HW1

**Problem 1: Bias-variance decomposition**

1. A graph on a piece of paper

   Description automatically generated
2. Expected Test MSE is the true test MSE of an infinite number of training sets. Training MSE is the mean (squared) difference between what we observed and the fitted model for the dataset the model was trained on. Bias is the error introduced by approximating f; it is the difference between what is predicted, Y hat = f hat x, and the true value of the E(Y)=f(x). Variance is the amount that the predictions, Y hat, would change if it was estimated using a different training set. Irreducible error is the error/noise that is observed in the data set and can’t be reduced.
3. As the flexibility of the method increases, bias gets smaller and there is a closer fit to the data. However, at some point, we will start to overfit which means we are modeling noise instead of the relationship between Y and X. So, the model will be a poor generalization to the new data and the variance will increase. This results in the u-shape curve of the expected test MSE. The training MSE will always decrease because the model is getting a closer fit to the data. The irreducible error will be constant. Even if we choose the optimal model, there will still be irreducible error we can’t reduce even with changing the flexibility.
4. An advantage of a flexible approach in supervised learning is its capability to capture complex and non-linear relationships in the data, leading to reduced bias and better model fit. A drawback is the risk of overfitting where the model might exhibit high variance on new data. So, interpretability of results can be more challenging. A more flexible approach is preferred when dealing with intricate data relationships and prioritizing accurate predictions for new data points. A less flexible approach is preferred when we are interested in the interpretability of the results.
5. Because the true relationship is linear, the linear regression model would be the one that accurately represents the underlying data generating process. The cubic regression model brings complexity due to the cubic terms which is not necessary for capturing the true linear relationship. The cubic regression model has more parameters to estimate and provides it the flexibility to fit more intricate patterns in the data, even if those patterns do not align with the true linear relationship. Because the data is generated from a linear relationship, the linear regression model would be a better fit and have a lower training MSE. The cubic regression might overfit the data to some degree, meaning there would be a higher training MSE.
6. When evaluating the models on test data, it could be expected that the linear regression model to have a lower test MSE than the cubic regression model. The linear model is better suited for getting the true linear relationship and is less likely to suffer from overfitting when tested on unseen data points.

**Problem 2: Interpreting MLR**

1. The given regression model is Salary = βˆ0 + βˆ1 \* GPA + βˆ2 \* IQ + βˆ3 \* Level where βˆ0 is the intercept, βˆ1 is the coefficient for GPA, βˆ2 is the coefficient for IQ, βˆ3 is the coefficient for Level. The option (ii.) for a fixed values of IQ and GPA, college graduates earn more, on average, than high school graduates. βˆ0 is the starting salary for high school graduates with GPA and IQ values of 0. But interpreting the intercept in this way isn’t practical because GPA and IQ are unlikely to be exactly 0 in this context. βˆ1 and βˆ2 represent the change in salary for a one-unit increase in GPA and IQ, respectively, while holding other variables constant. βˆ3 represents the difference in starting salary between college grads and high school grads while keeping GPA and IQ constant. Based on all this, the correct interpretation is that for a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates.
2. Salary = βˆ0 + βˆ1 \* GPA + βˆ2 \* IQ + βˆ3 \* Level

Given: βˆ0 = 50, βˆ1 = 20, βˆ2 = 0.07, βˆ3 = 35, Level for college graduate = 1, IQ = 110, GPA = 4.0

Salary = 50 + 20\*4.0 + 0.07\*110 + 35\*1 = 50 + 80 + 7.7 + 35 = 172.7

Predicted salary for college graduate with IQ of 110 and GPA of 4.0 is approximately $172,700.

