Companion Explanatory Notes
Linear Regression as an Explanatory Model for Credit Risk

Overview of the Notebook

In this notebook, we study a credit risk dataset using *linear regression* as a **purely explanatory baseline**. Our objective is *not* to build a production-ready fraud or default classifier, but rather to understand:

- how borrower and loan characteristics relate linearly to default outcomes,
- how regression coefficients should be interpreted in a binary-outcome setting.

The dependent variable, `loan_status`, is binary, which places our approach in the family of *Linear Probability Models (LPMs)*. While such models are inappropriate for final classification decisions, we use them because they remain valuable for interpretation, diagnostics, and pedagogical analysis. Throughout the notebook, we emphasize interpretation, residual analysis, and limitations rather than predictive accuracy.

Exercise 0: What Does Linear Regression Compute?

In this exercise, we are asked to interpret linear regression from a geometric perspective. The exercise focuses on understanding the meaning of the fitted vector

$$\hat{y} = X\hat{\beta}$$

when an outcome vector $y \in \mathbb{R}^n$ is approximated using explanatory variables collected in the matrix $X \in \mathbb{R}^{n \times p}$.

Specifically, the exercise asks us to identify what $\hat{y}$ represents among several possible interpretations, and to justify this interpretation using geometric reasoning. It further asks us to interpret the residual vector

$$r = y - \hat{y}$$

by describing where it lies relative to the column space of X, and what the magnitude of an individual residual $|r_i|$ indicates about a particular observation.

Conclusion

We conclude that the fitted vector $\hat{y}$ is the projection of y onto the column space of X. Equivalently, $\hat{y}$ is the unique vector in $\text{col}(X)$ that is closest to y in Euclidean distance. This corresponds to option **B** in the multiple-choice question.

The residual vector $r = y - \hat{y}$ lies in the orthogonal complement of $\text{col}(X)$, meaning it is perpendicular to every column of X. A large absolute residual $|r_i|$ indicates that observation i is poorly explained by the linear model, suggesting that its outcome cannot be well approximated by any linear combination of the available explanatory variables.

Exercise 1: What Will Linear Regression Learn From This Data?

In this exercise, we consider how linear regression behaves when it is used to model a binary outcome representing loan default. The linear model produces a score of the form

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots,$$

which is interpreted as a linear approximation to the probability of default.

The exercise asks us to reason about the meaning of regression coefficients in this setting. In particular, we are asked to interpret what the *sign* of a coefficient indicates, and how domain knowledge should guide expectations about whether a variable ought to be positively or negatively associated with default risk.

Question 1: Interpreting the Sign of a Coefficient

The first question asks what it means for a regression coefficient to be positive or negative when the outcome variable represents default. This requires thinking about how changes in a single explanatory variable affect the fitted linear score produced by the model.

Question 2: Reasoning About Expected Signs

The second question asks us to predict the expected sign of coefficients for several variables based on intuition and domain knowledge, without relying on model output. This encourages us to think about whether increasing a variable should, in principle, be associated with higher or lower default risk.

Question 3: Relative Impact on the Linear Score

The third question asks which variables we expect to have the largest impact on the fitted linear score. This shifts attention from direction alone to the *strength* of the association, prompting us to consider which borrower or loan characteristics are likely to matter most in explaining default behavior.

Conclusion

We conclude that, in a linear model fit to a binary default outcome, the sign of a coefficient captures the *direction* of association between an explanatory variable and default risk. A positive coefficient indicates that higher values of the variable are associated with a higher predicted probability of default, while a negative coefficient indicates an association with lower default risk.

Importantly, the exercise emphasizes that expected signs and relative importance should be reasoned about *before* fitting the model, using economic or behavioral knowledge. Linear regression then serves as a tool to quantify these associations, rather than to replace domain intuition with purely mechanical estimation.

Exercise 2: Interpreting Coefficients

In this exercise, we interpret the estimated coefficients from a linear regression model fit to `loan_status`, after all numerical features have been scaled and categorical variables have been explicitly encoded. The exercise focuses on understanding what the magnitude and sign of these coefficients reveal about default risk.

Question 1: Dominant Risk Drivers

The first question asks us to identify which variable has the largest positive coefficient in the fitted model and to explain why this result is economically reasonable in a credit risk context. This requires connecting coefficient magnitude to the strength of association with default risk.

Question 2: Credit Grade Effects

The second question asks us to examine the coefficients corresponding to credit grade indicators, from `loan_grade_A` through `loan_grade_G`. We are asked to describe how default risk changes as credit grade worsens, and to relate this pattern to standard credit scoring intuition.

