Regression Analysis
Quiz-4

Likhit Garimella
lg836

Q.1)

a. False.
- Forward selection and backward elimination based on p-to-enter and p-to-remove may not always result in the same final model, even if the p-to-enter and p-to-remove thresholds are the same.
- This is because the order in which variables are added or removed can affect the final model.

b. True.
- In logistic regression, the coefficient estimates represent the change in the log-odds of the response variable for a one-unit increase in the corresponding explanatory variable.
- Therefore, if the estimated coefficient associated with X is -0.2, it means that the log-odds of Y=1 will decrease by 0.2 when X increases by one unit.

c. False.
- Increasing the cutoff value for classifying a new observation as an Event in a logistic regression model does not necessarily decrease the sensitivity of the model.
- The sensitivity, also known as the true positive rate, measures the proportion of actual positive cases correctly classified as positive.
- Adjusting the cutoff value may impact the model's sensitivity, specificity, accuracy, and other performance metrics in complex ways.

d. False.
- The availability of the output from all possible regressions for variable subset selection does not necessarily imply that stepwise selection methods should be avoided.
- Stepwise selection methods can still be useful, especially when dealing with a large number of variables, as they provide a more automated approach to select a subset of variables based on certain criteria.

e. False.
- The odds of an event occurring is defined as the probability of the event divided by the probability of the complement event.
- In this case, the odds of Win to Loss is given as 0.25.
- To calculate the probability of Win, you would divide the odds of Win (0.25) by the sum of the odds and 1 (0.25 + 1 = 1.25).
- Therefore, the probability of Win is 0.25/1.25 = 0.20.

Q.2)

a. given sensitivity and specificity values.

- From the given information, we have:
- Sensitivity = True Positives / (True Positives + False Negatives)
- Sensitivity = X / (X + 130)

130/205 = X / (X + 130)
130(X + 130) = 205X

$130X + 16900 = 205X$
$75X = 16900$
$X = 16900 / 75$
$X = 225.33$ (approx.)

- Specificity = True Negatives / (True Negatives + False Positives)
- Specificity = Y / (Y + 60)

$60/105 = Y / (Y + 60)$
$105(Y + 60) = 60Y$
$105Y + 6300 = 60Y$
$45Y = 6300$
$Y = 6300 / 45$
$Y = 140$

Therefore, $X \approx 225.33$ and $Y = 140$.

b. Positive predictive value (PPV) is the proportion of true positive cases among all the cases that tested positive.
PPV = True Positives / (True Positives + False Positives)
$PPV = X / (X + 60)$
$PPV = 225.33 / (225.33 + 60)$
$PPV \approx 0.789$

c. Negative predictive value (NPV) is the proportion of true negative cases among all the cases that tested negative.
NPV = True Negatives / (True Negatives + False Negatives)
$NPV = Y / (Y + 130)$
$NPV = 140 / (140 + 130)$
$NPV \approx 0.519$

d. Accuracy is the proportion of correct predictions out of all the cases.
Accuracy = (True Positives + True Negatives) / (Total Cases)
Accuracy = $(X + Y) / (X + Y + 60 + 130)$
Accuracy = $(225.33 + 140) / (225.33 + 140 + 60 + 130)$
Accuracy ≈ 0.689

Therefore, the answers are:
a. $X \approx 225.33$, $Y = 140$
b. Positive predictive value (PPV) ≈ 0.789
c. Negative predictive value (NPV) ≈ 0.519
d. Accuracy ≈ 0.689

Q.3)

- two regression models: one with four predictors X1, X2, X3, and X4, and the other with three predictors X1, X2, and X3.
- given information: SSR(X1, X2, X3, X4) = SSR(X1, X2, X3).

