Linear Regression Assumptions and Statistical Inference

Contents

Questions to discuss

1. What is simple and multiple linear regression?
2. What are the assumptions of linear regression?
3. What are the statistical inferences we can make from a linear regression model?

What is simple and multiple linear regression?

- Linear regression is a way to identify a relationship between the independent variable(s) and dependent variable
- We can use these relationships to predict values for one variable for the given value(s) of the other variable(s)
- The variable which is used in prediction is termed as independent/explanatory/regressor, and the predicted variable is termed as dependent/target/response variable.
- In the case of linear regression with a single explanatory variable, the linear combination can be expressed as

  response = intercept + constant * explanatory variable

- Multiple linear regression is just the extension of the concept of simple linear regression with one variable
- It can be represented by

  target = intercept + constant1*feature1 + constant2*feature2 + constant3*feature3 + ...

- The model aims to find the constants and intercept such that the hyperplane is the best fit

What are the assumptions of linear regression?

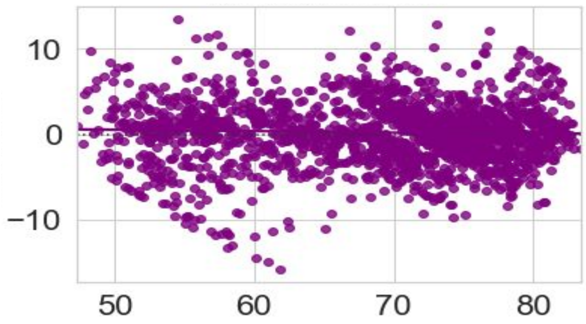
Assumption	How to test	How to fix
No multicollinearity in independent variables	Heatmaps of correlations or VIF (Variance inflation factor)	Remove correlated variables
There should be a linear relationship between dependent and independent variables	Plot residuals vs. fitted values and check the plot	Transform variables that appear non-linear (log, square root, etc.)
The residuals should be independent of each other	Plot residuals vs. fitted values and check the plot	Transform variables (log, square root, etc.)
Residuals must be normally distributed	Plot residuals or use Q-Q plot	Non-linear transformation of the independent or dependent variable
No heteroscedasticity, i.e., residuals should have constant variance	Use statistical test (like goldfeldquandt test)	Non-linear transformation of the dependent variable or add other important variables

Testing Multicollinearity using VIF

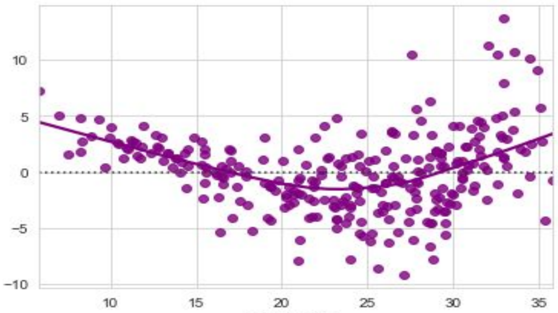
- Multicollinearity occurs when predictor variables in a regression model are correlated.
- When we have multicollinearity in the linear model, the coefficients that the model suggests are unreliable.
- We can detect or test for multicollinearity using the Variance Inflation Factor or VIF.
- Variance inflation factors measure the inflation in the variances of the regression parameter estimates due to collinearities that exist among the predictors.
    - If VIF is 1, then there is no correlation between the selected predictor and the remaining predictor variables, and hence, the variance of its coefficient is not inflated at all.
- General Rule of thumb:
    - 1 < VIF <= 5: There is low multicollinearity
    - 5 < VIF <= 10: There is moderate multicollinearity
    - VIF > 10: There is high multicollinearity

Testing Linearity and Independence using residuals vs fitted values plot

- Predictor variables must have a linear relation with the dependent variable.
- If the residuals are not independent, then the confidence intervals of the coefficient estimates will be narrower and make us incorrectly conclude a parameter to be statistically significant.
- We can check for linearity and independence by checking a plot of fitted values vs residuals.
  - If they don't follow any pattern, then we say the model is linear and residuals are independent.
  - Otherwise, the model is showing signs of non-linearity and residuals are not independent.



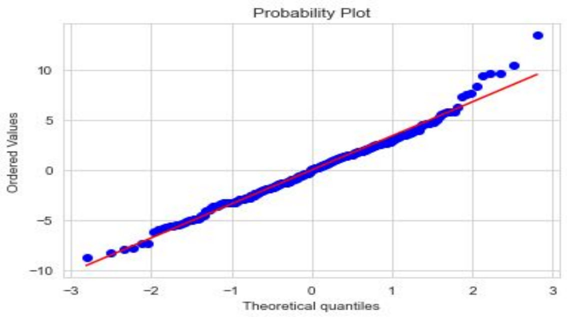
No pattern spotted



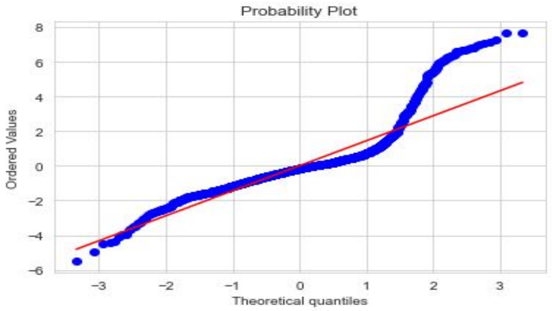
Some non-linearity spotted

Testing Normality using QQ plots

- If the error terms are not normally distributed, the confidence intervals of the coefficient estimates may become too wide or narrow.
- Non-normality suggests that there are a few unusual data points that must be studied closely to make a better model.
- The shape of the histogram of residuals can give an initial idea about the normality.
- We can also check normality via a Q-Q plot of residuals.
    - If the residuals follow a normal distribution, they will make a straight line plot, otherwise not.



