Week 6: Linear Regression

Francisco Richter and Ernst Wit

Introduction to Data Science (MSc)

1 Introduction to Linear Regression

Linear regression represents the cornerstone of statistical modeling and machine learning, providing both a fundamental framework for understanding relationships between variables and a foundation upon which more complex methods are built. The mathematical elegance of linear regression lies in its ability to provide exact analytical solutions while maintaining interpretability and theoretical rigor that extends to modern data science applications.

The theoretical importance of linear regression extends far beyond its apparent simplicity. The method embodies fundamental principles of statistical inference, including maximum likelihood estimation, hypothesis testing, and model selection, while providing a natural bridge between classical statistics and contemporary machine learning. Understanding linear regression deeply provides essential insights into bias-variance tradeoffs, regularization, and the geometric interpretation of statistical methods that are crucial for advanced data science practice.

From a computational perspective, linear regression serves as an ideal introduction to matrix algebra in statistics, demonstrating how high-dimensional problems can be solved efficiently through linear algebraic operations. The connection between geometric intuition and algebraic manipulation in linear regression provides essential preparation for understanding more complex methods such as principal component analysis, support vector machines, and neural networks.

2 Mathematical Foundations and Model Specification

2.1 Simple Linear Regression: Geometric and Algebraic Perspectives

The simple linear regression model provides a natural starting point for understanding the mathematical structure underlying all linear models. The model specification captures the essential idea that observed responses can be decomposed into systematic and random components, with the systematic component following a linear relationship.

Definition 2.1 (Simple Linear Regression Model)**.** *For observations* (x_i, y_i)*,* $i = 1, 2, \ldots, n$*, the simple linear regression model is:*

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where β_0 *is the intercept parameter representing the expected response when* $x = 0$*,* β_1 *is the slope parameter representing the expected change in response per unit change in* x*, and* ϵ_i *are independent error terms with* $E[\epsilon_i] = 0$ *and* $Var(\epsilon_i) = \sigma^2$*.*

The geometric interpretation of this model reveals that we are fitting a line through a cloud of points in two-dimensional space, with the line representing the systematic relationship and the

vertical deviations representing random variation. The parameters β_0 and β_1 determine the position and orientation of this line, while σ^2 quantifies the magnitude of random variation around the line.

The probabilistic interpretation requires additional distributional assumptions. Under the classical normal linear model, we assume $\epsilon_i \sim N(0, \sigma^2)$ independently, which implies that $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$. This normality assumption enables exact finite-sample inference and provides the foundation for confidence intervals and hypothesis tests.

2.2 Least Squares Estimation: Optimization and Geometric Interpretation

The method of least squares provides both an intuitive and mathematically elegant approach to parameter estimation. The criterion minimizes the sum of squared vertical deviations between observed and predicted values, corresponding to the geometric intuition of finding the line that best fits the data.

Theorem 2.1 (Least Squares Estimators for Simple Regression)**.** *The least squares estimators that minimize* $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$ *are:*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \tag{1}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{2}$$

where $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ *and* $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

The derivation through calculus reveals the normal equations that characterize the least squares solution. Taking partial derivatives of the sum of squared errors with respect to β_0 and β_1 and setting them equal to zero yields:

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{3}$$

$$\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{4}$$

These equations have profound geometric interpretation. The first equation states that the sum of residuals is zero, meaning the fitted line passes through the point $(\bar{x}, \bar{y})$. The second equation states that the residuals are orthogonal to the predictor variable, meaning the fitted line is positioned to minimize the total squared distance to the data points.

2.3 Statistical Properties of Least Squares Estimators

The theoretical properties of least squares estimators provide the foundation for statistical inference and reveal why this method has such broad applicability and appeal.

Theorem 2.2 (Gauss-Markov Theorem)**.** *Under the assumptions of linearity, independence, homoscedasticity, and zero mean errors, the least squares estimators are the Best Linear Unbiased Estimators (BLUE), meaning they have minimum variance among all linear unbiased estimators.*

This optimality result is remarkable because it holds without requiring normality assumptions. The proof relies on the method of Lagrange multipliers to show that any other linear unbiased estimator must have larger variance than the least squares estimator.

Under the additional assumption of normality, the least squares estimators have exact finite-sample distributions that enable precise statistical inference.

