Regression and Survival Analysis – Quiz

Statistical Data Analysis

Question 1

A researcher fits a simple linear regression $\hat{Y} = 41.5 + 7.5X$ predicting exam scores from study hours. If a student studies for 4 hours, what is the predicted score?

A. 49.0

B. 71.5

C. 30.0

D. 78.0

Question 1

A researcher fits a simple linear regression $\hat{Y} = 41.5 + 7.5X$ predicting exam scores from study hours. If a student studies for 4 hours, what is the predicted score?

A. 49.0

B. 71.5

C. 30.0

D. 78.0

Answer: B

Substituting $X = 4$ into the equation: $\hat{Y} = 41.5 + 7.5 \times 4 = 41.5 + 30 = 71.5$. The intercept 41.5 is the baseline score when study hours equal zero, and each additional hour adds 7.5 points.

Question 2

In OLS regression, the slope $\hat{\beta}_1$ is computed as $\text{Cov}(X, Y)/\text{Var}(X)$. Given $\sum(x_i - \bar{x})(y_i - \bar{y}) = 120$ and $\sum(x_i - \bar{x})^2 = 40$, what is $\hat{\beta}_1$?

A. 0.33

B. 80.0

C. 4,800

D. 3.0

Question 2

In OLS regression, the slope $\hat{\beta}_1$ is computed as $\text{Cov}(X, Y)/\text{Var}(X)$. Given $\sum(x_i - \bar{x})(y_i - \bar{y}) = 120$ and $\sum(x_i - \bar{x})^2 = 40$, what is $\hat{\beta}_1$?

A. 0.33

B. 80.0

C. 4,800

D. 3.0

Answer: D

The slope is $\hat{\beta}_1 = 120/40 = 3.0$. This means each one-unit increase in X is associated with a 3-unit increase in Y, based on the ratio of the cross-deviation to the sum of squared deviations of X.

Question 3

A simple linear regression yields $\hat{\beta}_1 = 5.2$ with $SE(\hat{\beta}_1) = 1.3$ on $n = 22$ observations. To test $H_0 : \beta_1 = 0$ at $\alpha = 0.05$ (two-tailed, $t_{0.025,20} = 2.086$), what is the correct conclusion?

A. Reject H_0 because $t = 4.0 > 2.086$
B. Fail to reject H_0 because $t = 4.0 < 5.0$
C. Fail to reject H_0 because the sample size is too small
D. Reject H_0 because SE is less than $\hat{\beta}_1$

Question 3

A simple linear regression yields $\hat{\beta}_1 = 5.2$ with $SE(\hat{\beta}_1) = 1.3$ on $n = 22$ observations. To test $H_0 : \beta_1 = 0$ at $\alpha = 0.05$ (two-tailed, $t_{0.025,20} = 2.086$), what is the correct conclusion?

A. Reject H_0 because $t = 4.0 > 2.086$

B. Fail to reject H_0 because $t = 4.0 < 5.0$

C. Fail to reject H_0 because the sample size is too small

D. Reject H_0 because SE is less than $\hat{\beta}_1$

Answer: A

The test statistic is $t = \hat{\beta}_1 / SE(\hat{\beta}_1) = 5.2/1.3 = 4.0$. Since $4.0 > 2.086$, we reject H_0 and conclude there is a statistically significant linear relationship between X and Y at the 5% level.

Question 4

If a regression model has $SSR = 400$ and $SST = 500$, what is the R^2 value?

A. 0.20
B. 0.50
C. 0.80
D. 1.25

Question 4

If a regression model has $SSR = 400$ and $SST = 500$, what is the R^2 value?

A. 0.20
B. 0.50
C. 0.80
D. 1.25

Answer: C
$R^2 = SSR/SST = 400/500 = 0.80$. This means 80% of the total variation in the response variable is explained by the regression model. Equivalently, $R^2 = 1 - SSE/SST = 1 - 100/500 = 0.80$.

Question 5

A model with 3 predictors and $n = 50$ observations has $R^2 = 0.82$. Using $R^2_{adj} = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$, what is the adjusted R^2?

A. 0.85
B. 0.82
C. 0.76
D. 0.81

Question 5

A model with 3 predictors and $n = 50$ observations has $R^2 = 0.82$. Using $R^2_{adj} = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$, what is the adjusted R^2?

A. 0.85
B. 0.82
C. 0.76
D. 0.81

Answer: D
Plugging in: $R^2_{adj} = 1 - \frac{(1-0.82)(49)}{46} = 1 - \frac{0.18 \times 49}{46} = 1 - \frac{8.82}{46} = 1 - 0.1917 \approx 0.81$. The adjusted R^2 is slightly lower than R^2 because it penalizes for the number of predictors in the model.

