Regression Assumptions

Thursday 25 April 2024 1:51 PM

→ Linear regression relies on several assumptions to ensure the validity & reliability of the estimates & inferences.

1. Linearity :

→ There exists a linear relationship b/w the independent variables & dependent variable.

→ If not linear, model won't be a good fit for data & parameter estimates will be meaningless.

In case of violation:

→ Bias - The inability of a model to truly capture the relationship in the training data

→ Reduced predictive power

→ Invalid hypothesis tests & confidence intervals

How to check :

→ Scatter plots / pair plots

→ Residual plots

- $(y-\hat{y})$ vs $\hat{y}$
- $(y-\hat{y})$ vs x

Points should be symmetrically/Randomly distributed around the line. With no discernible pattern

→ Polynomial terms : Add polynomial terms to your model & compare the model fit with the original linear model. If the new model with additional terms significantly improves the fit, it may suggest that the linearity assumption is not met.

→ Likelihood (LR) test

What to do ?

→ Transformations

→ Polynomial Regression

→ Piecewise Regression

→ Non-parametric or semi-parametric methods

2. Normality of errors :

→ The error terms (residuals) are assumed to follow a normal distribution with a mean of zero & a constant variance.

In case of violation :

→ Inaccurate hypothesis tests :

$$\text{F-statistic} = \frac{MSR}{MSE} = \frac{\sum(\hat{y}-\bar{y})^2/k}{\sum(y-\hat{y})^2/(n-k-1)}$$

follows F-distribution

$$\frac{\chi_1^2/df_1}{\chi_2^2/df_1}$$

(N^2)

→ Invalid confidence intervals :

$$\text{t-value}_{\beta_k} = \frac{\beta_k - 0}{\sqrt{\dfrac{\sum(y-\hat{y})^2}{(n-k-1)\sum(x_k-\bar{x}_k)^2}}}$$

$$t\text{-value}_{\beta_k} = \frac{\hat{\beta}_k - 0}{\sqrt{\dfrac{\sum (y-\hat{y})^2}{(n-k-1)\sum (x_k - \bar{x}_k)^2}}}$$

x

$e_i \sim N(0, \sigma^2)$ X

→ Model performance : reduced predictive accuracy

How to Check :

- ✓ Histogram
- ✓ Residual plot
- ✓ Q-Q plot → bow- & S-shaped X
- ✓ Shapiro-Wilk test
- Komolgorov-Smirnov test
- Anderson-Darling test
- ✓ Omnibus test
- ✓ Jarque-Bera test

What to do?

→ Model selection techniques

→ Robust Regression

→ Transformations may help

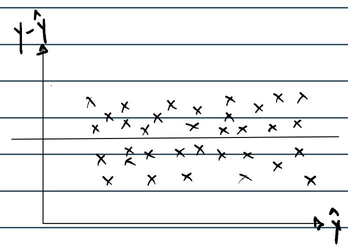
→ Non-parametric or semi-parametric methods

→ Use bootstrapping

Note: Remember that normality of residuals assumption is not always critical for linear regression, especially when the sample size is large, due to CLT.

3. Constant Variance (homoscedastic)

→ The spread of the error terms should be constant across all levels of the independent variables. If the spread of the residuals changes systematically, it leads to hetroscedasticity, which can affect the efficiency of the estimates.



→ plot the residuals vs independent variables to look for consistency there as well.

→ sns.regplot(y= residuals, x=y_pred)

→ from yellowbrick.regressor import ResidualsPlot, residuals_plot
res_plot = residuals_plot(LinearRegression(), X_train, y_train, X_test, y_test)

Issue?

→ While the parameters estimates are still unbiased under hetroskedasticity, $\underbrace{\text{standard errors}}$ in the coefficients cannot be relied upon.

→ standard deviation of sample parameter is standard error.

$\boxed{\text{sns.lmplot}}$

- inefficient estimates
- invalid confidence intervals
- inaccurate hypothesis tests

What happens if the residual analysis reveals hetroscedasticity?

→ Rebuild the model with different independent variable(s)
→ Perform transformations on non-linear data e.g. $\log(Y)$ or $\sqrt{Y}$
→ Fit a non-linear regression model but don't OVERFIT.
→ Weighted least squares

statistical tests for residuals

- Breusch-Pagan Test → H_0: Homoscedasticity is present; H_a: Hetroscedasticity exists
- White test
- NCV test

4. No multicollinearity

→ occurs where the independent variables are themselves related, (highly correlated) making it difficult to isolate the individual effects of each variable on the dependent variable.

