KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DEPARTMENT OF MATHEMATICS

STAT 513: Statistical Modeling

Term 241, Final Exam

Wednesday December 17, 2024, 7:00 PM

Time allowed **150 minutes**

Name: _____ ID #: _____

Question No	Full Marks	Marks Obtained
1	**30**	
2	**18**	
3	**22**	
Total	**70**	

Instructions:

- Mobiles are not allowed in exam. If you have your **mobile** with you, **turn it off** and put it **on the table/floor** so that it is visible to the proctor.
- Show all the calculation steps. There are points for the steps so if you miss them, you lose points. For multiple choice type questions, showing calculation steps is not required.
- Report **at least 4 decimal points** of your numerical answers.

Q1: (30 points) Answer the following 30 questions:

1. Why is extrapolation particularly risky in polynomial regression models of higher order?

(A) It introduces bias by forcing the model to fit data points outside the observed range.
(B) Extrapolation in high-order polynomials can result in an overly complex model.
(C) Polynomial models overfit the data, leading to high variance for out-of-sample predictions.
(D) The relationship modeled by the polynomial may change or behave differently outside the range of observed data.
(E) Extrapolation causes underfitting due to the complexity of the polynomial.

2. Which of the following is a key reason for selecting a spline model over a polynomial model?

(A) It requires fewer parameters to be estimated
(B) It avoids multicollinearity issues present in polynomial regression
(C) It provides more straightforward interpretability
(D) It reduces the residual standard error significantly
(E) It provides a better fit for data where the relationship changes behavior at different ranges of the predictor

3. Which of the following is a key consequence of multicollinearity in regression analysis?

(A) It leads to large variances for the estimated regression coefficients.
(B) It makes the regression coefficients more accurate.
(C) It reduces the correlation between the predictors and the response variable.
(D) It improves the stability of the least-squares estimates.
(E) It ensures unbiased estimates of regression coefficients.

4. Consider a scenario where a company is analyzing the relationship between monthly sales revenue and advertising expenses across several locations. Upon fitting a linear regression model, you notice that the residuals exhibit increasing variance as the advertising expenses grow larger. Which method would be most appropriate to handle this situation?

(A) Box-Tidwell, by transforming the model to satisfy the assumptions of constant variance and no correlation among errors.
(B) Polynomial regression, by adding higher-degree terms of advertising expenses.
(C) Ridge regression, by introducing regularization to reduce variance.
(D) Logistic regression, if the response variable is binary.
(E) Weighted Least Squares, by using weights inversely proportional to the variance of the residuals.

5. You are conducting a study to analyze the effect of different education levels (High School, Bachelor, Master and Doctorate) on income. You decide to use indicator variables to represent these education levels. How many indicator variables are required to represent this qualitative variable (education level) in your regression model?

(A) Two
(B) Four
(C) One
(D) Five
(E) Three

6. Which of the following correctly describes the probability calculation in a Probit model?
Note: $\Phi(.)$ represents the CDF of standard normal distribution.

(A) $\pi_i = \frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}$
(B) $\pi_i = \lambda e^{-\lambda X_i}$
(C) $\pi_i = \Phi(\beta_0 + \beta_1 X_i)$
(D) $\pi_i = \Phi\left(e^{\beta_0+\beta_1 X_i}\right)$
(E) $\pi_i = \frac{1}{1+e^{-\beta_0-\beta_1 X_i}}$

7. In backward elimination, which predictor is examined first for removal?

(A) The predictor with the highest correlation with the response variable.
(B) The predictor with the highest absolute t statistic.
(C) The predictor with the highest p-value.
(D) The predictor with the smallest coefficient magnitude.
(E) The predictor with the smallest p-value.

8. Which one of the following is true about forward selection, backward elimination and stepwise regression?

(A) Stepwise regression always produces the "best" model regardless of data quality.
(B) Different methods can lead to different final models.
(C) The procedure eliminates the need for exploratory data analysis.
(D) Results from these methods are invariant to the choice of significance levels.
(E) These methods guarantee the inclusion of all important predictors.

9. Which of the following is a limitation of using pairwise correlation between predictors to diagnose multicollinearity?

(A) Pairwise correlation may fail to indicate multicollinearity when there are more than two predictors involved.
(B) Pairwise correlation will always detect multicollinearity, regardless of the number of predictors.
(C) Pairwise correlation can only detect multicollinearity in nonlinear models.
(D) Pairwise correlation works well in models with more than three predictors.
(E) Pairwise correlation indicates multicollinearity only when the correlation is exactly 1 or -1.

10. In a Logit model $\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_i$ where X is a continuous predictor, what does the parameter β_1 represent?

