BITTIGER

DS 501 Data scientist express bootcamp

Week 3 [Ella]

BITTIGER

版权声明

BITTIGER

Copyright Policy

All content included on the Site or third-party platforms as part of the class, such as text, graphics, logos, button icons, images, audio clips, video clips, live streams, digital downloads, data compilations, and software, is the property of BitTiger or its content suppliers and protected by copyright laws.

Any attempt to redistribute or resell BitTiger content will result in the appropriate legal action being taken.

We thank you in advance for respecting our copyrighted content.

For more info:
see https://www.bittiger.io/termsofuse
and https://www.bittiger.io/termsofservice

Summary

- Important hypothesis testing in R
  - T test
  - Chi square test
- Steps to build a model
  - Problem statement
  - Feature processing
  - Feature engineering
  - Features selection
  - Model evaluation

Summary

- Build linear regression
  - Coefficient estimation
  - Residual variance, p value, F test
  - Residual diagnostics
  - Model performance

Hypothesis testing

- Two sample t test
  - Welch and student t test
  - Calculate t statistics and p value
  - Exercise
- Chi square test
  - Expected and observed values
  - Calculate t statistics and p value
  - Exercise

How to build model

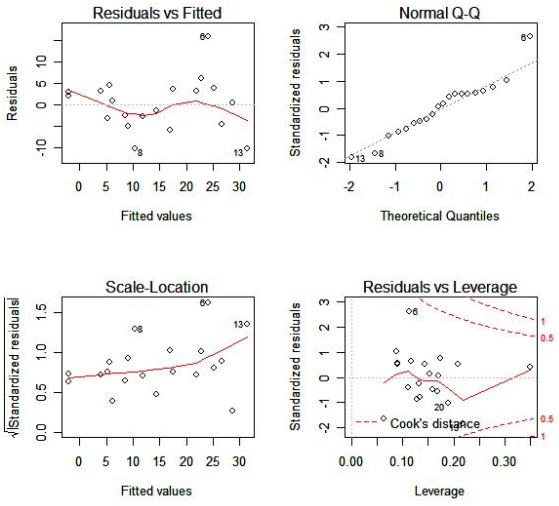
- 1. What features can be included in the model?
  - What features will be available?
    - Example: loan payment related features will not be available when predicting interest rate.
- 2. What features should be included?
  - Remove irrelevant features from intuition
  - Remove features with unique value per row, with same value across rows (no variance): Id, member_id, url...
  - Remove redundant features: dti_joint, annual_income_joint
  - Understand relationship between features and response
    - EDA

How to build model

- 3. Feature processing
  - Missing value imputation
  - Categorical features with too many levels
- 4. Feature engineering
  - Transform/process existing features
  - Generate new features
- 5. Choose models
- 6. Compare results
- Exercise

Linear Regression

- Build linear regression
  - Calculate coefficient LSE/MLE estimation
  - Calculate residual variance, p value, F test, R squared
  - Residual diagnostics
  - Evaluate model performance: RMSE
  - Exercise



https://stats.stackexchange.com/questions/5135/interpretation-of-rs-lm-output?noredirect=1&lq=1