→ Inference / interpretability is affected e.g. effect of β_k on y while keeping other variables constant → not possible with multicollinearity.
[Explainable AI] → Prediction is not effected as much

→ What is the purpose of the model — inference / prediction   overall model fit not affected

→ if multicollinearity exists
- Difficulty indentifying the most important predictors ①
- Inflated standard errors ② → decreases the statistical power & can make it challenging to determine the true relationship b/w the independent & dependent variables.
- Unstable & Unreliable estimates ③ → coefficients become sensitive to small changes in the data, making it difficult to interpret the results accurately. Coefficients are still unbiased.

→ Perfect multicollinearity occurs when one independent variable in a multiple regression model is an exact linear combination of one or more other independent variables.

i.e. $X_1 = \beta_0 + \beta_1 X_2 + \text{error}$ ✗

$\boxed{X_1 = \beta_0 + \beta_1 X_2}$ ✓ → perfect multicollinearity

e.g.

rainfall	temprature	sales
2	20	8
4	40	16

OLS

$\boxed{\beta = (X^TX)^{-1}X^Ty}$ → design matrix

$$\text{Sales} = \beta_0 + \beta_1 \text{rainfall} + \beta_2 \text{temprature} + \varepsilon_i$$

$$X = \begin{bmatrix} 1 & 2 & 20 \\ 1 & 4 & 40 \end{bmatrix} \quad Y = \begin{bmatrix} 8 \\ 16 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 1 \\ 2 & 4 \\ 20 & 40 \end{bmatrix}_{3\times2} \begin{bmatrix} 1 & 2 & 20 \\ 1 & 4 & 40 \end{bmatrix}_{2\times3} = \begin{bmatrix} 2 & 6 & 60 \\ 6 & 20 & 200 \\ 60 & 200 & 2000 \end{bmatrix}$$

$$\text{Det} = 2\begin{vmatrix} 20 & 200 \\ 200 & 2000 \end{vmatrix} - 6\begin{vmatrix} 6 & 200 \\ 60 & 2000 \end{vmatrix} + 60\begin{vmatrix} 6 & 20 \\ 60 & 200 \end{vmatrix}$$

$= 2(0) - 6(0) + 60(0)$

$= 0$ → Singular matrix, inverse cannot calculated
i.e. β matrix cannot be found

$$SE(\beta) = \sqrt{\text{diag}(\sigma^2 (X^TX)^{-1})} \uparrow$$ → covariance matrix

$$(X^TX)^{-1} = \frac{1}{\det(X^TX)} \text{adj}(X^TX) \uparrow$$

if we have perfect multicollinearity $\det(X^TX) = 0$, mostly that wont be the case. If we have strong multicollinearity then $\det(X^TX)$ will be very small

→ Types of multicollinearity
- structural (featuring engineering decisions) e.g. one-hot-encoding
- Data-driven (natural)

→ How to detect?

① Correlation b/w independent variables

② VIF (Variance inflation factor)

e.g.

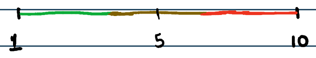
TV	NP	Radio	Sales

① $Radio = \beta_0^* + \beta_1^* TV + \beta_2^* NP + \varepsilon^*$

② $NP = \beta_0^* + \beta_1^* TV + \beta_2^* NP + \varepsilon^*$

③ $TV = \beta_0^* + \beta_1^* NP + \beta_2^* Radio + \varepsilon^*$

$\rightarrow R^2 \text{ score} \rightarrow VIF = \frac{1}{1-R^2} = 10$



③ Condition number

→ describes the ill conditioning of the (X^TX) matrix.

→ Typically, a condition number larger than 30 (or sometimes even larger than 10 or 20) is considered a warning sign of potential multicollinearity issues.

→ Note: A high condition number alone is not definitive proof of multicollinearity.

④ Tolerence $= 1-R^2$

$\boxed{T < 0.1}$

→ Solutions?

1) Collect more data
2) Remove one of the highly correlated features → Note: Beware of omitted variable bias! we will have variables outside the model which might be pulling the strings still of the remaining variables.
3) Combine correlated variables
4) Use partial least squares regression (PLS)

5. No Auto-correlation [Research] the error terms are also assumed to be independent

→ Durbin-Watson test