Regression Analysis Cheat Sheet

Graduate Resource Center, University of New Mexico

1 Regression basics

Definitions and Terms

Variable: Measurable characteristics that varies (by groups, individuals or time)

Dependent/Outcome Variable (DV): Presumed effect in an analysis

Independent/Explanatory Variable (IV): The presumed cause in an analysis

Control Variable/Covariate: Variables that are not studied but included in the model/analysis

Parameter: Unknown population characteristics

Best Fitting Line: When plotting data, the most appropriate line showing the relationship between dependent and independent variables

Residual: Deviations from the fitted line (estimated value) to the observed values (data point)

Error: Difference between the observed value and the true value (often unobserved)

Regression Coefficient: Describes the relationship between a DV and IV

Understanding Regression

Can we explain how much, on average, DV changes because of IV?

Can we predict, on average, what DV values might be for a value of IV?

Can we determine if the amount of change in one variable is related to, on average, the amount of change in another?

Model Building

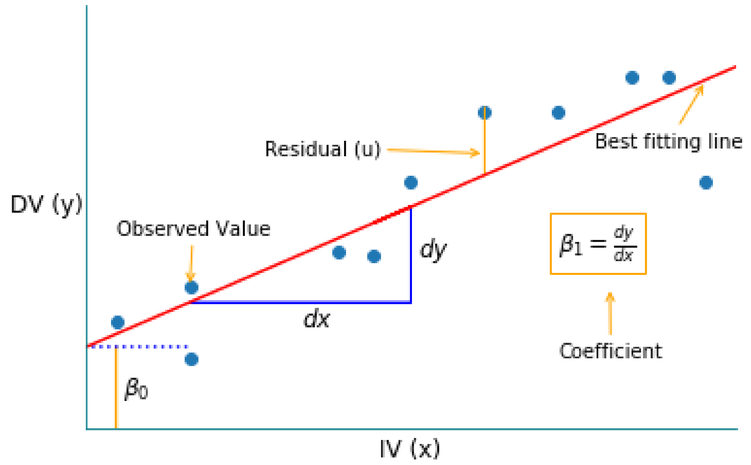
Model is a simplified representation of reality

When building a model, think:

1. What are the variables that interact?
2. How do those variables interact?
3. Apply underlying theories

2 Regression

Graphical Representation (univeriate linear model)



Mathematical Representation

1. Linear regression (DV is a continuous variable)

$$\overbrace{y_i}^{\text{DV}} = \overbrace{\beta_0}^{\text{Constant}} + \overbrace{\beta_0}^{\text{Coefficient}} \overbrace{x_i}^{\text{IV}} + \overbrace{\epsilon_i}^{\text{Error}}$$

2. Logistic Regression (DV is a binary variable)

$$\Pr(y_i = 1 \mid x_i) = \frac{1}{1+exp^{-(\beta_0+\beta_i x_i)}}$$

Steps to Regression Analysis

1. Choose a suitable model
2. Choose a suitable predictors
3. Select a technique for fitting the model
4. Fit the model
5. Assesses goodness of fit of the model
6. Interpret and Conclude - make causal inference, issue prediction

Need Help?

Contact Us

Graduate Resource Center
Mesa Vista Hall, Suite 1057
Phone: 505-277-1407
Email: unmgrc@unm.edu
Website: https://unmgrc.unm.edu/

3 Regression result

After Regression, what to look for

1. R-square (R^2): It shows the amount of variance of DV explained by IV (described in percent)

2. *p*-value of the model: It tests whether R^2 is different from 0. A value less than 0.05 shows statistically significant relationship between IV and DV

3. *p*-value of the coefficient: p-value of the hypothesis testing coefficient is different from 0 (H_0).

4. Coefficient (β_i): For each one-point increase in IV, the DV is expected to increases by β_i (or decreases if β_i is negative), *holding all the other independent variables constant*

Note: Coefficient of logistic regression is interpreted differently

5. Regression results also provides:

t-statistics: Coefficient divided by standard error
Standard error: Standard deviation of the coefficient
Confidence Interval of the coefficient - Usually 95%
Root mean squared error: Standard deviation of the regression

Check for Key Assumptions

1. Linearity: The relationship between IV and the mean of DV is linear.

2. Homoscedasticity: The variance of residual is the same for any of IV.

3. Normality: For any fixed value of IV, DV is normally distributed.

4. Multicollinearity: The independent variables are not not perfectly multicolinear (one IV should not be a linear function of another).

Note: Assumptions 1, 2 and 3 do not apply to logistic regression.

Selecting the "BEST" Model

Goal is to minimize the residual mean square (which maximizes R^2) - by comparing regression models

Use informtion criterion statistics -

Akaike's Information Criterion (AIC)
Bayesian Information Criterion (BIC)