The Categorized Regression Data

Posted on September 13, 2018

Introduction

The categorized regression data discussed in this essay come from the simulated *Credit* dataset used in *An Introduction to Statistical Learning*. The dataset contains information about 400 customers and is designed to demonstrate how linear regression can be used to explain or predict average credit-card balance. Its numerical variables include income, credit limit, credit rating, number of cards, age, and years of education. Its categorical variables include student status, marital status, home ownership or gender depending on the dataset edition, and a regional or ethnicity category. The response variable is *Balance* (James et al., 2023; Introduction to Statistical Learning, n.d.).

The original analysis correctly identifies the need to combine numerical and categorical predictors, assess multicollinearity, and examine regression assumptions. However, it contains unclear claims about category totals, treats weak pairwise correlation as sufficient reason to remove variables, and describes assumptions inaccurately. A reliable analysis should begin with a defined research question, encode categories transparently, inspect relationships without relying only on scatterplots, estimate a multiple regression model, evaluate diagnostics, and distinguish prediction from causal interpretation (Fox, 2016; Montgomery et al., 2021).

From Simple to Multiple Linear Regression

Simple linear regression models the expected value of a response using one predictor. If the purpose is to relate credit-card balance to income, the model can be written as:

$$\text{Balance}_i = \beta_0 + \beta_1 \text{Income}_i + \varepsilon_i$$

The intercept β_0 is the expected balance when income equals zero, although that value may not have a meaningful real-world interpretation if zero income lies outside the relevant range. The slope β_1 is the expected change in balance associated with a one-unit increase in income, measured in the dataset's income units. The error term ε represents individual variation not explained by the model (Montgomery et al., 2021).

Multiple regression adds predictors:

$$\text{Balance}_i = \beta_0 + \beta_1 \text{Income}_i + \beta_2 \text{Limit}_i + \beta_3 \text{Student}_i + \ldots + \varepsilon_i$$

Each coefficient is interpreted while holding the other included predictors constant. This conditional interpretation is the main advantage of multiple regression. It can estimate the association between

student status and balance among customers with comparable income and credit limits rather than comparing students and nonstudents without adjustment (Fox, 2016; Montgomery et al., 2021).

Understanding the Credit Variables

The original Credit dataset is simulated for teaching. It should not be described as an actual bank's customer file or used to make real lending decisions. The variables have different scales and meanings (James et al., 2023; Introduction to Statistical Learning, n.d.):

- Income is annual income in thousands of dollars.
- Limit is the customer's credit limit.
- Rating is a credit-rating measure strongly related to limit.
- Cards is the number of credit cards.
- Age is measured in years.
- Education is measured in years.
- Student is a categorical indicator with No and Yes levels.
- Married is a categorical indicator with No and Yes levels.
- Gender, Own, Ethnicity, or Region may appear depending on the R or Python edition of the teaching dataset.
- Balance is the average credit-card balance in dollars.

The existence of slightly different editions is important. An analyst must inspect the actual file rather than copy variable descriptions from another version. Category names, sample size, and units should be confirmed before modeling.

Representing Categorical Variables

Linear regression cannot use text labels directly. A variable with two categories is commonly represented by an indicator. For Student, the analyst may code No as 0 and Yes as 1. If No is the reference category, the coefficient for StudentYes estimates the expected difference in balance between students and nonstudents after adjustment for the other predictors (Fox, 2016).

A categorical variable with three levels requires two indicator variables. Suppose Ethnicity has African American, Asian, and Caucasian levels and Caucasian is selected as the reference. The model includes indicators for African American and Asian. Each coefficient compares that group with the reference while other predictors are held constant. Changing the reference category changes the displayed coefficients but not fitted values or overall model fit.

Categories should not be assigned arbitrary numeric ranks such as 1, 2, and 3 unless the variable is genuinely ordinal and equal spacing is defensible. Coding ethnicity or region as a continuous number would impose a meaningless order and slope.

Descriptive Analysis Before Modeling

Regression should not be the first interaction with the data. The analyst should inspect dimensions, data types, missing values, duplicates, ranges, category counts, and summary statistics. Histograms and boxplots can reveal skewness and outliers. Scatterplots can show whether relationships appear approximately linear and whether variance changes across the range (Fox, 2016; Montgomery et al., 2021).

Group summaries are useful for categorical variables. Mean and median balance can be compared across student status, marital status, and other categories. These unadjusted comparisons are descriptive, not causal. A higher mean among students could reflect differences in income, limits, age, or other variables.

Data quality should also be assessed. Impossible ages, negative limits, inconsistent category labels, or duplicated customer records must be resolved before modeling. Because the teaching dataset is simulated and generally clean, students should still describe these checks as part of a sound workflow.

Correlation and Multicollinearity

Income, Limit, and Rating are likely to be positively related to balance, but Limit and Rating are also strongly related to each other. When predictors contain overlapping information, coefficient estimates can become unstable. This is multicollinearity. It does not necessarily reduce the model's predictive accuracy, but it can inflate standard errors and make individual coefficients difficult to interpret (Fox, 2016; Montgomery et al., 2021).

