Categorical/Qualitative Predictors

DS 301

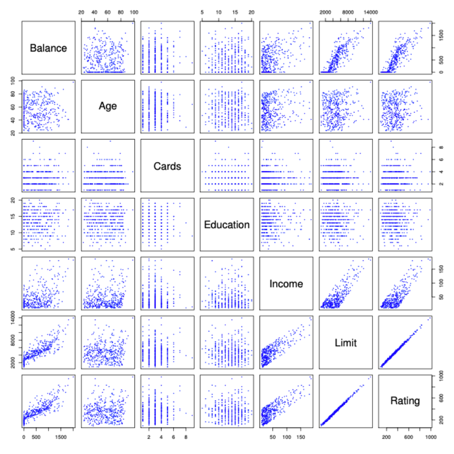
Iowa State University

Other considerations/potential issues in linear regression

- Dealing with categorical or qualitative predictors.
- Model Diagnostics
    - Non-constant variance of error.
    - Non-linear relationships.
- Multicollinearity.

Dealing with categorical or qualitative predictors

Example: credit dataset



Additionally there are 4 categorical (qualitative) predictors.

Credit dataset

Additionally there are 4 categorical (qualitative) predictors:

- `Own`: homeowner or not.
- `student`: student or not.
- `married`: married or not.
- `region`: East, South, West indicating geographical location.

Qualitative predictor with 2 categories

Y: credit card balance

- Suppose we wish to investigate differences in credit card balance between those who own a house and those who don't, holding all other predictors constant.
- If the qualitative predictor only has *two* levels, or possible values, then incorporating it into a regression model is very simple:

We create a dummy variable/indicator

$$x_{i1} = \begin{cases} 1 & \text{if homeowner} \\ 0 & \text{if not.} \end{cases}$$

Use the dummy variable as a predictor in our model:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \sum_{j=2}^{p} \hat{\beta}_j x_{ij}$$

$$= \begin{cases} \hat{\beta}_0 + \hat{\beta}_1 + \sum_{j=2}^{p} \hat{\beta}_j x_{ij} & \text{if homeowner} \\ \hat{\beta}_0 + \sum_{j=2}^{p} \hat{\beta}_j x_{ij} & \text{if not homeowner} \end{cases}$$

Qualitative predictor with 2 categories

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} = \begin{cases} \hat{\beta}_0 & \text{if non.owner} \\ \hat{\beta}_0 + \hat{\beta}_1 & \text{if homeowner} \end{cases}$$

In machine learning community, the creation of dummy variables to handle qualitative predictors is known as "one-hot encoding".

$$\longrightarrow \quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \sum_{j=2}^{p} \hat{\beta}_j X_{ij}$$

dummy variable: ownership

How do we interpret $\hat{\beta}_0$?

The avg credit card balance among those who do not own a home, holding all other predictors constant.

How do we interpret $\hat{\beta}_1$?

The avg. difference in credit card balance between homeowners and non.owners, holding all other predictors constant.

Qualitative predictor with 2 categories

Note: the decision to code ‘owners’ as 1 and ‘non-owners’ as 0 is arbitrary.

- It has no effect on how model fits your data. It will result in the exact same values for $\hat{Y}$. The final predictions will be identical regardless of your coding scheme.
- It does have an effect on the interpretation of your regression coefficients.

Qualitative predictor with more than 2 categories

Y: gpa   X: class rank

Region is a qualitative predictor that has 3 levels: East, South, and West.

Is there any problem with creating one variable that represents all 3 levels:

$$X_{\text{region}} = \begin{cases} 0 & \text{if region is East} \\ 1 & \text{if region is South} \\ 2 & \text{if region is West} \end{cases}$$

$\hat{Y} = \hat{B}_0 + \hat{B}_{Region} \cdot X_{Region} = \{$

$\hat{B}_0$: avg cc balance in east.

$\hat{B}_{Region}$: avg. difference in cc balance btwn east & south

• avg diff. in cc btwn south/west

$\Rightarrow$ assumes a constant difference between levels

$\rightarrow$ This is a very restrictive assumption.

Categorical predictor with more than 2 categories (Region)

- Choose one baseline category (east)

- qualitative predictors w/ K levels $\rightarrow$ $K-1$ dummy variables

$$x_{i1} = \begin{cases} 1 & \text{if from south} \\ 0 & \text{if not from south} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if from west} \\ 0 & \text{if not from west} \end{cases}$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \sum_{j=3}^{p} \hat{\beta}_j x_{ij}$$

$$= \begin{cases} \hat{\beta}_0 + \hat{\beta}_1 + \sum_{j=3}^{p} \hat{\beta}_j x_{ij} & \text{if south} \\ \hat{\beta}_0 + \hat{\beta}_2 + \sum_{j=3}^{p} \hat{\beta}_j x_{ij} & \text{if west} \\ \hat{\beta}_0 + \sum_{j=3}^{p} \hat{\beta}_j x_{ij} & \text{if east} \end{cases}$$

In summary

For each categorical predictor:

- Choose a baseline category (`R` will automatically choose one for you, but you can change this).
- For every other category, define a dummy variable.
- The model fit $\hat{f}$ and its predictions are independent of the choice of the baseline category and the coding scheme.
- However, the interpretation of the regression coefficients and associated hypothesis tests depend on the baseline category and the coding scheme.

Implementation

See R script: `MLR_CategoricalPredictors.R`