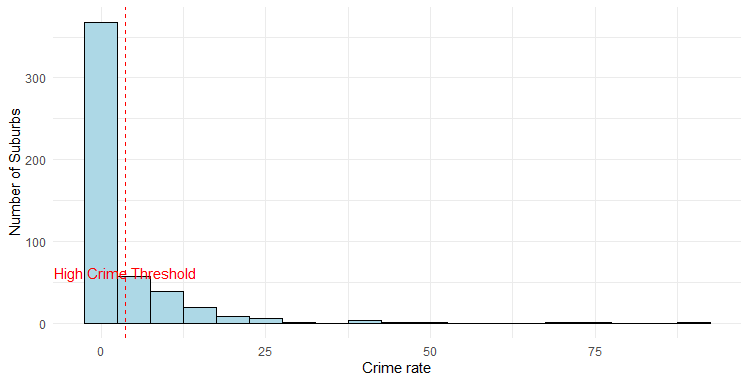
1. False. The size of the coefficient doesn’t determine the importance of a predictor’s effect on the response variable. Context, scaling, and interpretation are important factors in assessing the actual impact of a predictor on the response variable.

**Problem 3: Multiple linear regression**

1. There are 506 rows and 13 variables in the data set Boston. The variable lstat represents the lower status of the population (percent).
2. Average per capita crime rate across all suburbs in the data set is 3.613524
3. The average crime rate for suburbs near Charles River is 1.85167 and the average crime rate for suburbs away from Charles River is 3.744447. It is safer near the Charles River.
4. For this problem, I defined the high crime rates as those above the 75th percentile. Crime rate has the following summary statistics:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min. | 1st Quantile | Median | Mean | 3rd Quantile | Max |
| 0.00632 | 0.08204 | 0.25651 | 3.61352 | 3.67708 | 88.97620 |

Here is a plot of the number of suburbs with crime rates.



1. To check if there are any other predictors in the data set that are associated with per capita crime rate, I made a correlation matrix:

|  |  |
| --- | --- |
| Predictor | Correlation with crime |
| crim | 1.00000000 |
| rad | 0.62550515 |
| tax | 0.58276431 |
| lstat | 0.45562148 |
| nox | 0.42097171 |
| indus | 0.40658341 |
| age | 0.35273425 |
| ptratio | 0.28994558 |
| chas | -0.05589158 |
| zn | -0.20046922 |
| rm | -0.21924670 |
| dis | -0.37967009 |
| medv | -0.38830461 |

Higher crime rates are correlated with areas having greater accessibility to radical highways (rad) and higher property tax rates (tax). Lower crime rates are correlated with areas closer to employment centers (dis) and higher median home values (medv).

1. B^0 = -3.33054, standard error =0.69376, p-value = 2.09e-06

B^1 = 0.54880, standard error = 0.04776, p-value = <2e-16

The crime rate is calculated by = (-3.3305) + 0.5488(lstat)

1. A linear regression model was made to predict crime rates based on the percentage of lower status population in the Boston suburbs. The lm() function is used to build the model . model$coefficients provides the estimated coefficients in the model , showing the intercept and the effect of the predictor lstat on crime rates. This approach is reasonable since it tries to find a linear relationship between these variables. But the model’s validity relies on meeting assumptions like linearity and normality of the residuals.
2. After fitting a simple linear regression for each predictor in the dataset, these are the results:

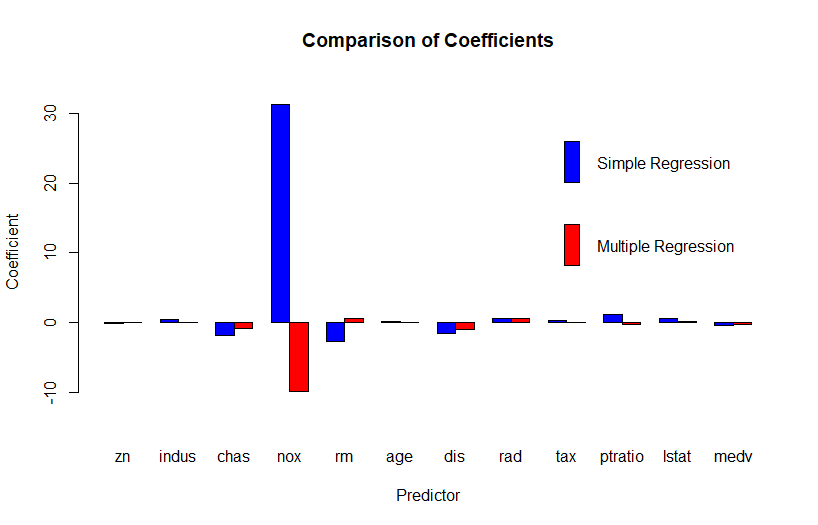
|  |  |  |  |
| --- | --- | --- | --- |
| Variable |  |  | p-value for predictor |
| zn | 4.45 | -0.07 | 5.51e-06­ |
| indus | -2.06 | 0.50 | < 2e-16 |
| chas | 3.74 | -1.89 | 0.209 |
| nox | -13.72 | 31.24 | < 2e-16 |
| rm | 20.48 | -2.68 | 6.34e-07 |
| age | -3.77 | 0.10 | 2.85e-16 |
| dis | 9.49 | -1.55 | < 2e-16 |
| rad | -2.28 | 0.61 | < 2e-16 |
| tax | -8.52 | 0.29 | < 2e-16 |
| ptratio | -17.64 | 1.15 | 2.94e-11 |
| lstat | -3.33 | 0.54 | < 2e-16 |
| medv | 11.79 | -0.36 | < 2e-16 |

All predictors have a p-value less than 0.05 except predictor chas. So, we could conclude that there is a statistically significant association between each predictor and crim except for the chas predictor in this case. Plots of these linear regression models are created in R code.

1. The coefficients represent the estimated effect of each predictor on the response variable. The intercept is 13.77. zn, dis, rad, and medv have statistically significant p-values (p <0.05), indicating a significant impact on the response variable. This model provides insights into the relationships between the predictors and the response variable, where some predictors show stronger associations than others.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Predictor | Coefficient | Standard Error | p-value | Significance |
| Intercept | 13.77 | 7.08 | 0.052 |  |
| zn | 0.045 | 0.018 | 0.015 | \* |
| indus | -0.058 | 0.083 | 0.485 |  |
| chas | -0.82 | 1.18 | 0.485 |  |
| nox | -9.95 | 5.28 | 0.060 |  |
| rm | 0.62 | 0.607 | 0.300 |  |
| age | -0.0008 | 0.017 | 0.962 |  |
| dis | -1.01 | 0.28 | 0.000372 | \* |
| rad | 0.61 | 0.08 | 0.00000000000858 | \* |
| tax | -0.003 | 0.005 | 0.465 |  |
| ptratio | -0.30 | 0.18 | 0.103 |  |
| lstat | 0.13 | 0.075 | 0.067 |  |
| medv | -0.22 | 0.05 | 0.00026 | \* |

1. Simple linear regression models only consider one predictor at a time while ignoring the presence of the other predictors. Multiple regression models consider all predictors simultaneously, adjusting for the influences of other predictors. The comparison plot provides evidence that using many simple linear regression models might not be sufficient compared to a MLR model, especially when dealing with multiple predictors that may interact or have interdependencies.



1. Training MSE = 42.49345

Test MSE = 41.19923

1. Training MSE = 43.97466

Test MSE = 39.62763

These results compare to the results from part k as using a subset of predictors yields slightly higher training MSE but a slightly lower test MSE compared to using all predictors in the model. This suggests that the model with the subset of predictors is performing slightly better on the test set, as indicated by the lower test MSE. But the differences in MSE values between the two approaches are relatively small. The choice of predictors can have varying impacts on the model’s performance and small changes in MSE might not necessarily indicate significantly better or worse models.

1. I think that this is not that surprising and that it is expected to see when using a subset of predictors in a linear regression model. When using a subset of predictors, we are basically simplifying the model by only considering a smaller set of features to predict the response variable.