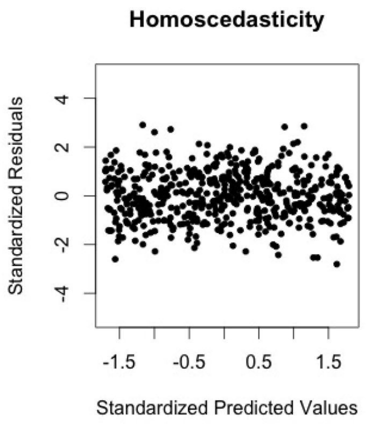
Close to normal

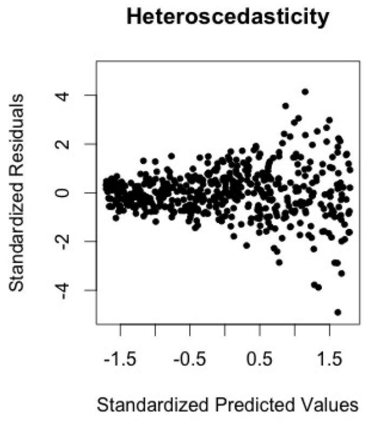


Not normal

Testing Homoscedasticity using goldfeldquandt test

- If the variance of the residuals is symmetrically distributed across the regression line, then the data is said to be homoscedastic. Else, they are heteroscedastic.
- Generally, non-constant variance arises in presence of outliers.
- The residual vs fitted values plot can be looked at to check for homoscedasticity.
  - In the case of heteroscedasticity, the residuals can form an arrow shape or any other non-symmetrical shape.
- The goldfeldquandt test can also be used.
  - If we get a p-value > 0.05 we can say that the residuals are homoscedastic. Otherwise, they are heteroscedastic.
  - Null hypothesis: Residuals are homoscedastic
  - Alternate hypothesis: Residuals have heteroscedasticity





What are the statistical inferences we can make from a linear regression model?

- **Confidence interval**: Give us the 95% confidence interval for the coefficient
    - The 95% confidence interval for Alcohol is [-0.146, -0.015].
- **p-values**: Give us an idea of whether a particular predictor variable has a statistically significant effect on the target variable or not. If p-value > 0,05, the variable is not statistically significant
    - The p-value for Alcohol is 0.017, indicating that it is statistically significant in predicting life expectancy

```
                            OLS Regression Results
==============================================================================
Dep. Variable:        Life expectancy   R-squared:                       0.843
Model:                            OLS   Adj. R-squared:                  0.842
Method:                 Least Squares   F-statistic:                     575.6
Date:                Fri, 01 Oct 2021   Prob (F-statistic):               0.00
Time:                        13:13:06   Log-Likelihood:                -5636.5
No. Observations:                2049   AIC:                         1.131e+04
Df Residuals:                    2029   BIC:                         1.143e+04
Df Model:                          19
Covariance Type:            nonrobust
====================================================================================================
                                      coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------------
const                             -23.1832     39.635     -0.585      0.559    -100.913      54.547
Year                                0.0395      0.020      1.994      0.046       0.001       0.078
Adult Mortality                    -0.0162      0.001    -17.819      0.000      -0.018      -0.014
Alcohol                            -0.0801      0.033     -2.395      0.017      -0.146      -0.015
Percentage expenditure              0.0003   4.96e-05      6.128      0.000       0.000       0.000
Hepatitis B                        -0.0163      0.004     -3.821      0.000      -0.025      -0.008
BMI                                 0.0346      0.006      5.860      0.000       0.023       0.046
Under-five deaths                  -0.0023      0.001     -3.794      0.000      -0.004      -0.001
Polio                               0.0346      0.005      6.960      0.000       0.025       0.044
Diphtheria                          0.0344      0.005      6.688      0.000       0.024       0.045
HIV/AIDS                           -0.3813      0.020    -18.604      0.000      -0.422      -0.341
Thinness 5-9 years                 -0.0777      0.029     -2.669      0.008      -0.135      -0.021
Income composition of resources     4.5468      0.721      6.302      0.000       3.132       5.962
Schooling                           0.6184      0.048     12.789      0.000       0.524       0.713
Status_Developing                  -2.6234      0.353     -7.442      0.000      -3.315      -1.932
Continent_Asia                      4.7406      0.281     16.862      0.000       4.189       5.292
Continent_Europe                    4.3902      0.411     10.694      0.000       3.585       5.195
Continent_North America             6.2753      0.360     17.417      0.000       5.569       6.982
Continent_Oceania                   2.7757      0.456      6.089      0.000       1.882       3.670
Continent_South America             4.4261      0.440     10.062      0.000       3.563       5.289
==============================================================================
Omnibus:                       80.001   Durbin-Watson:                   2.014
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              214.140
Skew:                          -0.138   Prob(JB):                     3.16e-47
Kurtosis:                       4.559   Cond. No.                     1.14e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.14e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

greatlearning
Power Ahead

Happy Learning !