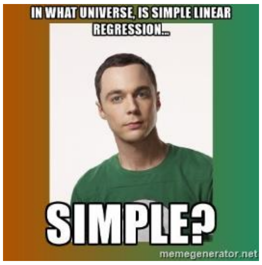
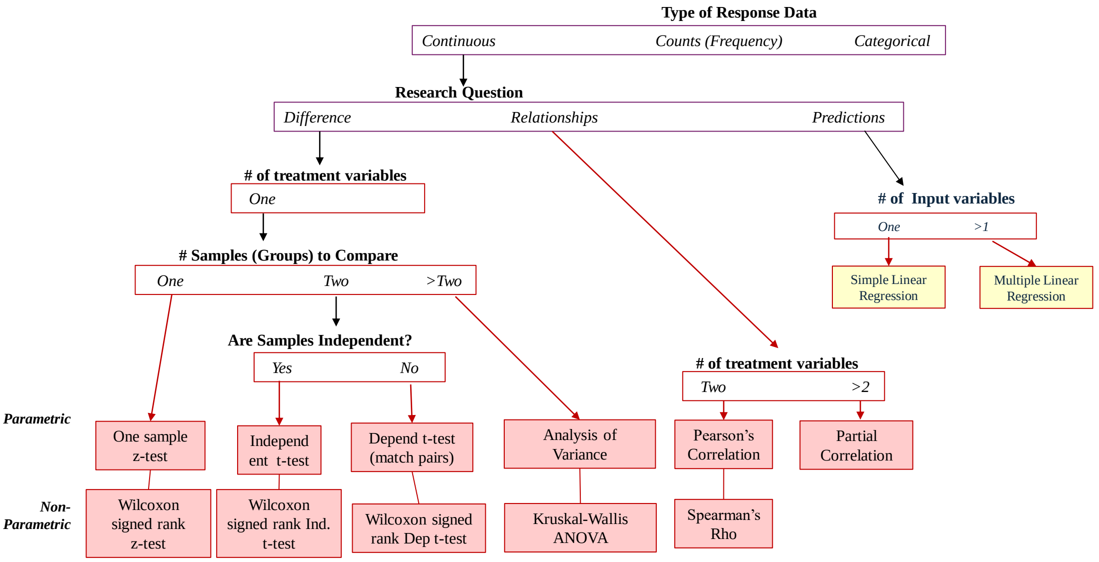
Modeling Using Linear Regressions



When would you use a linear regression?



Type of Response Data

Continuous | *Counts (Frequency)* | *Categorical*

Research Question

Difference | *Relationships* | *Predictions*

# of treatment variables

One

# Samples (Groups) to Compare

One | *Two* | *>Two*

Are Samples Independent?

Yes | *No*

# of treatment variables

Two | *>2*

# of Input variables

One | *>1*

Simple Linear Regression

Multiple Linear Regression

Parametric

One sample z-test

Independ ent t-test

Depend t-test (match pairs)

Analysis of Variance

Pearson's Correlation

Partial Correlation

Non-Parametric

Wilcoxon signed rank z-test

Wilcoxon signed rank Ind. t-test

Wilcoxon signed rank Dep t-test

Kruskal-Wallis ANOVA

Spearman's Rho

What is multiple linear regression (MLR)?

- **Multiple linear regression** is a generalization of simple linear regression, in the sense that this approach makes it possible to evaluate the linear relationships between a response variable (quantitative) and several explanatory variables (quantitative or qualitative).

When to use a MLR?

- How strong the relationship is between two or more **independent variables** and one **dependent variable** (e.g. how rainfall, temperature, and amount of fertilizer added affect crop growth)

Assumptions of multiple linear regression

- **Homogeneity of variance (homoscedasticity)**: the size of the error in our prediction doesn’t change significantly across the values of the independent variable.

- **Independence of observations**: the observations in the dataset were collected using statistically valid sampling methods, and there are no hidden relationships among variables.
  - In multiple linear regression, it is possible that some of the independent variables are actually correlated with one another, so it is important to check these before developing the regression model. If two independent variables are too highly correlated (r2 > ~0.6), then only one of them should be used in the regression model.

- The residual values are normally distributed. This can be checked by either using a normal probability plot and histogram, and Shapiro test.

- **Linearity**: the line of best fit through the data points is a straight line, rather than a curve or some sort of grouping factor.

5

Assumptions of multiple linear regression

- **Multicollinearity** meaning that the independent variables are not highly correlated with each other.

- Multicollinearity makes it difficult to identify which variables better explain the dependent variable.

- This assumption is verified by computing a matrix of Pearson’s bivariate correlations among all the independent variables. If there is no collinearity in the data, then all the values should be less than 0.8.

6

What is a Variation Inflation Factor?

- A variance inflation factor (VIF) provides a measure of multicollinearity among the independent variables in a multiple regression model.
- A variance inflation factor (*VIF*) quantifies how much the variance is inflated.

