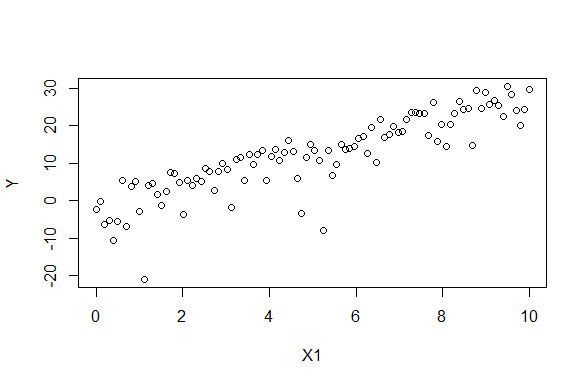
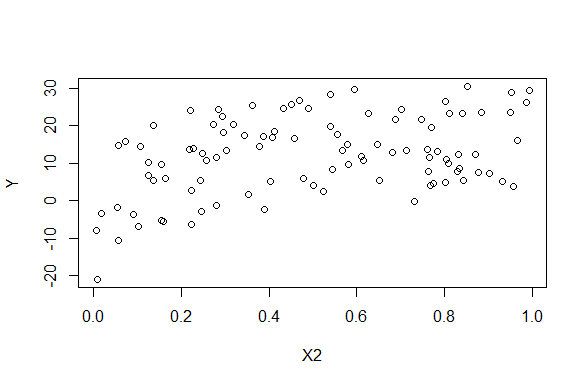
**Problem 1: Properties of Least Square Estimators via Simulations**

1. Β0 = 2, β1 = 3, β2 = 5
2. 100 observations of Yi from the true population regression model were generated and can be seen in the R script.
3. Scatter plot X1 and Y

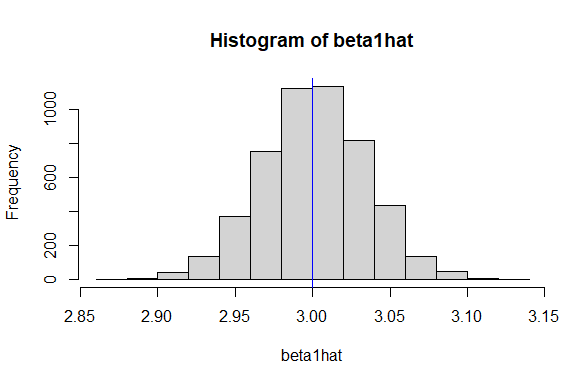


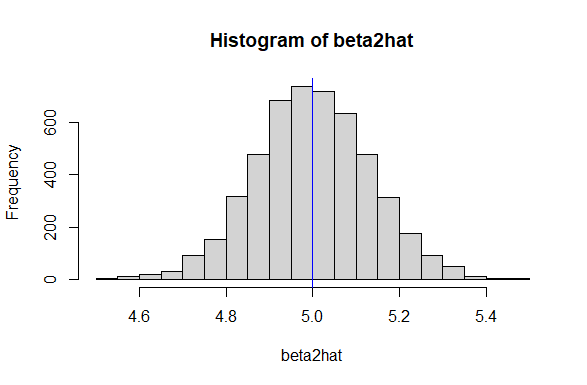
Scatter plot of X2 and Y



The scatterplot of X1 compared to Y shows a strong positive linear correlation. The scatterplot of X2 compared to Y shows a weak positive linear correlation.

1. Simulation was designed in the R script. The mean of beta1hat was 3.001126 which means that βˆ1 is an unbiased estimator of β1
2. Histogram of the sampling distribution of the βˆ1’s generated



1. Simulation was designed in the R script. The mean of beta2hat was 4.999846 which means that βˆ2 is an unbiased estimator of β2
2. Histogram of the sampling distribution of the βˆ2’s generated
3. A paper with writing on it

   Description automatically generated
4. E(σ hat^2) = n / (n-2) \* var(B^1)

n = 100

mean(beta1hat) from code = 3.001126

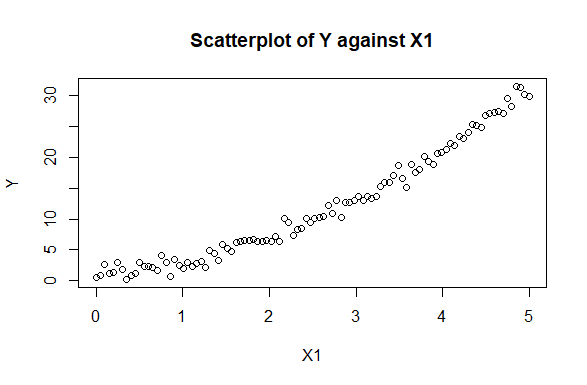
E(σ hat^2) = 100 / (100-2) \* (3.001126) = 100 / 98 \* 3.001126 ≈ 3.0666. So the unbiased estimator for Var(ei) for the simulated data is approximately 3.0666.