The original essay proposes removing Rating to reduce multicollinearity, which may be reasonable if Limit and Rating provide nearly the same information. The decision should be supported by a correlation matrix, variance inflation factors, model purpose, and subject-matter reasoning. A fixed rule that every variance inflation factor below 10 is harmless is too crude. Lower thresholds may be appropriate when interpretation is central.

Removing a variable solely because its simple correlation with Balance is weak can be mistaken. A predictor may become informative after adjustment, participate in an interaction, or reduce omitted-variable bias. Variable selection should consider the research question and out-of-sample performance, not only pairwise plots.

Choosing a Model

A reasonable explanatory model might begin with Income, Limit, Student, Cards, Age, Education,

Married, and the available demographic categories. Rating might be excluded when Limit is retained because of their redundancy, or the analyst might compare models containing one or the other. The model should be specified before extensive significance testing when possible.

Automatic procedures such as stepwise selection can produce unstable models and overly optimistic p-values. Better approaches include theory-guided selection, cross-validation, penalized regression, or comparison of a small number of defensible candidate models. If the purpose is prediction, a test set or repeated cross-validation is essential. Adjusted R-squared alone does not measure performance on new customers (James et al., 2023).

Interactions and Nonlinear Relationships

Multiple regression assumes additive effects unless interactions are included. The relationship between income and balance may differ by student status. An interaction model can include Income × Student. The Student coefficient then represents the group difference at the reference income value, while the interaction coefficient indicates whether the income slope differs between students and nonstudents (Fox, 2016).

Centering continuous variables around their means can make interaction coefficients easier to interpret. Polynomial terms or splines can model curvature when the relationship between Limit and Balance is not adequately linear. Such flexibility should be validated because complex models can fit noise.

Regression Assumptions

Linear regression relies on several conditions for reliable inference. The relationship between predictors and the conditional mean of the response should be adequately represented by the model. Errors should be independent when the sampling design supports that assumption. Residual variance should be reasonably constant for standard ordinary-least-squares standard errors. Strongly influential observations should be examined. For small-sample hypothesis tests and confidence intervals, approximate normality of residuals may also matter (Montgomery et al., 2021).

The assumption is not that all variables themselves must be normally distributed. Nor is “reality” a regression assumption. Autocorrelation is most relevant when observations are ordered in time or space; a cross-sectional customer dataset may instead require attention to clustering if customers share institutions or locations.

Diagnostic Plots

A residual-versus-fitted plot assesses nonlinearity and changing variance. A random cloud centered around zero is supportive, though not proof, of adequate specification. A funnel shape suggests heteroscedasticity. A curved pattern suggests omitted nonlinear structure. A normal quantile plot helps assess tail behavior, while leverage and Cook's distance help identify observations with unusual predictor values and substantial influence (Fox, 2016; Montgomery et al., 2021).

Outliers should not be deleted merely because they make a model less convenient. The analyst should determine whether the observation is a data error, a valid rare customer, or evidence that the model is incomplete. Results can be reported with and without highly influential cases when their treatment materially changes conclusions.

Interpreting Model Fit

R-squared is the proportion of observed variation in Balance explained by the fitted model in the sample. An R-squared of 0.62, for example, would mean that 62 percent of sample variation is explained, not that the model is 62 percent accurate or that the predictors cause 62 percent of balance. Adjusted R-squared applies a penalty for adding predictors, but it still describes in-sample fit (James et al., 2023; Montgomery et al., 2021).

Prediction error should be reported in meaningful units through measures such as root mean squared error or mean absolute error on held-out observations. Confidence intervals describe uncertainty in estimated relationships; prediction intervals are wider because they include individual variability.

Ethics and Protected Characteristics

Variables such as gender, ethnicity, age, and marital status require ethical and legal care in real credit applications. Their inclusion in a teaching dataset does not imply that lenders may use them freely. A model can reproduce historical discrimination, use proxy variables, or perform differently across groups. Real-world credit scoring requires applicable law, fairness testing, documentation, data governance, and human oversight.

The simulated Credit data are suitable for learning regression mechanics, not for drawing claims about actual demographic groups. Coefficients should not be interpreted as biological or moral differences. They describe patterns constructed in a simulated dataset under a particular model (James et al., 2023).

Conclusion

Analysis of categorized regression data requires more than inserting variables into a formula. The Credit dataset combines numerical predictors with binary and multi-level categories, making it useful for demonstrating reference coding, conditional interpretation, multicollinearity, interactions, diagnostics, and validation. Limit and Rating should be handled carefully because they may contain redundant information, while weak marginal correlation is not by itself a reason to remove a predictor. Regression assumptions concern the model's errors and functional form, not simple normality of every variable. Finally, the dataset is simulated and should not be used to justify real credit decisions or demographic stereotypes. A credible analysis makes its coding, assumptions, uncertainty, and intended use explicit (James et al., 2023; Fox, 2016; Montgomery et al., 2021).

References

James, Gareth, Daniela Witten, Trevor Hastie, Robert Tibshirani, and Jonathan Taylor. *An Introduction to Statistical Learning: With Applications in Python*. Springer, 2023.

Introduction to Statistical Learning. "Credit Card Balance Data." https://intro-stat-learning.github.io/ISLP/datasets/Credit.html

Fox, John. *Applied Regression Analysis and Generalized Linear Models*. Sage Publications, 2016.

Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to Linear Regression Analysis*. Wiley, 2021.