(A) The predicted probability of success given X
(B) The expected value of Y for a one-unit increase in X
(C) The indicator variable used to classify success or failure
(D) The predicted probability of failure given X
(E) The change in log-odds for a one-unit increase in X

11. In the context of LASSO Regression, what is a key feature that differentiates it from Ridge Regression?

(A) LASSO uses the sum of squared coefficients for the penalty term.
(B) LASSO does not require the use of cross-validation to select the tuning parameter.
(C) LASSO produces a closed-form solution for coefficient estimation.
(D) LASSO applies a penalty proportional to the absolute value of the coefficients.
(E) LASSO only works when multicollinearity is not present.

12. Which of the following statements is true regarding the hat matrix?

(A) The diagonal elements of the hat matrix measure the distance of the i^{th} observation from the center of the x-space.
(B) The sum of all diagonal elements of the hat matrix equals 0.
(C) An influential point is determined by the elements of hat matrix.
(D) Any diagonal element of hat matrix > 0 is considered a leverage point.
(E) The hat matrix is used only to measure the residuals in regression.

13. When calculating the influence of an observation using $DFBETAS_{j,i}$, what does this measure represent?

(A) The residual change for the ith observation after removing the jth predictor.
(B) The change in the regression coefficient for the jth predictor when the ith observation is removed.
(C) The overall change in the model fit when the ith observation is deleted.
(D) The change in the variance of the regression coefficients when the ith observation is deleted.
(E) The change in the fitted value for the ith observation when the jth predictor is excluded.

14. In the context of polynomial regression, which of the following is the main concern when using higher-order polynomial models?

(A) Ill-conditioning
(B) Underfitting
(C) Non-linearity of the relationship
(D) Homogeneity of the residuals
(E) Non-independence of observations

15. In a Logit model $\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_i$ where X is a continuous predictor, if $\hat{\beta}_1$ = 0.5, what is the change in odds ratio for a one-unit increase in X?

(A) $e^{0.5}$
(B) $\ln 0.5$
(C) 1
(D) 0.5
(E) $e^{\frac{0.5}{1-0.5}}$

16. In a regression model $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$, where X_1 is the continuous predictor and X_2 is a dummy variable for Gender, when testing if two regression equations for Males and Females are the same, the null hypothesis H_0 is:

(A) $\beta_1 = \beta_3 = 0$
(B) $\beta_2 = 0$
(C) $\beta_0 = \beta_1 = 0$
(D) $\beta_2 \neq \beta_3 \neq 0$
(E) $\beta_2 = \beta_3 = 0$

17. When using Iteratively Reweighted Least Squares (IRLS) in robust regression, which of the following steps is correct?

(A) Fit the model using standard least squares, and then adjust the model based on the initial residuals.
(B) Compute residuals, calculate weights based on these residuals, and solve the weighted least squares problem iteratively until the parameters stabilize.
(C) Remove all observations with large residuals and re-fit the model using the remaining points.
(D) The weights are calculated based on the original least squares solution without iterating.
(E) IRLS does not require any iteration and uses the same procedure as traditional least squares.

18. The Box-Cox transformation requires selecting a value for the transformation parameter (λ). Which of the following is the interpretation of the value $\lambda = 0$?

(A) The transformation is equivalent to taking the natural logarithm of the response variable.
(B) The transformation is equivalent to raising the response variable to the power of 2.
(C) The transformation leaves the response variable unchanged.
(D) The transformation is equivalent to applying the inverse of the response variable.
(E) The transformation makes the response variable normally distributed with zero mean and unit variance.

19. When building a polynomial regression model, which of the following is typically recommended when encountering ill-conditioning in the design matrix $\boldsymbol{X}$?

(A) Fitting a higher-order polynomial
(B) Using a different regression method
(C) Removing outliers
(D) Transforming the response variable
(E) Centering the predictors

20. In the case of multicollinearity between two predictors, X_1 and X_2, how does the variance of the least-squares estimate of the regression coefficients behave as the correlation between the predictors approaches 1 or -1?

(A) The variance increases without bound as the correlation approaches ± 1.
(B) The variance decreases as the correlation approaches ± 1.
(C) The variance remains constant regardless of the correlation between the predictors.
(D) The variance becomes negative as the correlation approaches ± 1.
(E) The variance approaches zero as the correlation approaches ± 1.

21. In the context of weighted least squares, which of the following methods can be used to estimate the weights when prior knowledge of the weights is not available?

(A) Estimating the weights using the standard deviation of the residuals from an initial ordinary least squares fit.
(B) Assuming constant variance across all observations and applying OLS directly.
(C) Using K-Nearest Neighbors (KNN) to estimate the local variance of the observations and then calculating the weights.
(D) Using cross-validation to determine the best weights for the model.
(E) Performing a Box-Cox transformation to estimate the weights.