Theorem 2.3 (Sampling Distributions Under Normality)**.** *Under the normal linear model assumptions:*

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right) \tag{5}$$

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right) \tag{6}$$

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2} \tag{7}$$

where $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$ *is the residual mean square.*

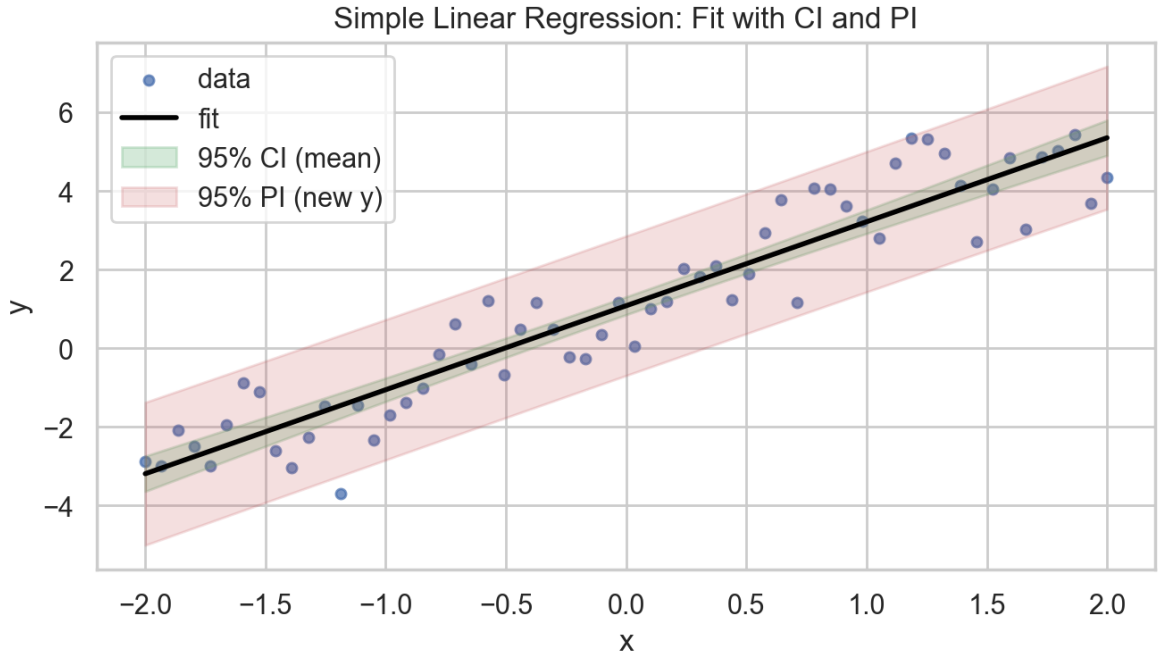

Figure 1: Simple linear regression fit with 95% confidence band for the mean response (green) and 95% prediction band for a new observation (red). Prediction intervals are wider as they include irreducible noise.

These distributional results reveal several important insights. The variance of $\hat{\beta}_1$ is inversely proportional to S_{xx}, indicating that more spread in the predictor variable leads to more precise slope estimation. The variance of $\hat{\beta}_0$ depends on both the sample size and the distance of $\bar{x}$ from zero, showing that intercept estimation is most precise when the predictor values are centered around zero.

3 Multiple Linear Regression: Matrix Formulation and Theory

3.1 Matrix Representation and Geometric Interpretation

The extension to multiple predictors requires matrix notation to maintain mathematical clarity and computational efficiency. The matrix formulation reveals the geometric structure underlying multiple regression and provides the foundation for understanding more advanced methods.

Definition 3.1 (Multiple Linear Regression Model)**.** *The multiple linear regression model with p predictors is:*

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{y}$ is the $n \times 1$ response vector, $\mathbf{X}$ is the $n \times (p+1)$ design matrix with first column of ones, $\boldsymbol{\beta}$ is the $(p+1) \times 1$ parameter vector, and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$.

The design matrix $\mathbf{X}$ has the structure:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

The geometric interpretation extends the two-dimensional line-fitting problem to fitting a hyperplane in $(p+1)$-dimensional space. The response vector $\mathbf{y}$ lies in n-dimensional space, while the fitted values $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ lie in the column space of $\mathbf{X}$, which is a $(p+1)$-dimensional subspace of $\mathbb{R}^n$.

3.2 Least Squares Solution and Projection Theory

The least squares solution in multiple regression has an elegant geometric interpretation as an orthogonal projection of the response vector onto the column space of the design matrix.

Theorem 3.1 (Matrix Form of Least Squares Estimator)**.** *The least squares estimator is:*

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

provided that $\mathbf{X}^T\mathbf{X}$ is invertible, which requires that $\mathbf{X}$ has full column rank.

The matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is called the hat matrix because it transforms the observed responses into fitted values: $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$. The hat matrix is idempotent ($\mathbf{H}^2 = \mathbf{H}$) and symmetric, reflecting its role as an orthogonal projection operator.

