1. Upload the "salary\_data.csv" data set.

2. Build the plot to look at the relationship between the variables. What will be the dependent variable (outcome), what will be the independent variable (predictor)?

3. Perform linear regression analysis (fit a simple linear regression model between the variables). Draw the best-fit regression line.

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. Check the assumption "Normality of residuals" using histogram and normality tests; and "Zero mean of residuals". Don't forget to look at the Q-Q plot from the previous question.

6. Obtain parameters of the regression line (the intercept, the slope of the line); check the significance. Fill up the check list.

7. Obtain criteria for the model evaluation (Adjusted R-squared, RSE, AIC, the 95% confidence intervals). Fill up the check list.

8. After checking all the assumptions, what conclusion can you make?

9. Take away the outlier (number 5 on the previous plots) that has a high influence on the regression line.

To identify the outlier, first, look at histograms of the variables.

10. Delete the outlier using the tidyverse package.

11. Now, when you have data without the outlier, fit an adjusted simple linear regression model (repeat steps 2-7).

12. Fill up the check list, compare the models, choose a better model and draw your conclusions.

|  |  |  |
| --- | --- | --- |
| Check list | **Model\_version\_1** | **Model\_version\_2** |
| **Assumptions after Linear regression:** | | |
| Plot 1: Linearity of the data, independence of residuals |  |  |
| Plot 2: Normality of residuals  +histogram + normality tests |  |  |
| Zero mean of residuals |  |  |
| Plot 3: Constant variance of residuals |  |  |
| Plot 4: No influential outliers |  |  |
| **Results interpretation and model evaluation:** | | |
| Parameters of the regression:  - intercept (α)  - slope of the line (β) |  |  |
| Significance of β and the model |  |  |
| Criteria for the model evaluation:  Adjusted R2; RSE; 95% CI; AIC |  |  |
| Conclusion based on the chosen model: | The assumptions are \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_; the model and the independent variable (\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_) are \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (p\_\_\_\_\_\_\_\_\_\_\_\_).  The \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ variable explains \_\_\_\_\_\_\_\_% of the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ variability, RSE equals \_\_\_\_\_\_\_\_\_\_\_\_\_\_.  The estimate of the β-coefficient equals \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  (95% CI [\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_]), the intercept α equals \_\_\_\_\_\_\_.  Y(\_\_\_\_\_\_\_\_\_\_\_\_)=\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_(for each one-unit shift of \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_increases by \_\_\_\_\_\_\_\_\_\_\_). | |

|  |  |  |
| --- | --- | --- |
| Check list | **Model\_version\_1** | **Model\_version\_2** |
| **Assumptions after Linear regression:** | | |
| Plot 1: Linearity of the data, independence of residuals | met | met |
| Plot 2: Normality of residuals  +histogram + normality tests | not met | met |
| Zero mean of residuals | met | met |
| Plot 3: Constant variance of residuals | met | met |
| Plot 4: No influential outliers | not met | met |
| **Results interpretation and model evaluation:** | | |
| Parameters of the regression:  - intercept (α)  - slope of the line (β) | α= -28.63, β=0.62  Y(productivity)=  -28.63+0.62\*X(salary) | α= -41.75, β=0.71  Y(productivity)=  -41.75+0.71\*X(salary) |
| Significance of β and the model | p<0.001 | p<0.001 |
| Criteria for the model evaluation:  Adjusted R2; RSE; 95% CI; AIC | R2adj.=33%, RSE=7.95 (thousand dollars per year), 95%CI [0.44;0.79], AIC=702.5 | R2adj.=51%, RSE=6.22 (thousand dollars per year), 95%CI [0.57;0.85], AIC=646.7 |
| Conclusion based on the chosen model: | The assumptions are met the model and the independent variable (salary) are significant (p<0.001).  The salary variable explains 51% of the productivity variability, RSE equals 6.22 thousand dollars per year.  The estimate of the β-coefficient equals 0.71 (95% CI [0.57;0.85]), the intercept α equals -41.75.  Y(productivity)= -41.75+0.71\*X(salary) (for each one-unit shift of salary productivity increases by 0.71). | |