Midterm 1

**Question 1: Consulting**

1. The estimate for B2 is 327.58649.

we know that t = Bhat / SE so t\*SE = Bhat

= 6.377 \* 51.37 = 327.58649

1. H0: B2 = 0 versus H1: B2 ≠ 0

Test statistic = 6.377

Null distribution: t distribution with 4997 degrees of freedom

P-value: 2.49e-10

Conclusion: We have evidence at the α = 0.01 significance level that the regression coefficient associated with X2 is significantly different from 0

1. We can use cross validation to check the model’s performance on multiple subsets of the data. This essentially involves splitting the data into training and testing sets multiple times and evaluating the model’s performance across these different splits. CV helps get a more robust estimate of the model’s performance, considering the potential variations due to different training data.
2. While p-values can be useful for assessing the significance of individual predictors, subset selection methods have an ability to account for interaction effects, handle collinearity among predictors, have model simplicity, and reduce overfitting, making it a valuable tool for selecting the most relevant predictors in a regression model.
3. Part a: Redundant predictors introduce multicollinearity, making it challenging for the model to distinguish the individual effects of correlated predictors.

Part b: violations of the assumption of normality occur. This can affect the validity of statistical inference and reliability of the estimates.

Part c: this can indicate the presence of omitted variables. This can lead to biased estimates and unreliable statistical inferences.

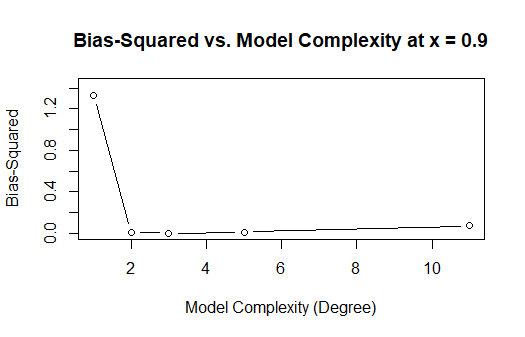
**Problem 2:**

1. The columns are the models M1, M2, M3, M4, M5. The rows are the 5 predicted values

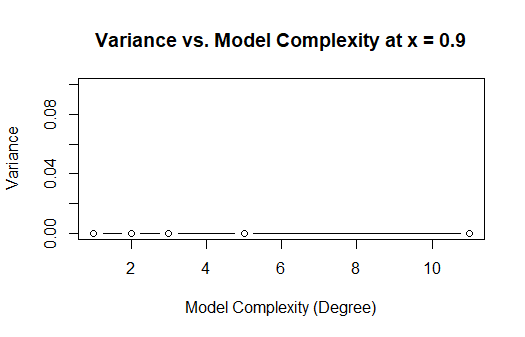
A screenshot of a computer screen

Description automatically generated

1. Here are the square bias results



1. Here are the variance results



1. The irreducible error is approximately 1 because the error is generated from a normal distribution with mean 0 and standard deviation 1.
2. Expected Test MSE for the 5 models:

M1: 2.117303

M2: 1.022738

M3: 1.002052

M4: 1.017395

M5: 1.016199

1. Bias(fhat(x0)): we were calculating the squared bias for different models at a specific point (x=0.9) to assess how well each model captures the underlying population pattern. Var(fhat(x0)) refers to the variance of the estimated or predicted values produced by the model at a specific point x. As model complexity increases, bias squared initially decreases and then increases and it could be because of overfitting. Variance slightly increases as complex models are flexible and can fit data closely.
2. Irreducible error estimates for the 5 models:

2.3527570, 0.9098303, 0.8656705, 0.8537671, 0.8532190