Regression Analysis
Final Exam

Likhit Garimella
lg836

Q.1)

1.01 False. A rejection of the null hypothesis H0: β1=0 in a simple linear model indicates that the predictor variable X is statistically significant in predicting the response variable Y, but it does not necessarily imply that the simple linear model is more useful than the model Yi = β0 + εi. The usefulness of a model depends on various factors beyond the significance of individual predictors.

1.02 True.

1.03 False. In logistic regression, an estimated probability of an outcome does not directly translate to a prediction of whether the event occurred or did not occur. To make such predictions, a threshold is typically set, and if the estimated probability exceeds the threshold, it is predicted that the event occurred; otherwise, it is predicted that the event did not occur.

1.04 False. The area under the Receiver Operating Characteristic (ROC) curve is commonly used for evaluating and comparing classification models, particularly in logistic regression or other models where the outcome is binary or categorical. It is not directly applicable for comparing different linear regression models, which are typically used for predicting continuous outcomes.

1.05 False. The correlation between X1 and Y being positive does not guarantee a positive coefficient for X1 in the model with Y, X1, and X2, even if the X'X matrix is diagonal. The presence of other predictors (X2 in this case) can affect the coefficient of X1 due to multicollinearity or other model interactions.

1.06 False. R2 = 0 implies that the variation in the response variable Y is not explained by the linear relationship with the predictor variable X. However, it does not necessarily imply that Y and X are not linearly related. It could indicate that the linear model is not an adequate representation of the relationship.

1.07 False. Removing a point with a leverage value (hii) of 1 does not guarantee a higher R2. The leverage value represents the influence of an observation on the regression model, but its removal can affect the model in various ways depending on other factors such as the position of the observation in the predictor space and the relationship between the variables. R2 may increase, decrease, or remain the same after removing an observation.

1.08 False. Taking the natural logarithm of Y does not directly reduce multicollinearity in a multiple regression. Multicollinearity refers to high correlations among the predictor variables, and it can be addressed by various techniques such as variable selection, ridge regression, or orthogonalization methods. Taking the natural log of Y may transform the relationship between Y and the predictors but does not directly address multicollinearity.

1.09 False. The expression SSReg(X1|X2) / SSE(X2) is not related to the correlation between X1 and X2 or their independence. It is not expected to be close to 0 unless there is a specific relationship between the variables in the context of the model.

1.10 True.

1.11 True.

1.12 False. Repeated application of leave-one-out cross-validation may produce slightly different estimations of error due to randomization processes involved in the technique. It involves training the model multiple times, each time leaving out one data point as a validation set. The variability in the validation set selection can lead to slightly different estimates of error.

1.13 False. Decreasing the confidence level does not decrease the width of a confidence interval on β1 from a simple linear regression. The confidence level determines the range of the interval, not its width. A higher confidence level will result in a wider interval, providing a larger range of plausible values for the parameter.

1.14 True.

1.15 True.

1.16 False. If VIF(X1) is large, it indicates high multicollinearity between X1 and the other predictors, which means that X1 is strongly correlated with the other predictors. In such a case, regressing X1 on the other X's will yield a high R2, close to 1, rather than close to 0.

1.17 True.

1.18 True.

1.19 True.

1.20 False. Multicollinearity occurs when the regressors (predictors) in a regression model are highly correlated with each other, rather than with the response variable. It is the correlation between predictors that leads to multicollinearity, which can cause issues such as unstable coefficient estimates and difficulties in interpreting the individual contributions of predictors to the response variable.

Q.2)

- The formula for variance (var($\hat{\beta}$j)) that is mentioned is relevant in the context of multicollinearity, which is a concept covered in our course.
- Multicollinearity refers to the presence of high correlation among predictor variables in a regression model.
- In multiple regression analysis, the variance of the estimated regression coefficient ($\hat{\beta}$j) represents the uncertainty or variability associated with the estimated coefficient.
- When multicollinearity is present, it can cause issues in interpreting the individual effects of predictor variables on the response variable and can lead to unstable or unreliable estimates.

- The formula that is mentioned, $\text{var}(\hat{\beta}j) = (1 / (1 - Rj^2)) * (\sigma^2 / SXiXj)$, provides a way to quantify the variance of the estimated coefficient ($\hat{\beta}j$) for a particular predictor variable (Xj) in the presence of multicollinearity.
- Here, Rj^2 represents the coefficient of determination obtained when regressing the predictor variable Xj on the other predictors in the model.
- The term σ^2 represents the variance of the error term in the regression model, and SXiXj represents the sum of squares of the predictor variable Xj.
- By understanding and calculating the variance of the estimated coefficient in the presence of multicollinearity, you can assess the precision and reliability of the estimated effect of a specific predictor variable.
- It helps you evaluate the potential impact of multicollinearity on the regression results and make informed decisions about variable selection, model specification, and interpretation of the regression coefficients.