22. In a typical regression flow chart, what is the primary goal of exploratory data analysis?

(A) To split the data into training and test sets based on a random process.
(B) To identify patterns, correlations, and anomalies in the data that may inform model specification.
(C) To determine the optimal number of predictors to include in the model.
(D) To transform variables into the appropriate format for training the model.
(E) To calculate the performance metrics of the regression model.

23. Which of the following is a key reason to perform transformations during regression modeling?

(A) To increase the correlation between predictors.
(B) To ensure that predictors have equal variances.
(C) To minimize the number of outliers in the model.
(D) To maximize the MSE of model.
(E) To address violations of model assumptions.

24. Which link function is used in a Poisson regression model to relate the mean response to the linear predictor?

(A) Identity link
(B) Inverse Normal CDF link
(C) Log link
(D) Inverse Logistic CDF link
(E) Exponential link

25. In regression analysis, which of the following best describes an influential point?

(A) A point that is distant from the center of the x-space.
(B) A point that falls on the regression line.
(C) A point that has a high leverage.
(D) A point that is unusual in both y and x.
(E) A point that does not significantly change the regression model.

26. When applying a cubic spline model, how are the knots typically chosen?

(A) By fitting the model with as many knots as possible
(B) Based on minimizing the residual sum of squares
(C) Based on domain knowledge or data behavior at specific thresholds
(D) Using automatic selection criteria like AIC or BIC
(E) By choosing equidistant points along the range of the data

27. What is the primary criterion for adding a predictor during forward selection?

(A) The predictor must have the lowest variance among all available predictors.
(B) The predictor must be selected by domain experts without statistical testing.
(C) The predictor must minimize the adjusted R-squared of the model.
(D) The predictor must have zero correlation with other predictors in the model.
(E) The t-test corresponding to the predictor must be significant according to a pre-defined α.

28. Stepwise regression differs from forward selection by:

(A) Eliminating predictors based only on their correlation with the response variable.
(B) Reassessing the contribution of predictors already in the model and removing them if necessary.
(C) Including all predictors initially and not removing any throughout the procedure.
(D) Using only subjective criteria for predictor inclusion.
(E) Ignoring partial F-statistics for model refinement.

29. Which of the following statements is not true about robust regression through Huber's function?

(A) The weights decrease as the magnitude of residuals increases.
(B) The function is less sensitive to outliers compared to ordinary least squares regression.
(C) The function uses a threshold to switch between full weight and reduced weight.
(D) The values with residuals close to zero are assigned full weight.
(E) Extreme outliers are assigned zero weight.

30. In a regression model $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$, where X_1 is the continuous predictor and X_2 is a dummy variable for Gender, when testing if two regression equations for Males and Females are the same, which test should be used?

(A) Partial F-test
(B) t-test
(C) Chi-square test
(D) Z-test
(E) Likelihood ratio test

Q2: A dataset is collected from a public health study aimed at identifying factors associated with heart disease diagnosis. The response variable indicates whether an individual was diagnosed with heart disease based on their health profile. The predictors include:

`CholesterolLevel`: The cholesterol level of an individual (measured in mg/dL, continuous predictor).
`SmokingStatus`: Whether the individual is a "`Non-Smoker`", "`Former Smoker`", or "`Current Smoker`".

Response Variable:
`HeartDisease`: Indicates whether the individual was diagnosed with heart disease (`1` for `Yes`, `0` for `No`).

The data scientist fits some models and the outputs are given below:

```
> mylogit <- glm(HeartDisease ~ CholesterolLevel + factor(SmokingStatus),
+                data = health_data, family = binomial)
> summary(mylogit)

Call:
glm(formula = HeartDisease ~ CholesterolLevel + factor(SmokingStatus),
    family = binomial, data = health_data)

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -2.23311    1.23918  -1.802  0.07153 .
CholesterolLevel                  0.01869    0.00598   3.126  0.00177 **
factor(SmokingStatus)Former Smoker -1.29346    0.62904  -2.056  0.03976 *
factor(SmokingStatus)Non-Smoker    -1.63092    0.57772  -2.823  0.00476 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 174.28  on 126  degrees of freedom
Residual deviance: 154.74  on 123  degrees of freedom
```

```
> myprobit <- glm(HeartDisease ~ CholesterolLevel + factor(SmokingStatus),
+                data = health_data, family = binomial(link = "probit"))
> summary(myprobit)

Call:
glm(formula = HeartDisease ~ CholesterolLevel + factor(SmokingStatus),
    family = binomial(link = "probit"), data = health_data)

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -1.392652   0.735814  -1.893  0.05840 .
CholesterolLevel                  0.011574   0.003523   3.285  0.00102 **
factor(SmokingStatus)Former Smoker -0.785906   0.370985  -2.118  0.03414 *
factor(SmokingStatus)Non-Smoker    -1.003298   0.336100  -2.985  0.00283 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 174.28  on 126  degrees of freedom
Residual deviance: 154.51  on 123  degrees of freedom
```

a) (3 pts.) Can we interpret the coefficient of `CholesterolLevel` in the model named `mylogit` i.e. `0.01869`.? If yes, give the interpretation.