- The coefficient of determination (R-squared, R2) is a measure that explains the proportion of the variance in the dependent variable that can be explained by the independent variables in the model.
- The adjusted R-squared (R2adj) is a variant of R-squared that takes into account the number of predictors in the model to avoid overfitting.
- To compare the R2adj values for the two models, we need to look at their formula:
- For the model with four predictors (X1, X2, X3, and X4): R2adj(X1, X2, X3, X4) = 1 - (SSR(X1, X2, X3, X4) / SST) * ((n - 1) / (n - p - 1)), where SST is the total sum of squares and n is the number of data points, and p is the number of predictors (4 in this case).
- For the model with three predictors (X1, X2, and X3): R2adj(X1, X2, X3) = 1 - (SSR(X1, X2, X3) / SST) * ((n - 1) / (n - p - 1)), where p is the number of predictors (3 in this case).
- Given that SSR(X1, X2, X3, X4) = SSR(X1, X2, X3), we can see that both models have the same numerator in their adjusted R-squared formulas.
- The only difference between the two models lies in the denominator, as they have different numbers of predictors (p).
- Since the denominator in the adjusted R-squared formula for the model with three predictors (X1, X2, and X3) is smaller than the denominator in the formula for the model with four predictors (X1, X2, X3, and X4), the adjusted R-squared value for the three-predictor model will be larger.
- Therefore, R2adj for the model with X1, X2, X3 is larger than R2adj for the model with X1, X2, X3, and X4.
- This result is because the three-predictor model has a higher degree of freedom, which tends to increase the adjusted R-squared value and better accounts for potential overfitting.

Q.4)

- The sequence described, where x20 enters the model first, followed by x1 and x11, and then x20 is subsequently removed, can occur due to the nature of stepwise procedures and the specific criteria used for variable selection and removal.
- Stepwise procedures are commonly used in statistical modeling to select variables for inclusion in a model based on their significance or predictive power.
- These procedures involve iteratively adding or removing variables from the model based on certain criteria, such as p-values or information criteria like AIC or BIC.
- In our case, the researcher started with an empty model and began the stepwise procedure by considering all 20 variables.
- The procedure determined that x20 was the most significant variable and entered it into the model.
- Next, the researcher re-evaluated the model with x20 and considered the remaining variables.
- At this point, x1 and x11 were found to be the next most significant variables and were added to the model.
- Now, the stepwise procedure re-evaluates the model with x20, x1, and x11.
- It is at this stage that the stepwise procedure determined that x20 is no longer significant or does not contribute significantly to the model after accounting for x1 and x11.
- Therefore, x20 is removed from the model, resulting in a final model with only x1 and x11.
- This sequence can occur because the stepwise procedure evaluates the variables in a forward and backward manner, considering both the addition and removal of variables at each step based on the defined criteria.

- The specific combination of variables selected and removed depends on their individual significance and their interactions with other variables in the model.

Q.5)

- To find the Maximum Likelihood Estimator (MLE) for β, we need to maximize the likelihood function.
- In logistic regression, the likelihood function is given by the product of the probabilities of the observed outcomes.

Given the logistic regression model:
Prob(Y=1) = exp(βx) / (1 + exp(βx))
Let's calculate the likelihood for each data point:
For (X, Y) = (0, 0):
Prob(Y=0) = 1 - Prob(Y=1) = 1 - exp(β * 0) / (1 + exp(β * 0)) = 1 / (1 + 1) = 1/2
For (X, Y) = (0.1, 1):
Prob(Y=1) = exp(β * 0.1) / (1 + exp(β * 0.1))
For (X, Y) = (1, 1):
Prob(Y=1) = exp(β * 1) / (1 + exp(β * 1))
For (X, Y) = (10, 1):
Prob(Y=1) = exp(β * 10) / (1 + exp(β * 10))
Now, the likelihood function by multiplying the probabilities of the observed outcomes:
L(β) = (1/2) * [exp(β * 0.1) / (1 + exp(β * 0.1))] * [exp(β * 1) / (1 + exp(β * 1))] * [exp(β * 10) / (1 + exp(β * 10))]

- To find the MLE for β, we need to maximize this likelihood function.
- Since we have only two possible choices for β, 0 or 1, we can calculate the likelihood function for both cases and choose the one that gives the maximum likelihood.