Q.3)

If SSReg(X1, X2) = SSTO then SSReg(X1|X2)/SSE(X2) = __________.

- The expression provided is related to a statistical analysis called Analysis of Variance (ANOVA).
- In ANOVA, “SSReg” represents the sum of squares of the regression, “SSTO” represents the total sum of squares, and “SSE” represents the sum of squares of the error.
- To calculate the ratio SSReg(X1|X2)/SSE(X2), we need to understand the meaning of the notation “X1|X2”.
- This notation typically denotes the dependent variable (X1) given the independent variable (X2).
- Therefore, SSReg(X1|X2) represents the sum of squares of the regression for X1 when X2 is included in the model, and SSE(X2) represents the sum of squares of the error for X2.
- Given that SSReg(X1, X2) = SSTO, we can rewrite the expression as SSReg(X1|X2)/SSE(X2) = SSTO/SSE(X2).
- Hence, without specific information or assumptions about the relationship between X1 and X2 or the data being analyzed, it is not possible to simplify the expression any further.
- The specific values of SSTO and SSE(X2) would be needed to compute the ratio for this.

Q.4)

a.
To compute the Type 1 and Type 3 SSReg for X2, we need to consider the order of predictor inclusion in the model.
Type 1 SSReg for X2: This represents the additional sum of squares explained by X2 after accounting for the effects of other predictors.
Type 3 SSReg for X2: This represents the sum of squares explained by X2 when all other predictors are already in the model.

From the given results, we can see that X2 is included in the following terms:
- X2 alone: SSReg = 76.19
- X1 X2: SSReg = 198.41
- X2 X3: SSReg = 80.78
- X1 X2 X3: SSReg = 199.86

To compute the Type 1 SSReg for X2, we need to exclude the contributions of other predictors:
Type 1 SSReg for X2 = SSReg(X2 alone) - SSReg(X1 X2) - SSReg(X2 X3) - SSReg(X1 X2 X3)
= 76.19 - 198.41 - 80.78 - 199.86
= -402.86
Note that a negative value indicates that X2 does not contribute additional sum of squares after accounting for the effects of other predictors.

To compute the Type 3 SSReg for X2, we consider X2 with all other predictors already in the model:
Type 3 SSReg for X2 = SSReg(X2 alone) + SSReg(X1 X2) + SSReg(X2 X3) + SSReg(X1 X2 X3)
= 76.19 + 198.41 + 80.78 + 199.86
= 555.24
Therefore, the Type 1 SSReg for X2 is -402.86, and the Type 3 SSReg for X2 is 555.24.

b.
The X-matrix associated with the reduced model testing H0: $\beta_0 = 0$, $\beta_3 = 0$ would have the following first row:

$X_0 = 1$ (intercept term)
$X_1 = X_1$ (value of X_1 for the first observation)
$X_2 = X_2$ (value of X_2 for the first observation)
$X_3 = 0$ (since we are testing H0: $\beta_3 = 0$)

So, the first row of the X-matrix would be $[1, X_1, X_2, 0]$.

c.
The X-matrix associated with the reduced model testing H0: $\beta_2 = \beta_3 = 0$ would have the following first row:

$X_0 = 1$ (intercept term)
$X_1 = X_1$ (value of X_1 for the first observation)
$X_2 = 0$ (since we are testing H0: $\beta_2 = 0$)
$X_3 = 0$ (since we are testing H0: $\beta_3 = 0$)

So, the first row of the X-matrix would be $[1, X_1, 0, 0]$.

d.
The X-matrix associated with the reduced model testing H0: $\beta_1 = \beta_2 = 0$, $\beta_0 = 2$ would have the following first row:

$X_0 = 1$ (intercept term)
$X_1 = 0$ (since we are testing H0: $\beta_1 = 0$)
$X_2 = 0$ (since we are testing H0: $\beta_2 = 0$)
$X_3 = 0$ (since we are testing H0: $\beta_3 = 0$)

So, the first row of the X-matrix would be [1,1, 0, 0].

Q.5) handwritten upload (P.T.O)

Q.6)

a.

The saturated model equation for the given data would be:
$E(Y) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4$
Here, β_0, β_1, β_2, β_3, and β_4 are the coefficients of the predictor variables X_1, X_2, X_3, and X_4, respectively.

b.