**Problem 2: Review of regression concepts**

1. False. The student used E(Yi) when they should’ve used Yi. Instead of using a particular Yi, they used the true mean of Yi in their model.
2. False. The training MSE being smaller than the test MSE is not always guaranteed, although in general that is what we expect. The relationship between training and test MSE depends on the model’s performance and its ability to generalize to new data.
3. False. The expected test MSE is the expected value of the squared difference between the true test set values and the predicted values evaluated on the test set. It is not computed using training set values.
4. True. By reducing complexity, the variance can decrease at the cost of potentially increasing bias. If this reduction in variance outweighs the increase in bias, it can lead to an overall improvement in expected test MSE.
5. False. The expected test MSE can’t be smaller than the irreducible error. Expected test MSE = Irreducible error + variance + bias­2. Since the irreducible error can’t be eliminated by any model, the expected test MSE will always be greater than or equal to the irreducible error.
6. True. The training MSE can be smaller than the irreducible error. If the model is a well fit, then it’s possible for the training MSE to be smaller than the irreducible error.
7. The proposed linear regression model may be problematic because of multicollinearity, which is when two or more predictors in the model are highly correlated, making it difficult for the model to differentiate the individual effects of each predictor on the response variable. The new predictor ‘unhappiness’ is defined as the patient’s average of their disease severity score and anxiety score. Because of this, it’s highly likely to be strongly correlated with both ‘severe’ and ‘anxiety’. This high correlation between ‘unhappiness’ and the other predictors can lead to multicollinearity in the model. In linear regression, a matrix is full rank when its columns (corresponding to predictors) are linearly independent. When multicollinearity is present, the matrix of predictors may not be full rank.
8. Behavior of the RSS depends on whether the new predictor contributes valuable information, leads to multicollinearity, or doesn’t make a significant difference. Adding a new predictor that adds valuable information to the model, allowing it to better fit the data, the model’s predictors become closer to the actual observed values, leading to a decrease in the RSS. Adding a new predictor doesn’t provide valuable information beyond the existing predictors, it won’t lead to a reduction in the squared differences between observed and predicted values, resulting in an unchanged RSS. In cases of multicollinearity, the model may fit the training data worse than a simpler model without the new predictor, causing an increase in the RSS.

**Problem 3: Expected test MSE**

1. 100 observations for Yi are generated in the R script.



1. Predicted values for Model M1:

3.000537, 3.118931, 3.069246, 3.137434, 2.709149

Predicted values for Model M2:

3.126383, 3.242894, 3.19413, 3.25665, 2.822571

Predicted values for Model M3:

3.134523, 3.234978, 3.107329, 3.379868, 2.906938

Predicted values for Model M4:

3.223927, 3.244904, 3.324007, 3.309466, 3.006489

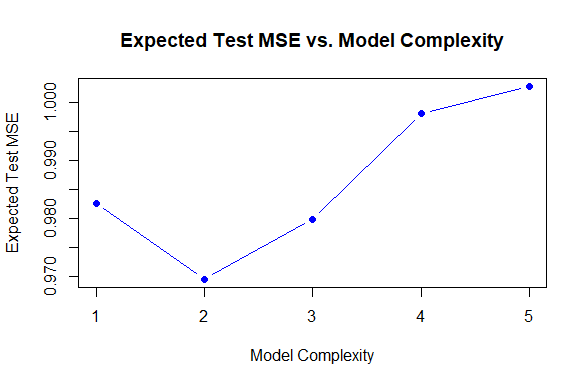
Predicted values for Model M5:

3.258938, 3.24573, 3.236349, 3.29144, 3.083052

1. The first five values in the test set are: 2.132794, 2.415852, 3.149316, 2.029063, 4.121106
2. The five expected test MSEs are: 0.9826527, 0.9694608, 0.9798791, 0.9981143, 1.0028426

Model 2 has the smallest expected test MSE of 0.9694608.

1. Plot of expected test MSE vs model complexity:



1. Model complexity increases 1 to 5 where Model 1 is the simplest (linear regression) and Model 5 is the most complex (polynomial regression with a 5th degree). Model 1 is the simplest model and has high bias. It assumes a linear relationship between the predictors and the response and may not capture the underlying non-linear patterns in the data. But it has low variance as it’s not very sensitive to small fluctuations in the training data. Model 2 introduces a 2nd order polynomial term, allowing it to capture some non-linear patterns. It has a better balance between bias and variance compared to Model 1, resulting in a lower expected test MSE. Models 3 to 5 have higher degree polynomial models and the complexity increases. While they can fit the training data very closely and reduce bias, they also become more prone to overfitting and have higher variance.