•VIF equal to 1 = variables are not correlated
•VIF between 1 and 5 = variables are moderately correlated
•VIF greater than 5 = variables are highly correlated

7

Multiple linear regression formula

$$y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \epsilon$$

- Y = the predicted value of the dependent variable
- B_0= the y-intercept (value of y when all other parameters are set to 0)
- $\beta_1 X_1$= the regression coefficient (β_1) of the first independent variable (X_1)
- $\beta_n X_n$ = do the same for however many independent variables you are testing
- ε= the regression coefficient of the last independent variable= model error (a.k.a. how much variation there is in our estimate of)

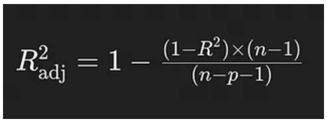
8

R-Squared and Adjusted R-Squared

- The r-squared value quantifies the proportion of the variance in the dependent variable that is predictable from the independent variables.

R-Squared and Adjusted R-Squared

- Adjusted r-squared, on the other hand, adjusts for the number of predictors in the model. It penalizes excessive complexity by incorporating a correction factor based on the number of predictors and sample size. Consequently, adjusted r-squared is often more reliable for comparing models with different numbers of predictors.



$$R^2_{\text{adj}} = 1 - \frac{(1-R^2)\times(n-1)}{(n-p-1)}$$

•*n* represents the number of observations
•*p* signifies the number of predictors in the model
•*R*2 denotes the ordinary r-squared value

Basic MLR in R: model

- **# Fit the multiple linear regression model**

```
mod <- lm(y ~ x1 + x2 + x3, data = my_data)
```

or

```
mod <- lm(dep_var ~ indep_var1 + indep_var2 + indep_var3, data =
my_data)
```

- **# View the summary of the model**

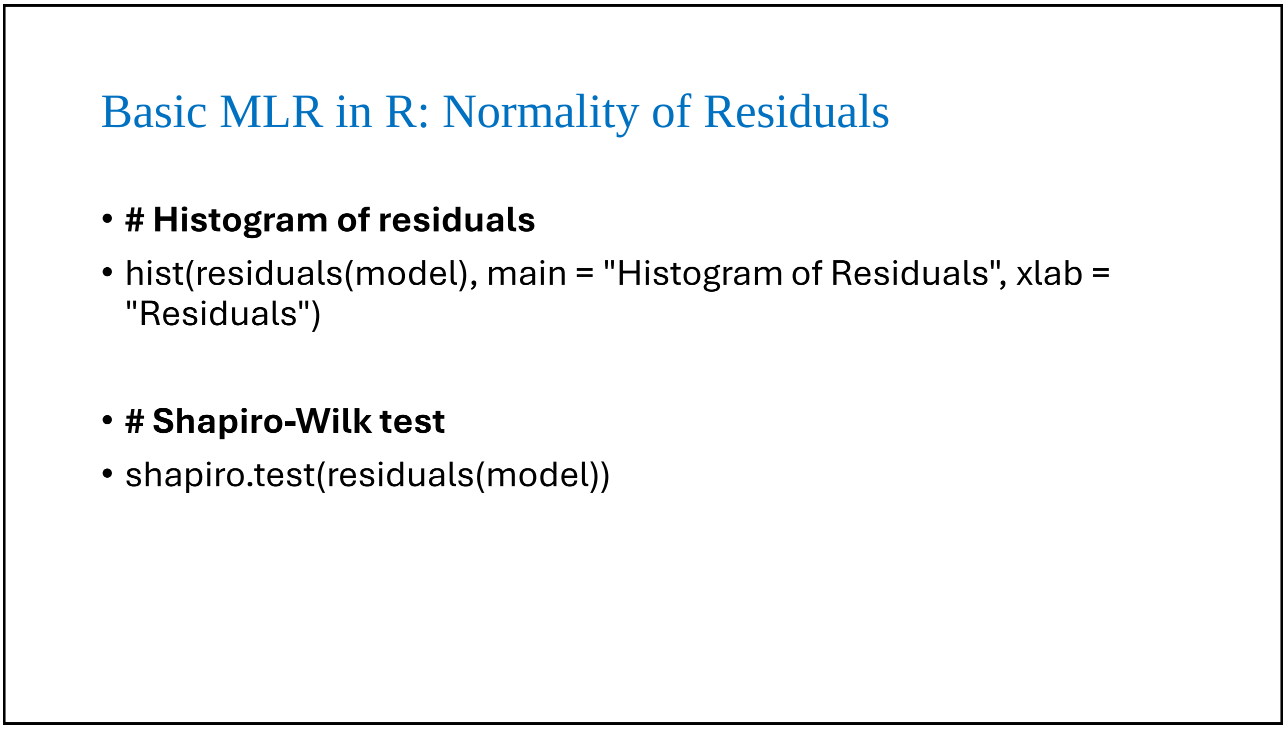
```
summary(model)
```

11

Basic MLR in R: check for multicollinearity

- **# Variance Inflation Factor (VIF)**
- library(car)
- vif(mod)

12



Basic MLR in R: Normality of Residuals

- **# Histogram of residuals**
- hist(residuals(model), main = "Histogram of Residuals", xlab = "Residuals")

- **# Shapiro-Wilk test**
- shapiro.test(residuals(model))

13