- In this case, since there are 4 data points and all the x-values are distinct, we have a perfect fit, which means the model can perfectly predict the observed values of Y.
- As a result, the residuals (the differences between the observed values and the predicted values) would be zero for all the data points.
- In ordinary least squares (OLS) regression, the estimate of σ^2 (the variance of the errors) is calculated by summing the squared residuals and dividing by the degrees of freedom.
- However, with a perfect fit, the residuals are zero, and dividing by zero is undefined.
- Therefore, when there is a perfect fit (as in the saturated model), the estimate of σ^2 is undefined or infinite.

Q.7)

Given data:
Suppose a researcher has n observations from an experiment:

a) How do you think LASSO regression will treat this error in programming?
- Lasso regression typically incorporates shrinkage in the structure.
- In this type of regression, the data values are reduced to a central value.
- This central value in most cases is the mean of the data values.
- It includes regularization in the regression technique.
- When the researcher duplicates the X1 column of the matrix prior to analysis, LASSO regression will result in sparsity of the end result.

b) How do you think Ridge regression will treat this error in programming?
- Ridge regularization is another term for L2 regularization.
- It is considered as one of the methods by which models are tuned when issues of having specific multi collinearity are predominant.
- This regression model will curtail the large variance that is induced by the duplication.

Q.8)

- To find the likelihood function for three observations on the random variable X, we need to multiply the individual probabilities of observing each value in the sample.
- Let's denote the three observations as x_1, x_2, and x_3. The likelihood function, denoted as $L(\lambda)$, is given by:

$L(\lambda) = P(X=x_1|\lambda) * P(X=x_2|\lambda) * P(X=x_3|\lambda)$

Substituting the given probability distribution function, we have:

$L(\lambda) = [\lambda\text{^}(x_1) / ((e\text{^}\lambda) - 1) * x_1!] * [\lambda\text{^}(x_2) / ((e\text{^}\lambda) - 1) * x_2!] * [\lambda\text{^}(x_3) / ((e\text{^}\lambda) - 1) * x_3!]$

Simplifying further, we can rewrite this as:

$L(\lambda) = \lambda\text{^}(x_1 + x_2 + x_3) / ((e\text{^}\lambda) - 1)\text{^}(3) * (x_1! * x_2! * x_3!)$

Therefore, the likelihood function for the three observations is:

$L(\lambda) = \lambda\text{^}(x_1 + x_2 + x_3) / ((e\text{^}\lambda) - 1)\text{^}(3) * (x_1! * x_2! * x_3!)$

Q.9)

Like many other R packages, the simplest way to obtain “glmnet” is to install it directly from CRAN. Type the following command in R console:

install.packages(“glmnet”, repos = "http://cran.us.r-project.org")

Users may change the “repos” options depending on their locations and preferences. Other options such as the directories where to install the packages can be altered in the command. For more details, see:
help(install.packages)

Here the R package has been downloaded and installed to the default directories.

Q.10)

a.
To compute the researcher's estimate of β, we first need to find the median of Y's at X=2 and X=5.

For X=2:
Y = {4, 6, 7}
Median = (4 + 6 + 7) / 3 = 17 / 3 = 5.67

For X=5:
Y = {17, 18, 20}
Median = (17 + 18 + 20) / 3 = 55 / 3 = 18.33

The researcher's estimate of β is the slope of the straight line connecting these two medians:
β = (Y_median_X=5 - Y_median_X=2) / (X=5 - X=2)
  = (18.33 - 5.67) / (5 - 2)
  = 12.66 / 3
  = 4.22
Therefore, the researcher's estimate of β is approximately 4.22.

b.

To graph the empirical distribution function Fhat of the residuals resulting from the researcher's fitted model, we need to calculate the residuals first.

Residuals (ε_i) are given by:
ε_i = Y_i - β X_i

Using the estimated β = 4.22, we can calculate the residuals for each observation:
ε_1 = 4 - (4.22 * 2) = 4 - 8.44 = -4.44
ε_2 = 6 - (4.22 * 2) = 6 - 8.44 = -2.44
ε_3 = 7 - (4.22 * 2) = 7 - 8.44 = -1.44
ε_4 = 17 - (4.22 * 5) = 17 - 21.1 = -4.1
ε_5 = 18 - (4.22 * 5) = 18 - 21.1 = -3.1
ε_6 = 20 - (4.22 * 5) = 20 - 21.1 = -1.1

Now, let's graph the empirical distribution function Fhat of the residuals:
Residuals: {-4.44, -2.44, -1.44, -4.1, -3.1, -1.1}
Sorted Residuals: {-4.44, -4.1, -3.1, -2.44, -1.44, -1.1}
Fhat = {1/6, 2/6, 3/6, 4/6, 5/6, 6/6}

The graph of Fhat will show a step function with the x-axis representing the sorted residuals and the y-axis representing the cumulative probability.

c.
To compute the first bootstrapped sample estimate of β, we use the given set of random numbers:
Random numbers: {3, 1, 2, 4, 6, 6}

This set represents a resampling with replacement from the original dataset.
We take the corresponding Y values for these randomly selected X values and fit a line using robust regression to estimate β.

