**\*\* CODE IN THE SAS FILE \*\***

**3) (i) Fit the bivariate regression model of Claim on Age. Obtain:**

**- a scatterplot of Claim against Age with the fitted regression line superimposed**

**- a plot of studentised residuals against fitted values**

**Critically discuss these two plots and draw appropriate conclusions regarding the adequacy**

**of the systematic component of the fitted model. (13 marks)**

**🡪 (a) Fitting the Bivariate Regression Model (Claim on Age)**

**(b) Scatterplot of Claim vs. Age with Fitted Regression Line:**

a) Claim and Age are positively correlated with each other as shown by the scatter plot.

An increase in Age also leads to an increase in average claim amount.

b) The fitted regression line (shown by the red line) conforms closely to the data points, indicating a strong fit.

c) There is no obvious curvature, which means a simple linear model would be appropriate.

**(c) Plot of Studentized Residuals vs. Fitted Values:**

a) The residuals appear to be randomly scattered around zero, which is a good sign.

b) However, there are some residuals that are larger in magnitude, suggesting that the variance might not be constant (potential heteroscedasticity).

c) No strong pattern (e.g., funnel shape) is present, meaning that no major violations of linear regression assumptions are evident.

Model Adequacy of Regression:

i) Good Fit:

- The R-square value (0.9867) is very large; the 98.67% variations in Claim can be explained by Age.

- The F-statistic (2148.48, p<0.0001) validates that Age is a significant predictor of Claim.

ii) Intercept and Slope:

- The intercept (-374.86, p<0.0001) is negative but cannot be interpreted as Age kicks in only from 50.

- The coefficient for Age (15.41, p<0.0001) implies that with every subsequent year of age, the average claim increase by approximately £15.41.

iii) Residual Analysis:

- Residuals show no dissimilar patterns and confirm the linear model is appropriate.

- Some larger studentized residuals indicate possible outliers or slight heteroscedasticity, but as a whole, the residual distribution does not display any serious violations.

**Conclusion & Recommendations**

- Linear regression model fits the data well so that almost all of the variance is explained.

- No increased evidence of non-linearity; perhaps some mild heteroscedasticity.

**Next Steps:**

- Check different transformations (such as using LnClaim instead of Claim) to see if they improve the model.

- Look at outlier diagnostics (Cook's distance, leverage) to ensure that no influential points affect the results.

**3) (ii) Now fit the bivariate regression model of LnClaim on Age. Obtain:**

**• A scatterplot of LnClaim against Age with the fitted regression line superimposed**

**• A plot of studentised residuals from this regression against the corresponding fitted**

**values**

**• A histogram of the studentised residuals**

**• A normal probability plot of the studentised residuals**

**Carefully explaining your methodology, investigate the adequacy of the new fitted**

**regression model. Which model would you recommend for these data and why? (12 marks)**

🡪 **Investigation of the Adequacy of the Fitted Regression Model:**

The fitted regression model evaluates the relationship between LnClaim (log-transformed claim values) and Age using simple linear regression.

The adequacy of the model is assessed through regression diagnostics, including residual analysis and normality checks.

**Methodology:**

1. Model Fitting:

- The regression equation takes the form: LnClaim=ß0+ß1×Age+?

- The scatterplot of LnClaim vs. Age shows a positive linear trend, suggesting that as Age increases, LnClaim also increases.

- The log transformation of Claim was likely performed to normalize the distribution and stabilize variance.

2. Residual Analysis:

- Studentized residuals vs. fitted values plot (homoscedasticity check):

- The points are randomly scattered around zero without a funnel pattern, confirming constant variance (homoscedasticity).

- There are no clear patterns, suggesting the linear assumption holds.

- Histogram of studentized residuals (normality check):

- The residuals follow a bell-shaped distribution.

- The slight right skew is present but not extreme.

- The standard deviation is close to 1, confirming appropriate studentization.

- Normal probability plot (Q-Q plot):

- The points align well with the expected normal line, except for minor deviations at the tails.

- This indicates approximate normality of residuals.

3. Goodness-of-Fit Tests for Normality

- Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling tests suggest no significant deviation from normality (p-values > 0.05).

- Mean of residuals is close to zero, indicating no systematic bias.

**Evaluation of Model Adequacy:**

Strengths:

a) Linear Relationship Holds – The scatterplot confirms a strong linear trend between Age and LnClaim.

b) Homoscedasticity Verified – The residuals show constant variance across fitted values.

c) Residuals Approximate Normality – Statistical tests and visualizations support the normality assumption.

d) No Major Outliers – Most studentized residuals are within ±2, with no extreme deviations.

**Limitations:**

a) Slight Positive Skew – The histogram and Q-Q plot indicate a small deviation from normality.

b) Potential Influential Observations – If leverage and DFBETA values are examined, they might highlight specific influential cases.