Question 6

A researcher compares two multiple regression models: Model A ($R^2 = 0.88$, $R^2_{adj} = 0.75$, 10 predictors) and Model B ($R^2 = 0.83$, $R^2_{adj} = 0.80$, 3 predictors). Which model should be preferred and why?

A. Model B, because it has a higher adjusted R^2
B. Model A, because it has a higher R^2
C. Model B, because it has fewer predictors regardless of fit
D. Model A, because it uses more predictors

Question 6

A researcher compares two multiple regression models: Model A ($R^2 = 0.88$, $R^2_{adj} = 0.75$, 10 predictors) and Model B ($R^2 = 0.83$, $R^2_{adj} = 0.80$, 3 predictors). Which model should be preferred and why?

A. Model B, because it has a higher adjusted R^2
B. Model A, because it has a higher R^2
C. Model B, because it has fewer predictors regardless of fit
D. Model A, because it uses more predictors

Answer: A
Model B is preferred because its adjusted R^2 (0.80) exceeds that of Model A (0.75). The large gap between R^2 and R^2_{adj} in Model A indicates that many of its 10 predictors are not contributing meaningfully, and the model is likely overfitting the data.

Question 7

Which of the following residual plot patterns indicates that the linearity assumption of regression is violated?

A. A random scatter of points around zero
B. A U-shaped or curved pattern in the residuals
C. Residuals that are exactly zero for all observations
D. A constant band of residuals with uniform width

Question 7

Which of the following residual plot patterns indicates that the linearity assumption of regression is violated?

A. A random scatter of points around zero

B. A U-shaped or curved pattern in the residuals

C. Residuals that are exactly zero for all observations

D. A constant band of residuals with uniform width

Answer: B

A U-shaped or curved pattern in the residual plot indicates the true relationship between X and Y is non-linear, violating the linearity assumption. A random scatter around zero indicates the linearity assumption is satisfied, while a widening fan shape would indicate heteroscedasticity.

Question 8

A Durbin-Watson test statistic of $DW = 0.8$ suggests which problem with regression residuals?

A. Heteroscedasticity
B. Multicollinearity
C. Positive autocorrelation
D. Non-normality

Question 8

A Durbin-Watson test statistic of $DW = 0.8$ suggests which problem with regression residuals?

A. Heteroscedasticity
B. Multicollinearity
C. Positive autocorrelation
D. Non-normality

Answer: C
A Durbin-Watson statistic ranges from 0 to 4, with values near 2 indicating no autocorrelation. A value of 0.8 is well below 2, indicating positive autocorrelation among residuals. This commonly occurs in time series data where consecutive errors are correlated.

Question 9

A regression model shows a fan-shaped residual plot where variance increases with fitted values. Which remedial measure is most appropriate?

A. Add more predictor variables
B. Remove all outliers from the dataset
C. Increase the sample size
D. Apply a log transformation to the response variable

Question 9

A regression model shows a fan-shaped residual plot where variance increases with fitted values. Which remedial measure is most appropriate?

A. Add more predictor variables

B. Remove all outliers from the dataset

C. Increase the sample size

D. Apply a log transformation to the response variable

Answer: D

A fan-shaped residual pattern indicates heteroscedasticity (non-constant variance). A log transformation of Y is a standard remedy that stabilizes variance. Alternative approaches include weighted least squares or using robust standard errors.

Question 10

A predictor X_3 has a Variance Inflation Factor of $VIF = 8.5$. What does this indicate and what proportion of X_3's variance is explained by the other predictors?

A. Moderate multicollinearity; $R_3^2 \approx 0.88$
B. No concern; $R_3^2 = 0.12$
C. Severe multicollinearity; $R_3^2 \approx 0.92$
D. Moderate multicollinearity; $R_3^2 \approx 0.85$

Question 10

A predictor X_3 has a Variance Inflation Factor of $VIF = 8.5$. What does this indicate and what proportion of X_3's variance is explained by the other predictors?

A. Moderate multicollinearity; $R_3^2 \approx 0.88$
B. No concern; $R_3^2 = 0.12$
C. Severe multicollinearity; $R_3^2 \approx 0.92$
D. Moderate multicollinearity; $R_3^2 \approx 0.85$

Answer: A
Since $VIF_j = 1/(1 - R_j^2)$, we get $R_3^2 = 1 - 1/8.5 = 1 - 0.1176 \approx 0.88$. A VIF between 5 and 10 indicates moderate multicollinearity, meaning about 88% of the variance in X_3 can be predicted from the other predictors in the model.

