1. Upload the "mortality\_data\_ver2.csv" data set.

2. Build the plot to look at the relationship between numerical variables.

What will be the dependent variable (outcome), what will be the independent variable (predictor)?

3. Perform multiple linear regression analysis with three independent variables (predictors).

Draw the best-fit regression line between the numerical variables.

4. Check the main assumptions of the model, use the four main plots for checking:

Plot 1.Linearity of the data, independence of residuals

Plot 2.Normality of residuals using Q-Q plot

Plot 3.Constant variance of residuals

Plot 4. No influential outliers

5. The assumption "Normality of residuals" is already checked, normally distributed.

6. After checking all the assumptions, what conclusion can you make? Are they met?

Yes.

7. Obtain parameters of the regression (α, β1, β2, β3); check significance. Fill it in the check list.

8. Obtain criteria for the model evaluation (Adjusted R-squared, RSE, the 95% confidence intervals). Fill it in the check list.

9. Fill up the check list.

|  |  |
| --- | --- |
| Check list | **Model** |
| **Assumptions after Linear regression:** |  |
| Plot 1: Linearity of the data, independence of residuals |  |
| Plot 2: Normality of residuals  +histogram + normality tests | Met |
| Zero mean of residuals | Met |
| Plot 3: Constant variance of residuals |  |
| Plot 4: No influential outliers |  |
| **Results interpretation and model evaluation:** |  |
| Parameters of the regression:  - intercept (α)  - β1, β2, β3 |  |
| Significance of β1, β2, β3 and the model |  |
| Build the multiple regression formula |  |
| Criteria for the model evaluation:  Adjusted R2; RSE; 95% CI |  |
| Calculate mortality of a 70 y.o. smoking man. |  |
| Calculate mortality of 10 y.o. non-smoking girl. |  |
| Calculate mortality of 30 y.o. non-smoking man. |  |
| Calculate mortality of 17 y.o. smoking girl. |  |

|  |  |
| --- | --- |
| Check list | **Model** |
| **Assumptions after Linear regression:** |  |
| Plot 1: Linearity of the data, independence of residuals | Met |
| Plot 2: Normality of residuals  +histogram + normality tests | Met |
| Zero mean of residuals | Met |
| Plot 3: Constant variance of residuals | Met |
| Plot 4: No influential outliers | Met |
| **Results interpretation and model evaluation:** |  |
| Parameters of the regression:  - intercept (α)  - β1, β2, β3 | α= 50.21, β1= 0.20, β2= 9.85, β3= 4.78 |
| Significance of β1, β2, β3 and the model | All p-values <0.001 |
| Build the multiple regression formula | Y(mortality)=50.21+0.20\*Age+9.85\*Smoker+4.78\*Gender |
| Criteria for the model evaluation:  Adjusted R2; RSE; 95% CI | R2adj.= 0.63, RSE=5.01,  95% CI: β1 [0.18; 0.22], β2[9.23; 10.47], β3[4.16; 5.40] |
| Calculate mortality of a 70 y.o. smoking man. | Y(mortality)=50.21+0.20\*70+9.85\*1+4.78\*1=78.84 |
| Calculate mortality of 10 y.o. non-smoking girl. | Y(mortality)=50.21+0.20\*10+9.85\*0+4.78\*0=52.21 |
| Calculate mortality of 30 y.o. non-smoking man. | Y(mortality)=50.21+0.20\*30+9.85\*0+4.78\*1=60.99 |
| Calculate mortality of 17 y.o. smoking girl. | Y(mortality)=50.21+0.20\*17+9.85\*1+4.78\*0=63.46 |