Key Concepts and Terminology

Linear Regression Modelling-Machine Learning Primer

Correlation: In linear regression and machine learning, correlation is a test used to determine the best predictor for the dependent variable.

Bivariate normality: Checking to see if both variables are normal.

Linear Regression model: The equation used to fit the data and help in predicting the outcome variable with high predictive precision being the target.

Line of Best fit: Mathematical predictor equation. Aka, the regression line. The line is only as good as the data.

Least-Squares Method: Used to determine the equation of the regression line by error minimization (using the lowest possible residual values).

$$\text{LSM} = \sum \text{Negative residuals} + \text{Positive residuals} = 0$$

Auto-correlation: Applicable in linear regression is the absence of independence of observations, as a result, absence of independence of residuals.

T-test: Tells whether the coefficients of the regression equation differ from 0 thus helps understand the significance of the magnitude of coefficients.

Anova: Anova in linear regressions produces an F-ratio that tells if the regression model is a statistically significant predictor of the outcome variable from the predictor variable.

Correlation Coefficient: Thought of as the measure of effect size in correlation.

$$Regression\ variance\ \% = \frac{\sum (y_{ip} - \bar{y})^2}{\sum (y_i - \bar{y})^2} * 100; y_{ip} = predicted\ FEV_L,$$
$$y_i = sample\ FEV_L\ value$$

$$Regression\ variance\ \% = \frac{\sum (y_{ip} - 2.637)^2}{\sum (y_i - 2.637)^2} * 100$$

$$Regression\ variance\ \% = \frac{366.1053}{489.8586} * 100 = 74.739\%$$

Key Concepts and Terminology

Regression Equation: slope = 0.121 and Y-intercept = 4.388

$$y = .121x - 4.388$$

F-ratio: A measure used to determine the statistical significance of the linear regression model ($p < 0.001$ is significant).

$$F - ratio = \frac{MS_M}{MS_R}, MS_M = SOS_{regression}, MS_R = SOS_{residuals}$$

Slope: The coefficient of the independent variable in a linear model.

$$m = r * \frac{S_y}{S_x}$$

Y-Intercept: It is the value of the dependent variable when there is no effect (0 effect) from the independent variable. The value is minimized by picking a good predictor. The variable r below is Pearson's regression coefficient.

$$Y - intercept = \bar{X}_y - r * \bar{X}_x$$

Models and Transformation Decisions

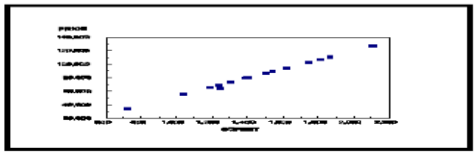

Figure 1

x

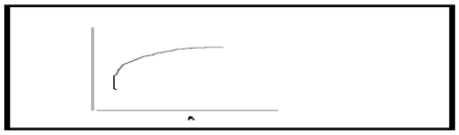

Figure 2

$\mathbf{Log(x)}$ or $\sqrt{x}$

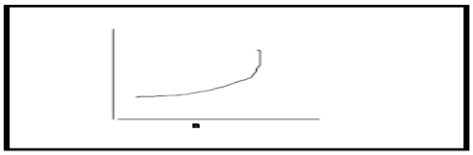

Figure 3

X^2 or $Exp(X)$

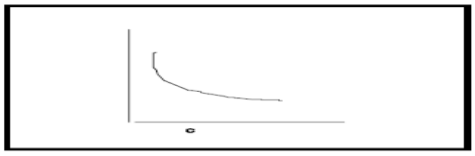

Figure 4

$\frac{1}{X}$ or $Exp(-X)$

Key Concepts and Terminology

Dummy variable: Used when coding categorical variables and rule is p-1 dummy variables where p = # of categories.

Data set size: $2 * \left(\frac{p}{n}\right)^{\frac{1}{2}} > 1$, $large\ data\ set$

Heteroscedasticity: Not identical distributions of error terms or residuals (error variances are NOT identical).

SPEC test (SAS): The null hypothesis posits independence of error terms and homoscedasticity (identical distribution of error terms). A p-value > 0.05 indicates acceptance of the NULL. (Prob>Chi-Sq of ..>0.05 is the desired outcome).

Durbin-Watson Test: A test for first order correlation of error terms. Value ranges from 0 to 4 with 2.0 being the sweet spot; and <2 = positive first order correlation VS >2 = negative first order correlation. The statistic is valid for large data sets.

Shapiro-Wilk: Null = normality hence normality of residuals is achieved with p > 0.01.

Multicollinearity: Correlation of X variables.

Variance Inflation Factor (VIF): Tests for multicollinearity. VIF > 10 indicates multicollinearity.

Outliers: observations that have high impact on the predictive model.

Cook's distance: A determinant of outliers. Cooks d > |2| should be reviewed.

Other outlier Measures (SAS): DFFITs > $2 * \left(\frac{p}{n}\right)^{\frac{1}{2}}$ for large data sets. Dfbetas > 1 (small to medium data sets) and > $\frac{2}{\sqrt{n}}$ (large data sets).

Weighted outliers: Assigning weights (ranking outliers) to outliers in trying to reduce their effect.

Order of Assumptions Check (Workflow): Most Severe to Least Severe (subjective)

1. Normality of residuals (S-W)
2. Homoscedasticity and autocorrelation (IID)- (Durbin Watson)
3. Multicollinearity (VIF)
4. Outliers(Cooks d, DFFIT, DFBetas)
5. Test Model Fit

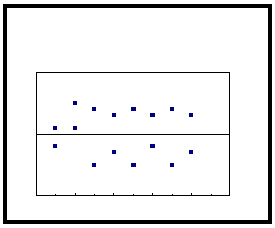

Figure 5

$$Linear$$

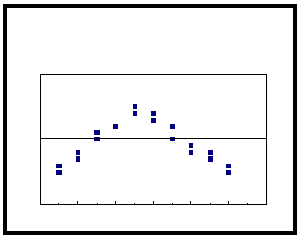

Figure 6

$$X \longrightarrow X^2$$

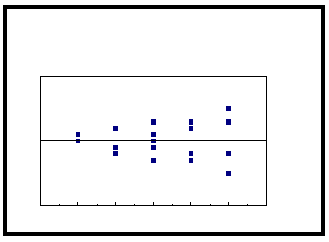

Figure 7

$$Y \longrightarrow Log(Y)$$

Key Concepts and Terminology

Interpreting Linear Regression:

Root MSE (Lower is better)

Type III SS p-value < 0.05 (indicates significance)

Source	DF	Type III SS	Mean Square	F Value	Pr > F
units_1	1	1476168.867	1476168.867	17059.0	<.0001
units_10	1	4290056.112	4290056.112	49576.9	<.0001
dummy	3	64.174	21.391	0.25	0.8634
unitssq	1	1270645.011	1270645.011	14683.9	<.0001

R-square and Adj R-square(>= 0.7)

Adjusted R-squared: Provides a penalty to the R-squared value if variables without strong correlation are added to the model (k is the # of predictors). Only calculated if there are several predictors that show poor correlation to the dependent variable.

$$Adj\ R^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}$$