**Model Recommendation:**

- The LnClaim ~ Age model is adequate and should be preferred over a direct Claim ~ Age model because:

1. Better Normality & Variance Stabilization – The log transformation improves the distribution of residuals.

2. No Heteroscedasticity Issues – The residual plot confirms constant variance.

3. Interpretability – The coefficients provide insight into the multiplicative effects of Age on Claim amounts.

**Alternative Considerations:**

- If further improvements are needed, we could:

1. Add polynomial terms (Age²) if nonlinearity is suspected.

2. Check interactions with other variables.

3. Consider other transformations if residual skewness is problematic.

**Final Verdict:**

The LnClaim ~ Age model is statistically valid and provides a better fit than using Claim directly.

It should be recommended for further analysis and decision-making.

**3) iii) Write down the fitted model obtained in (ii) above, commenting briefly on the value of the**

**estimated slope. Explain how the model can be used to estimate the mean value of a travel**

**insurance claim for an over-50's claimant from his/her age.**

**Obtain and print out the estimated mean values of LnClaim for each observation on the data**

**set, together with corresponing 95% confidence limits. Use this model to estimate the mean**

**value claimed for a 70-year old claimant. Give and interpret an associated 95% confidence**

**interval.**

**What other information about travel insurance claims by over-50's should the company also**

**investigate in order to be able to complete its claim costings? (10 marks)**

**🡪 (a) Fitted Model: LnClaim=ß0+ß1×Age+?**

Were,

- ß0 (Intercept) represents the estimated LnClaim when Age = 0 (not practically meaningful in this case).

- ß1 (Slope) represents the change in LnClaim per unit increase in Age.

Interpreting the Slope (??1?):

- If ß1>0 - As Age increases, LnClaim increases (positive relationship).

- If ß1<0 - As Age increases, LnClaim decreases (negative relationship).

- If ß1=˜0 - No relationship between Age and LnClaim.

**(b) Using the Model to Estimate the Mean Value of a Travel Insurance Claim for Over-50s:**

- The regression model allows estimation of mean claim amounts for any age over 50 by plugging the age into the equation.

- The claim values are in logarithmic form (LnClaim), so to obtain the actual mean claim, we need to exponentiate the prediction:

Estimated Claim = eLnClaim

- This helps the company predict expected claim amounts for different ages over 50.

**(c) Obtain Estimated Mean Values of LnClaim for Each Observation with 95% Confidence Limits:**

We now compute:

- Predicted values of LnClaim

- 95% Confidence Limits for each prediction

Summary:

- Age is a strong predictor of LnClaim (R² = 99%), meaning claim amounts increase with age.

- Each additional year in age increases LnClaim by ~2.22%.

- Predicted LnClaim values and 95% confidence intervals show that for each age, we are 95% confident about the estimated mean claim amount.

- Example: At Age 60, the predicted LnClaim is 6.3153, with a confidence range of (6.3039, 6.3267).

- The model is highly accurate, with narrow confidence intervals, confirming that older individuals tend to have higher claim amounts.

**(d) Estimate the Mean Value for a 70-Year-Old Claimant:**

To estimate the mean claim amount for a 70-year-old, we:

- Add Age = 70 to the dataset.

- Use the model to predict LnClaim and its 95% confidence interval.

- Convert LnClaim back to the original scale using exponentiation

**(e) Interpreting the 95% CI FOR AGE=70:**

- In this case, we can say that we are 95% confident that the true mean of `LnClaim` for individuals aged 70 years falls in this interval.

This means that if we were to take many samples and compute this interval, about 95% of the resulting intervals would actually contain the true mean value of `LnClaim` for 70-year-olds.

- The narrow width of the CI indicates a high precision for the estimate, especially with the very high R² value of 0.99, which means that `Age` can explain so much of the variation of `LnClaim` in such a case!

**(f) Other Factors the Company Should Investigate:**

Although age is a reasonably strong predictor of `LnClaim`, the company should seek other pertinent factors for improving claim prediction and risk assessment.

Some of the important areas requiring investigation are as follows:

1. Demographic Factors

- Gender: Analyze the difference in claims between males and females.

- Occupation: Some professions may have lesser and higher risks.

- Income Level: The higher-income segment may pursue a higher policy cover.

2. Health-Related Factors

- Pre-existing Medical Conditions: Chronic diseases may increase claim amount.

- Smoking and Drinking Level: These factors could aggravate chances of a claim.

- BMI (Body Mass Index): Increased claims could arise because of ramifications associated with health arising from obesity.

3. Travel Behavior

- Travel Destination: Some destinations might present higher risks (e.g., political instability and natural disasters).

- Length of Travel: Longer trips might present claims with higher chances.

- Reason of Travel: Different levels of risks may apply to business versus leisure travel.

4. Policy-Specific Factors

- Coverage Type: Higher cover might mean higher claims.

- Claim History: Frequent claimants may stand at a greater risk.

- Deductibles and Premiums: Relation between the cost of policy and claims.

5. Economic & Environmental Factors

- Inflation & Medical Costs: Healthcare costs might affect claims.

- Climate & Weather Conditions: Bad weather may hinder travel safety.