The residual vector $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$ lies in the orthogonal complement of the column space of $\mathbf{X}$. This orthogonality relationship, expressed as $\mathbf{X}^T\mathbf{e} = \mathbf{0}$, generalizes the normal equations from simple regression and provides the foundation for all least squares theory.

3.3 Statistical Properties in Multiple Regression

The statistical properties of the multiple regression estimator extend the results from simple regression while revealing additional complexity arising from the correlation structure among predictors.

Theorem 3.2 (Properties of Multiple Regression Estimator)**.** *Under the classical assumptions:*

$$E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta} \quad \textit{(unbiasedness)} \tag{8}$$

$$Cov(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \quad \textit{(covariance matrix)} \tag{9}$$

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}) \quad \textit{(exact distribution)} \tag{10}$$

The covariance matrix $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$ reveals how the precision of parameter estimates depends on the design matrix structure. The diagonal elements give the variances of individual parameter estimates, while the off-diagonal elements give the covariances between estimates. High correlation among predictors leads to large off-diagonal elements in $(\mathbf{X}^T\mathbf{X})^{-1}$, resulting in high variance and correlation among parameter estimates.

4 Statistical Inference and Hypothesis Testing

4.1 Estimation of Error Variance and Degrees of Freedom

The estimation of error variance requires careful attention to degrees of freedom, which reflect the number of independent pieces of information available for estimation after accounting for the parameters that have been estimated.

Theorem 4.1 (Unbiased Estimation of Error Variance)**.** *The residual mean square:*

$$\hat{\sigma}^2 = \frac{RSS}{n-p-1} = \frac{\mathbf{e}^T\mathbf{e}}{n-p-1}$$

is an unbiased estimator of σ^2, where $n-p-1$ represents the error degrees of freedom.

The degrees of freedom calculation reflects the constraint that the residuals must satisfy $p+1$ orthogonality conditions corresponding to the normal equations. Each constraint reduces the effective sample size by one, leaving $n-p-1$ degrees of freedom for error estimation.

The distribution theory for the error variance estimator provides the foundation for constructing confidence intervals and hypothesis tests.

Theorem 4.2 (Distribution of Error Variance Estimator)**.** *Under the normal linear model:*

$$\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1}$$

and $\hat{\sigma}^2$ is independent of $\hat{\boldsymbol{\beta}}$.

4.2 Individual Parameter Tests and Confidence Intervals

Testing individual regression coefficients requires combining the normal distribution of the parameter estimates with the chi-square distribution of the error variance estimate, leading to t-distributions that account for the uncertainty in both components.

Theorem 4.3 (t-Tests for Individual Parameters)**.** *For testing $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$:*

$$T = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \sim t_{n-p-1}$$

where $SE(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 C_{jj}}$ and C_{jj} is the j-th diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$.

The standard error formula reveals how the precision of individual parameter estimates depends on both the error variance and the design matrix structure. Large values of C_{jj} indicate that the j-th predictor is highly correlated with other predictors, leading to imprecise estimation and potential multicollinearity problems.

Confidence intervals for individual parameters follow directly from the t-distribution:

$$\hat{\beta}_j \pm t_{\alpha/2,n-p-1} \cdot \mathrm{SE}(\hat{\beta}_j)$$

These intervals provide ranges of plausible values for the true parameters and can be used to assess the practical significance of estimated effects.

4.3 Overall Model Tests and Analysis of Variance

Testing the overall significance of the regression model requires comparing the fitted model to the null model containing only an intercept. This comparison leads to the F-test, which has a natural interpretation in terms of explained versus unexplained variation.

Theorem 4.4 (F-Test for Overall Model Significance)**.** *For testing* $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ *versus* H_1 *: at least one* $\beta_j \neq 0$*:*

$$F = \frac{MSR}{MSE} = \frac{SSR/p}{RSS/(n-p-1)} \sim F_{p,n-p-1}$$

where $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ *is the regression sum of squares.*

The F-statistic has a natural interpretation as the ratio of explained variation per parameter to unexplained variation per degree of freedom. Large values indicate that the predictors explain a substantial amount of variation relative to the residual variation.

The analysis of variance (ANOVA) decomposition provides a systematic framework for understanding the sources of variation in the response:

$$\text{SST} = \text{SSR} + \text{RSS}$$

where $\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the total sum of squares. This decomposition forms the basis for the coefficient of determination and related measures of model fit.

5 Model Assessment and Diagnostic Methods

5.1 Measures of Model Fit

The coefficient of determination provides a standardized measure of the proportion of variation explained by the regression model.