Question 11

In a regression with three predictors, X_1 has $VIF = 2.1$, X_2 has $VIF = 15.3$, and X_3 has $VIF = 14.8$. What is the best course of action?

A. Remove X_1 because it has the lowest VIF
B. No action needed because VIF values are always acceptable in multiple regression
C. Remove all three predictors and start over
D. Investigate the relationship between X_2 and X_3 and consider removing one of them

Question 11

In a regression with three predictors, X_1 has $VIF = 2.1$, X_2 has $VIF = 15.3$, and X_3 has $VIF = 14.8$. What is the best course of action?

A. Remove X_1 because it has the lowest VIF

B. No action needed because VIF values are always acceptable in multiple regression

C. Remove all three predictors and start over

D. Investigate the relationship between X_2 and X_3 and consider removing one of them

Answer: D

VIF values exceeding 10 for both X_2 and X_3 indicate serious multicollinearity between them. The best approach is to examine their correlation, understand their substantive meaning, and consider removing one or combining them. X_1 with $VIF = 2.1$ has no multicollinearity problem.

Question 12

In logistic regression, the model predicts log-odds. If the predicted log-odds for a patient is 1.386, what is the predicted probability of the event?

A. 1.39
B. 0.50
C. 0.80
D. 0.20

Question 12

In logistic regression, the model predicts log-odds. If the predicted log-odds for a patient is 1.386, what is the predicted probability of the event?

A. 1.39
B. 0.50
C. 0.80
D. 0.20

Answer: C
The probability is obtained by applying the logistic function: $p = e^{1.386}/(1 + e^{1.386}) = 4.0/5.0 = 0.80$. Logistic regression outputs log-odds, which must be transformed through the sigmoid function to get probabilities between 0 and 1.

Question 13

A logistic regression model for disease risk includes age with coefficient $\hat{\beta} = 0.04$. What is the odds ratio for a 10-year increase in age?

A. $e^{0.4} \approx 1.49$

B. $e^{0.04} \approx 1.04$

C. $0.04 \times 10 = 0.40$

D. $e^{10} \approx 22{,}026$

Question 13

A logistic regression model for disease risk includes age with coefficient $\hat{\beta} = 0.04$. What is the odds ratio for a 10-year increase in age?

A. $e^{0.4} \approx 1.49$

B. $e^{0.04} \approx 1.04$

C. $0.04 \times 10 = 0.40$

D. $e^{10} \approx 22{,}026$

Answer: A

For a 10-year increase, the log-odds change is $0.04 \times 10 = 0.4$, so the odds ratio is $e^{0.4} \approx 1.49$. This means the odds of disease increase by approximately 49% for every 10-year increase in age. The single-unit OR would be $e^{0.04} \approx 1.04$.

Question 14

A logistic regression classifier has AUC = 0.92 on the test set. A second model has AUC = 0.51. Which statement is correct?

A. Both models perform well because AUC ¿ 0.50

B. The first model discriminates well; the second performs no better than random guessing

C. The second model is better because its AUC is closer to 0.50

D. AUC cannot be compared across different models

Question 14

A logistic regression classifier has AUC = 0.92 on the test set. A second model has AUC = 0.51. Which statement is correct?

A. Both models perform well because AUC ¿ 0.50

B. The first model discriminates well; the second performs no better than random guessing

C. The second model is better because its AUC is closer to 0.50

D. AUC cannot be compared across different models

Answer: B

AUC = 0.5 corresponds to random guessing (no discriminative ability), while AUC = 1.0 indicates perfect discrimination. The first model (AUC = 0.92) has excellent discriminative power, while the second (AUC = 0.51) performs essentially at chance level and has no practical predictive value.

Question 15

A patient enrolled in a clinical trial is still alive when the study ends after 36 months. This patient's survival time is classified as:

A. Uncensored, because the patient survived the entire study
B. Left censored, because the event did not occur
C. Right censored, because the true survival time exceeds the observed follow-up
D. Interval censored, because the event falls within a time window

Question 15

A patient enrolled in a clinical trial is still alive when the study ends after 36 months. This patient's survival time is classified as:

A. Uncensored, because the patient survived the entire study

B. Left censored, because the event did not occur

C. Right censored, because the true survival time exceeds the observed follow-up

D. Interval censored, because the event falls within a time window

Answer: C

This is right censoring, the most common type in survival analysis. We know the patient survived at least 36 months, but the true survival time is unknown and exceeds the observation period. Right censoring occurs when the study ends or a patient is lost to follow-up before experiencing the event.

Question 16

Why is it incorrect to simply exclude censored observations from a survival analysis and compute the mean survival time from uncensored observations only?