Question 3: Borrower Stability

The third question focuses on variables related to borrower characteristics, specifically `person_income`, `person_emp_length`, and `person_age`. We are asked to interpret the signs of these coefficients and to assess whether they align with intuition about financial stability and repayment capacity.

Question 4: Loan Purpose and Ownership

The fourth question considers categorical variables representing loan intent or ownership. We are asked to interpret what a negative coefficient means under the chosen encoding scheme, and to reason about why loans associated with lower average risk may still experience defaults.

Question 5: Surprising or Subtle Results

The final question highlights the near-zero and symmetric coefficients for `cb_person_default_on_file_N` and `cb_person_default_on_file_Y`. We are asked to propose explanations for why a historically important risk indicator might appear weak in this linear model.

Conclusion

From this exercise, we conclude that the magnitude of a regression coefficient reflects the relative importance of a variable in the linear approximation to default risk, once all features are placed on a common scale. Variables with large positive coefficients are those most strongly associated with increased default risk, while negative coefficients indicate associations with reduced risk.

We observe that credit grade coefficients follow an ordered pattern, with worse grades corresponding to higher default risk. This confirms that the model has learned a meaningful and economically sensible structure from the data. Borrower characteristics related to income, employment stability, and age tend to have negative coefficients, aligning with the idea that financial stability reduces default risk.

Negative coefficients for certain loan purposes indicate that, relative to the baseline category, these loans are associated with lower average default risk. However, the presence of defaults among such loans highlights that regression coefficients capture average tendencies rather than guarantees.

Finally, the weak and symmetric coefficients for prior default indicators suggest that their information may already be captured by other correlated variables, or that the linear model is unable to fully express their nonlinear or interaction effects. This underscores an important limitation of linear regression: even well-known risk factors may appear muted when their effects are indirect or overlapping with other features.

Exercise 3: How Variable Importance Changes When Information Is Removed

In this exercise, we investigate how linear regression redistributes importance across variables when a key source of information is removed. In the previous model, a composite measure of financial stress emerged as the strongest driver of default risk. This variable combined information about loan size and borrower income into a single feature.

To examine how the model responds to overlapping information, we refit the linear regression model after removing this composite variable. The exercise then asks us to analyze how the remaining coefficients change as a result.

Question 1: Changes in Coefficient Magnitude

The first question asks which variables experience the largest changes in coefficient magnitude after the composite variable is removed. This encourages us to identify which features were previously

sharing explanatory power with the removed variable.

Question 2: Redistribution Between Related Variables

The second question focuses on how the coefficients of variables directly related to loan size and borrower income change relative to the previous model. We are asked to compare their magnitudes before and after the removal.

Question 3: Why Coefficients Become Larger

The third question asks us to explain why removing a variable that summarizes multiple risk factors causes other coefficients to increase in magnitude. This requires reasoning about how linear regression allocates explanatory power when predictors are correlated.

Question 4: Implications for Variable Importance

The final question asks what this behavior reveals about how linear regression determines which variables appear important when several predictors contain overlapping or redundant information.

Conclusion

We conclude that when a composite variable capturing multiple aspects of risk is removed, linear regression reallocates its explanatory role to the remaining variables that contain related information. As a result, coefficients for these variables increase in magnitude, reflecting their newly assumed responsibility for explaining variation in the outcome.

This exercise illustrates that coefficient size in a linear model does not represent intrinsic importance, but rather importance *conditional on the set of variables included*. When predictors are correlated, linear regression distributes explanatory power among them, and removing one predictor forces the model to compensate by amplifying others.

Overall, this highlights a key limitation of linear regression for interpreting variable importance: coefficients are sensitive to feature selection, and large coefficients may arise not because a variable is uniquely informative, but because it is the best remaining proxy for omitted information.

Exercise 4: Understanding Residuals

In this exercise, we examine residuals produced by the linear regression model. After fitting the model, we compute predicted values $\hat{y}$ and define the residual for each observation as

$$\text{residual}_i = y_i - \hat{y}_i.$$

Residuals measure how much each loan's observed outcome deviates from what the linear model explains using the available features.

The exercise asks us to interpret both the sign and magnitude of residuals, and to reason about what residual patterns reveal about model limitations and missing information.

Question 1: Large Positive Residuals

The first question asks what it means for a loan to have a large positive residual. This corresponds to cases where the observed outcome is worse than what the model predicts.