b) (4 pts.) With reference to model named `mylogit`, discuss in detail the significance of `factor(SmokingStatus)Non-Smoker`. The value of coefficient is `-1.63092`, what is the interpretation of a negative coefficient in this case? Moreover, the p-value is small indicating the significance of coefficient. What is the practical interpretation of this significance?

c) (3 pts.) Can we interpret the coefficient of `CholesterolLevel` in the model named `myprobit` i.e. `0.011574`.? If yes, give the interpretation.

d) (4 pts.) Following are some new commands and their respective outputs:

```
> wald.test(b = coef(mylogit), Sigma = vcov(mylogit), Terms = 3:4)
Wald test:
----------

Chi-squared test:
X2 = 8.0, df = 2, P(> X2) = 0.019
```

What exactly is tested in the above output and what conclusions do you draw? Write the null and alternative hypotheses and explain in detail.

H_0:

H_1:

Decision:

Conclusion:

e) (4 pts.) Below is the data for two new patients, for whom we aim to predict the `HeartDisease`:

```
> new_patients <- data.frame(
+   CholesterolLevel = c(220, 180),
+   SmokingStatus = c("Current Smoker", "Non-Smoker")
+ )
>
> predict(mylogit, newdata = new_patients, type = "link")
         1          2
 1.8794042 -0.4992424
```

Prediction of `HeartDisease` for Patient 1:

- Yes
- No

Justification of your answer:

Prediction of `HeartDisease` for Patient 2:

- Yes
- No

Justification of your answer:

Name: _____ ID #: _____ Version: ____

Q3: A Student performance survey is designed to examine the factors influencing academic student performance. The dataset consists of 3748 student records, with each record containing information about various features and a performance index.

Features:
Hours Studied (HS): The total number of hours spent studying by each student.
Previous Scores (PS): The scores obtained by students in previous tests.
Extracurricular Activities (EA): Whether the student participates in extracurricular activities (Yes or No).
Sleep Hours (SH): The average number of hours of sleep the student had per day.
Sample Question Papers Practiced (SQPP): The number of sample question papers the student practiced.

Target Variable:
Performance Index (PI): A measure of the overall performance of each student. The performance index represents the student's academic performance and has been rounded to the nearest integer. The index ranges from 10 to 100, with higher values indicating better performance.

Report your answer correct up to 6 decimal points.

a) (2 pts.) Construct an indicator variable against Extracurricular Activities for "`Yes`" category, keeping "`No`" as default. Fit a multiple linear model for predicting Performance Index of student based on Hours Studied, Previous Scores, Extracurricular Activities, Sleep Hours and Sample Question Papers Practiced. Write down the fitted model.

b) (2 pts.) Compute the $R^2_{\text{Prediction}}$ for the fitted model, write computational details and interpret your answer. You can use the following formulas:
$R^2_{\text{Prediction}} = 1 - \frac{PRESS}{SST}$ where $PRESS = \sum_{i=1}^{n} e^2_{(i)}$ and $e_{(i)} = \frac{e_i}{1-h_{ii}}$.

c) (3 pts.) Evaluate whether ridge regression is needed for this problem. Use appropriate diagnostics to support your decision and provide a clear justification based on your findings.

d) (3 pts.) Evaluate whether Weighted Least Squares (WLS) are needed for this problem. Use appropriate diagnostics to support your decision and provide a clear justification based on your findings.

e) (4 pts.) Apply the WLS using replicates on response variable generated through KNN. Write down the equation of fitted WLS model.

f) (3 pts.) Fit robust regression model to the dataset using Bisquare function. After fitting the model, identify the observation that received the minimum weight. What is the value of this minimum weight?

g) (2 pts.) Determine the roll number of the student corresponding to the observation that received the minimum weight.

h) (3 pts.) Using the model fitted in (g), obtain a 90% interval estimate for the performance index of a student who studied for 5 hours, scored 70 in previous tests, participates in extracurricular activities, sleeps 6 hours per day and practiced 5 sample question papers. Write the fitted model, computed interval, and its interpretation.