A. It would make the hazard function constant
B. It would increase the sample size artificially
C. It would violate the normality assumption
D. It would underestimate survival because censored patients survived at least as long as their censoring time

Question 16

Why is it incorrect to simply exclude censored observations from a survival analysis and compute the mean survival time from uncensored observations only?

A. It would make the hazard function constant

B. It would increase the sample size artificially

C. It would violate the normality assumption

D. It would underestimate survival because censored patients survived at least as long as their censoring time

Answer: D

Censored patients are known to have survived at least until their censoring time, so excluding them discards valuable information and biases survival estimates downward. The Kaplan-Meier method correctly handles censoring by including these patients in the risk set until the point of censoring, then removing them.

Question 17

In a Kaplan-Meier analysis, 10 patients are at risk at time $t = 7$ and 1 event occurs. At the previous event time, $\hat{S} = 0.80$. What is $\hat{S}(7)$?

A. 0.90
B. 0.72
C. 0.80
D. 0.08

Question 17

In a Kaplan-Meier analysis, 10 patients are at risk at time $t = 7$ and 1 event occurs. At the previous event time, $\hat{S} = 0.80$. What is $\hat{S}(7)$?

A. 0.90
B. 0.72
C. 0.80
D. 0.08

Answer: B

Using the product-limit formula: $\hat{S}(7) = \hat{S}(t_{prev}) \times (n - d)/n = 0.80 \times (10 - 1)/10 = 0.80 \times 0.90 = 0.72$. The survival estimate drops by the factor $(n_i - d_i)/n_i$ at each event time, and these factors multiply cumulatively.

Question 18

A log-rank test comparing two treatment groups yields $\chi^2 = 6.84$ with 1 degree of freedom ($\chi^2_{0.05,1} = 3.841$). The median survival times are 24 months (Group A) and 15 months (Group B). What is the correct interpretation?

A. There is no difference because median survival times overlap

B. Group B is preferred because it has a shorter median survival

C. Group A has significantly better survival ($p < 0.05$), with a 9-month median advantage

D. The test is invalid because the medians are not equal

Question 18

A log-rank test comparing two treatment groups yields $\chi^2 = 6.84$ with 1 degree of freedom ($\chi^2_{0.05,1} = 3.841$). The median survival times are 24 months (Group A) and 15 months (Group B). What is the correct interpretation?

A. There is no difference because median survival times overlap

B. Group B is preferred because it has a shorter median survival

C. Group A has significantly better survival ($p < 0.05$), with a 9-month median advantage

D. The test is invalid because the medians are not equal

Answer: C

Since $\chi^2 = 6.84 > 3.841$, we reject H_0 at $\alpha = 0.05$, confirming a statistically significant difference in survival between groups. Group A's median survival of 24 months is 9 months longer than Group B's 15 months, indicating significantly better survival outcomes for Group A.

Question 19

In a Cox proportional hazards model, $h(t|x) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2)$. What does $h_0(t)$ represent?

A. The survival function for all patients
B. The probability of the event at time t
C. The average hazard ratio across groups
D. The baseline hazard when all covariates equal zero

Question 19

In a Cox proportional hazards model, $h(t|x) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2)$. What does $h_0(t)$ represent?

A. The survival function for all patients

B. The probability of the event at time t

C. The average hazard ratio across groups

D. The baseline hazard when all covariates equal zero

Answer: D

$h_0(t)$ is the baseline hazard function, representing the hazard when all covariates are set to zero. The Cox model is semi-parametric because $h_0(t)$ is left unspecified (non-parametric part), while the covariate effects are modeled parametrically through the exponential term.

Question 20

A Cox model estimates a hazard ratio of $HR = 2.3$ (95% CI: 1.4–3.8) for a binary treatment variable. Which interpretation is correct?

A. The treatment group has 2.3 times the instantaneous risk of the event at any given time
B. The treatment group survives 2.3 times longer on average
C. The survival probability is 2.3 times higher in the treatment group
D. The median survival in the treatment group is 2.3 months

Question 20

A Cox model estimates a hazard ratio of $HR = 2.3$ (95% CI: 1.4–3.8) for a binary treatment variable. Which interpretation is correct?

A. The treatment group has 2.3 times the instantaneous risk of the event at any given time

B. The treatment group survives 2.3 times longer on average

C. The survival probability is 2.3 times higher in the treatment group

D. The median survival in the treatment group is 2.3 months

Answer: A

A hazard ratio of 2.3 means the treatment group has 2.3 times the instantaneous risk (hazard) of the event compared to the reference group at any point in time. The 95% CI of 1.4–3.8 excludes 1.0, confirming this is statistically significant. Note that a higher hazard ratio corresponds to worse survival, not better.