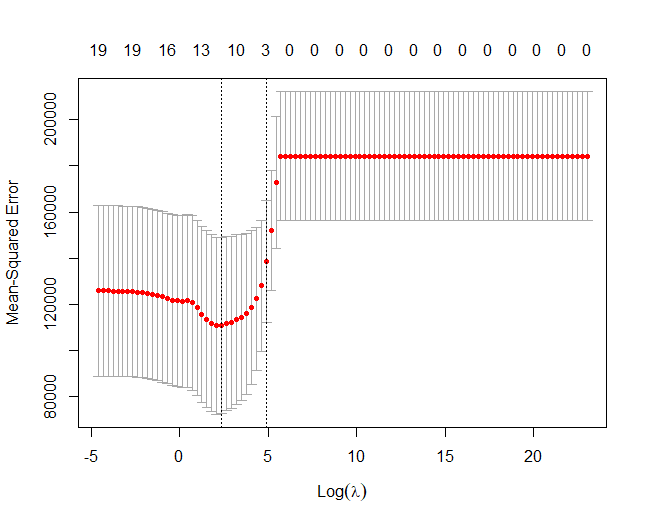
By definition, the irreducible error is independent of the model, and thus independent of the value of λ

**Problem 2: Regularized Regression Models**

1. See R script
2. λridgemin = 305.3856



1. λridgelse = 6579.332
2. λlassomin = 10.72267



1. λlassolse = 132.1941
2. λridgemin test MSE = 140081.9

λridgelse test MSE = 165523.3

λlassomin test MSE = 143527.7

λlassolse test MSE = 157773.7

1. Ridge regression models have all the predictors while lasso can zero them out when needed. Using lse for ridge regression made the coefficients smaller but the intercept larger. Using lse for lasso did the same and also got rid of some of the predictors compared to min for lasso.

λridgemin λridgelse

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

λlassomin λlassolse

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

1. αenet = 0.01

λenet = 305.3856

1. elastic net test MSE = 140023.6
2. Of the models ridge, lasso and elastic net, the elastic net model performs the best in terms of prediction. It has the lowest test MSE and because it has two tuning parameters. This dual regularization allows elastic net to have a balance between feature selection and coefficient shrinkage, making it the best model
3. Features to focus on: Hits, Walks, CHmRun, CRuns, CRBI, LeagueN, DivisionW, PutOuts and Errors

**Problem 3: Bootstrap**

1. µhat = 22.53281
2. standard error of µhat = 0.4088611
3. standard error of µhat using bootstrap = 0.4041568

The standard error using bootstrap vs the analytical formula is very similar.

1. 95% confidence interval for mean of medv is (21.74444, 23.32906)

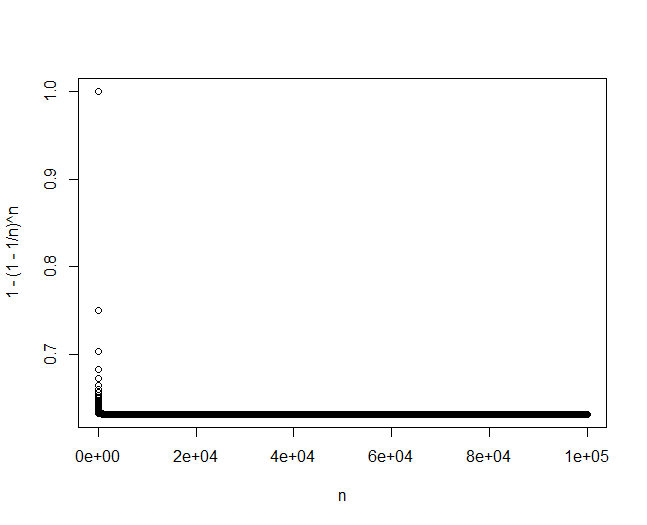
Using analytical formulas, the 95% confidence interval for mean of medv is (21.72953, 23.33608)

The two are very similar.

1. µhatmed = 21.2
2. standard error of µhatmed using bootstrap= 0.3765876
3. µhat0.1 = 12.75
4. standard error of µhat0.1 = 0.492203. This shows how far the estimate of the 10th percentile of medv is from the true value.

**Problem 4: Properties of Bootstrap**

1. The probability that the first bootstrap observation is the j-th observation from the original sample is 1/n
2. The probability that the first bootstrap observation is not the j-th observation from the original sample is 1-1/n.
3. The probability that the j-th observation from the original sample is not in the bootstrap sample is (1-1/n)n
4. 1-(1-1/5)5 = 0.67232
5. 1-(1-1/100)100 = 0.63397
6. 1-(1-1/10000)10000 = 0.632139
7. The graph quickly reaches an asymptote of approximately 0.632



1. The probability that a bootstrap sample of size n contains the jth observation converges

lim n 🡪 ∞ 1-(1-1/n)n = 1 – 1/e ≈ 0.632