X values: {2, 2, 2, 5, 5, 5}
Y values: {4, 4, 4, 18, 18, 18}

Using robust regression, we can estimate β by fitting a line to the medians of the Y values at X=2 and X=5.

For X=2:
Y = {4, 4, 4}
Median = 4

For X=5:
Y = {18, 18, 18}
Median = 18

The bootstrapped estimate of β is the slope of the straight line connecting these medians:
β = (Y_median_X=5 - Y_median_X=2) / (X=5 - X=2)
  = (18 - 4) / (5 - 2)
  = 14 / 3

≈ 4.67

Therefore, the first bootstrapped sample estimate of β for the researcher is approximately 4.67.

Q.5)

Given experiment has been carried out 4 times

Data is:-

Observation	Response (y)	Explanatory Variable		
1	y_1	$x_1 = 0$	$x_1^2 = 0$	(0,1)
2	y_2	$x_2 = 2$	$x_2^2 = 4$	(2,3)
3	y_3	$x_3 = 4$	$x_3^2 = 16$	(4,14)
4	y_4	$x_4 = 6$	$x_4^2 = 36$	(6,38)

a) Model assumed is $\Rightarrow E(y) = \beta_1 x + \beta_2 x^2 \rightarrow (1)$

Given data is also assumed to follow (1), hence:

$$y_1 = \beta_1(0) + \beta_2(0)$$
$$y_2 = \beta_1(2) + \beta_2(4)$$
$$y_3 = \beta_1(4) + \beta_2(16)$$
$$y_4 = \beta_1(6) + \beta_2(36)$$

In matrix form, $\boxed{Y = X\beta}$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 0 \\ 2 & 4 \\ 4 & 16 \\ 6 & 36 \end{bmatrix}}_{(4\times 2)} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

Hence, $X = \begin{bmatrix} 0 & 0 \\ 2 & 4 \\ 4 & 16 \\ 6 & 36 \end{bmatrix}$, and shape is (4×2)

4 rows
2 columns.

(b)

$$X' \cdot X = \begin{bmatrix} 0 & 2 & 4 & 6 \\ 0 & 4 & 16 & 36 \end{bmatrix}_{(2\times4)} \begin{bmatrix} 0 & 0 \\ 2 & 4 \\ 4 & 16 \\ 6 & 36 \end{bmatrix}_{(4\times2)}$$

$$X' \cdot X = \begin{bmatrix} 56 & 288 \\ 288 & 1568 \end{bmatrix}_{(2\times2)}$$

(c) $X' \cdot Y =$

In matrix form, $\boxed{Y = X\beta}$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 2 & 4 \\ 4 & 16 \\ 6 & 36 \end{bmatrix}_{(4\times2)} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}_{(2\times1)}$$

$$Y = \begin{bmatrix} 0 \\ 2\beta_1 + 4\beta_2 \\ 4\beta_1 + 16\beta_2 \\ 6\beta_1 + 36\beta_2 \end{bmatrix}_{(4\times1)}$$

$$X' \cdot Y = \begin{bmatrix} 0 & 2 & 4 & 6 \\ 0 & 4 & 16 & 36 \end{bmatrix}_{(2\times4)} \begin{bmatrix} 0 \\ 2\beta_1 + 4\beta_2 \\ 4\beta_1 + 16\beta_2 \\ 6\beta_1 + 36\beta_2 \end{bmatrix}_{(4\times1)}$$

$$= \begin{bmatrix} 4\beta_1 + 8\beta_2 + 16\beta_1 + 64\beta_2 + 36\beta_1 + 216\beta_2 \\ 8\beta_1 + 16\beta_2 + 64\beta_1 + 256\beta_2 + 216\beta_1 + 1296\beta_2 \end{bmatrix}_{(2\times1)}$$

$$X' \cdot Y = \begin{bmatrix} 56\beta_1 + 288\beta_2 \\ 288\beta_1 + 1568\beta_2 \end{bmatrix}_{(2\times2)}$$