Question 2: Large Negative Residuals

The second question asks what a large negative residual represents. This focuses on cases where the observed outcome is better than what the model predicts.

Question 3: Small Residuals for Defaulted Loans

The third question asks why some defaulted loans may still have small residuals. This encourages us to distinguish between adverse outcomes and poor model performance.

Question 4: Missing Information

The final question asks what kinds of information may be missing for loans with very large residuals. This highlights the role of unobserved factors and model misspecification.

Conclusion

We conclude that a large positive residual corresponds to a loan that defaulted despite being predicted by the model as relatively low risk, indicating that the model underestimated default risk for that observation. Conversely, a large negative residual corresponds to a loan that was predicted to be risky but ultimately did not default.

Some defaulted loans have small residuals because the model already assigns them a high predicted risk, meaning the observed outcome aligns closely with the model's expectations. In such cases, default does not indicate a failure of the linear approximation.

Loans with very large residuals likely depend on information not captured by the available features, such as sudden income shocks, behavioral factors, reporting errors, or nonlinear interactions between variables. Residual analysis therefore serves as a diagnostic tool for identifying where the linear model fails to capture important aspects of the data-generating process.

Note on Residual Analysis in a Linear Probability Model

Residual analysis in this notebook is used strictly as an *explanatory and diagnostic* tool. Although the outcome variable is binary, linear regression still computes the orthogonal projection of y onto the column space of X, and residuals are therefore well-defined.

We do *not* interpret residuals as satisfying classical regression assumptions. Instead, residuals are interpreted qualitatively to highlight unexplained variation and model limitations.

Exercise 5: Model Metrics

In this exercise, we summarize the performance of the linear regression model using two standard metrics: Mean Squared Error (MSE) and the coefficient of determination R^2.

The Mean Squared Error is defined as

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

and measures the average squared difference between the observed loan outcome and the linear score produced by the model. The R^2 statistic measures the fraction of variation in the outcome explained by the model, relative to a baseline that predicts the same value for every observation.

The exercise asks us to interpret these metrics conceptually, rather than as absolute indicators of model quality.

Question 1: Interpreting MSE

The first question asks what a larger MSE indicates about the average size of the model's residuals. This focuses on understanding MSE as a summary of prediction error magnitude.

Question 2: Why Error Is Expected

The second question asks why a non-zero MSE is expected when modeling loan default using a linear model. This requires reasoning about the inherent unpredictability of default outcomes and the limitations of linear approximation.

Question 3: Role of R^2

The third question asks how R^2 helps contextualize the error measured by MSE. This shifts attention from absolute error to the proportion of systematic variation captured by the model.

Conclusion

We conclude that a larger MSE indicates larger average residuals, meaning that the model's predictions deviate more substantially from observed outcomes. In the context of loan default, a non-zero MSE is unavoidable because default is driven by many unobserved, stochastic, and behavioral factors that cannot be fully captured by a linear combination of observed features.

The R^2 statistic complements MSE by indicating how much of the variation in loan default is explained by the model relative to a naïve baseline. A moderate or low R^2 does not imply model failure in this setting; instead, it reflects the inherent noise in binary default outcomes. Together,

MSE and R^2 provide a high-level diagnostic summary of how much structure the linear model is able to capture, rather than a definitive measure of predictive adequacy.

Final Methodological Note: The Linear Probability Model (LPM)

Although Logistic Regression is the standard modeling choice for binary outcomes such as loan default, this notebook deliberately employs Linear Regression. This approach is known in econometrics as the *Linear Probability Model (LPM)*.

Why Use a Linear Model for a Binary Outcome?

1. **Interpretability of Marginal Effects.** In an LPM, coefficients are directly interpretable as changes in the predicted probability of default. For example, a coefficient of 0.05 on a standardized feature implies that a one-standard-deviation increase in that feature is associated with an approximate five percentage point increase in default risk.

2. **Structural Discovery and Transparency.** The LPM serves as a transparent baseline that reveals the linear structure of the data. Before introducing nonlinear or highly flexible models, the linear framework allows us to clearly see which variables matter, in which direction, and by how much.

3. **Understanding Information Overlap.** As demonstrated in Exercise 3, the linear framework allows us to use geometric intuition to understand how explanatory power is redistributed when correlated variables are removed.

Final Conclusion

We do not recommend linear regression as a final production classifier for credit scoring. Instead, we use the Linear Probability Model as a tool for inference, feature ranking, and understanding the linear component of default risk. Within this role, the LPM provides a principled and interpretable foundation upon which more complex and nonlinear models can be meaningfully built.