For β = 0:
L(0) = (1/2) * [exp(0.1 * 0) / (1 + exp(0.1 * 0))] * [exp(0.1 * 1) / (1 + exp(0.1 * 1))] * [exp(0.1 * 10) / (1 + exp(0.1 * 10))]
    = (1/2) * (1 / (1 + 1)) * (1 / (1 + exp(0.1))) * (1 / (1 + exp(1)))
    = 0.5 * 0.5 * 0.909 * 0.731
    ≈ 0.166
For β = 1:
L(1) = (1/2) * [exp(0.1 * 1) / (1 + exp(0.1 * 1))] * [exp(0.1 * 1) / (1 + exp(0.1 * 1))] * [exp(0.1 * 10) / (1 + exp(0.1 * 10))]
    = (1/2) * (1 / (1 + exp(0.1))) * (1 / (1 + exp(0.1))) * (1 / (1 + exp(1)))
    = 0.5 * 0.909 * 0.909 * 0.731
    ≈ 0.243

- Comparing the likelihoods for β = 0 and β = 1, we can see that L(1) ≈ 0.243 is greater than L(0) ≈ 0.166.
- Therefore, the maximum likelihood estimate for β in the logistic regression model is β = 1.

Q.6)

a. x1 —> x4 —> x5 —> x2 —> x3
b. Same sequence as (a)
c. x1 —> x3 —> x2 —> x5 —> remove x1 <u>STOP</u>
d. x1 —> x3 —> x2 —> x5 —> remove x1 <u>STOP</u> same sequence as (c)

e. Type 1 AIC for x4 = AIC_with_x4 - AIC_without_x4
f. Type 3 AIC for x4 = AIC_full - AIC_reduced

Q.7)

a)
- The Wald test statistic for testing H0: $\beta4 = 0$ can be calculated using the formula:
- Wald statistic = $(\beta4 - 0)$ / Std. Error($\beta4$)
- From the given output, the estimate for $\beta4$ is 1.452990, and the standard error for $\beta4$ is 0.463061.
- => Wald statistic = $(1.452990 - 0) / 0.463061 \approx 3.136$

b)
- To test the hypothesis H0: all variables can be dropped from the model, we can perform a likelihood ratio test (LRT) by comparing the residual deviance of the full model to the residual deviance of the null model (which includes only the intercept term).
- The LRT statistic is calculated as:
- LRT statistic = Deviance(full model) - Deviance(null model)
- From the given output, the residual deviances are:
- Residual deviance (full model) = 184.21
- Residual deviance (null model) = 202.82
- => LRT statistic = $184.21 - 202.82 \approx -18.61$
- To perform the test, we compare the LRT statistic to the chi-squared distribution with degrees of freedom equal to the difference in degrees of freedom between the two models (full model - null model).
- In this case, the difference is $163 - 167 = -4$, but since degrees of freedom cannot be negative, we take the absolute value.
- We can then compare the LRT statistic to the critical value of the chi-squared distribution at the desired significance level (e.g., 0.05).
- If the LRT statistic is larger than the critical value, we reject the null hypothesis that all variables can be dropped from the model.

c)
- To test the hypothesis H0: $\beta2 = \beta3 = \beta4 = 0$, we need to compare the full model (with all predictors) to a reduced model that only includes the intercept term (null model).
- We can perform a likelihood ratio test (LRT) by comparing the residual deviance of the full model to the residual deviance of the null model.
- If the answer is not already available from the output, we would need to run two models: one with all predictors (X1, X2, X3, X4) and another with only the intercept term.

d)
- The Odds Ratio (OR) shown for X4 in the overall model indicates the multiplicative change in the odds of the response variable (Y) associated with a one-unit increase in X4, compared to the reference category.
- From the output, the Odds Ratio for X4: 1 vs 0 is 4.76.
- This means that, holding all other predictors constant, the odds of Y being 1 are approximately 4.76 times higher for the group represented by X4 = 1 compared to the group represented by X4 = 0.
- In other words, the presence of X4 (X4 = 1) is associated with a significantly higher likelihood of the response variable Y being 1, according to the logistic regression model.