Definition 5.1 (Coefficient of Determination)**.**

$$R^2 = \frac{SSR}{SST} = 1 - \frac{RSS}{SST}$$

represents the proportion of total variation in the response explained by the regression model.

While R^2 provides an intuitive measure of model fit, it has the limitation of always increasing when additional predictors are added to the model, regardless of their true relevance. The adjusted R^2 addresses this limitation by penalizing model complexity.

Definition 5.2 (Adjusted Coefficient of Determination)**.**

$$R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{SST/(n-1)} = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

The adjusted R^2 can decrease when irrelevant predictors are added, making it more suitable for model comparison and selection.

5.2 Residual Analysis and Assumption Checking

Residual analysis provides the primary tool for assessing model adequacy and identifying violations of regression assumptions. Different types of residuals provide different insights into model performance.

Standardized residuals account for the heteroscedasticity induced by the regression fit:

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

where h_{ii} is the i-th diagonal element of the hat matrix. These residuals have approximately unit variance and can be compared to standard normal quantiles.

Studentized residuals provide even better standardization by using leave-one-out estimates of the error variance:

$$t_i = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}}$$

where $\hat{\sigma}_{(i)}$ is the error standard deviation estimated with observation i removed.

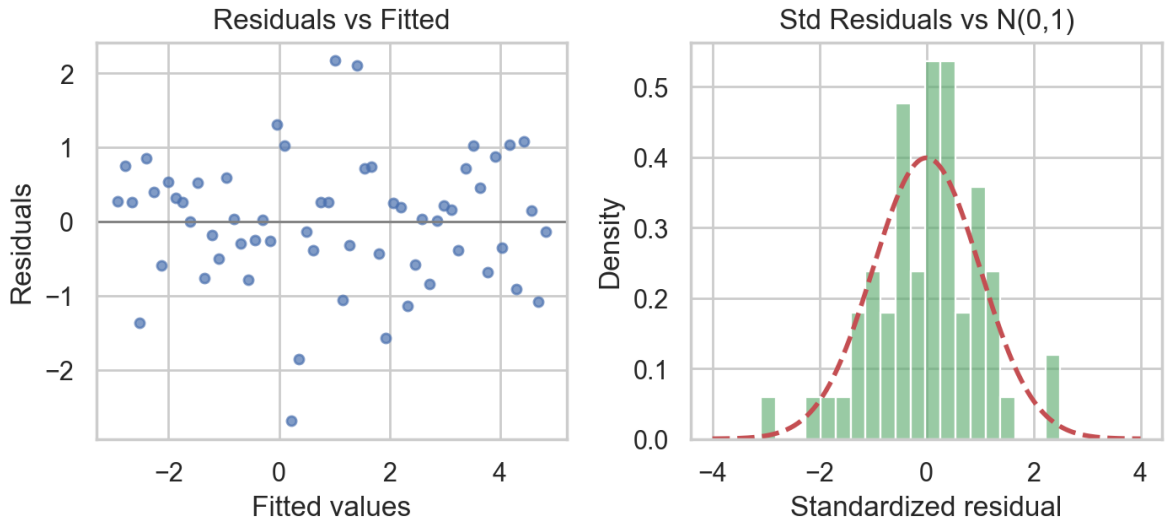

Figure 2: Left: residuals versus fitted values to assess linearity and homoscedasticity. Right: standardized residuals compared to the standard normal density.

5.3 Influential Observations and Leverage

The identification of influential observations requires understanding how individual data points affect the regression fit. Leverage measures the potential for influence based on predictor values, while influence measures the actual impact on the fitted model.

Definition 5.3 (Leverage)**.** *The leverage of observation i is:*

$$h_{ii} = \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i$$

where $\mathbf{x}_i$ is the i-th row of the design matrix. High leverage points have $h_{ii} > 2(p+1)/n$.

Leverage depends only on the predictor values and identifies observations that are outlying in the predictor space. High leverage points have the potential to strongly influence the regression fit, but their actual influence depends on whether their response values are consistent with the pattern established by other observations.

Definition 5.4 (Cook's Distance)**.** *Cook's distance measures the actual influence of observation i:*

$$D_i = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})^T (\mathbf{X}^T \mathbf{X})(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{(p+1)\hat{\sigma}^2}$$

where $\hat{\boldsymbol{\beta}}_{(i)}$ is the parameter estimate with observation i removed.

Cook's distance can be computed efficiently without actually refitting the model:

$$D_i = \frac{r_i^2}{p+1} \cdot \frac{h_{ii}}{1 - h_{ii}}$$

This formula reveals that influence depends on both the size of the residual and the leverage of the observation.

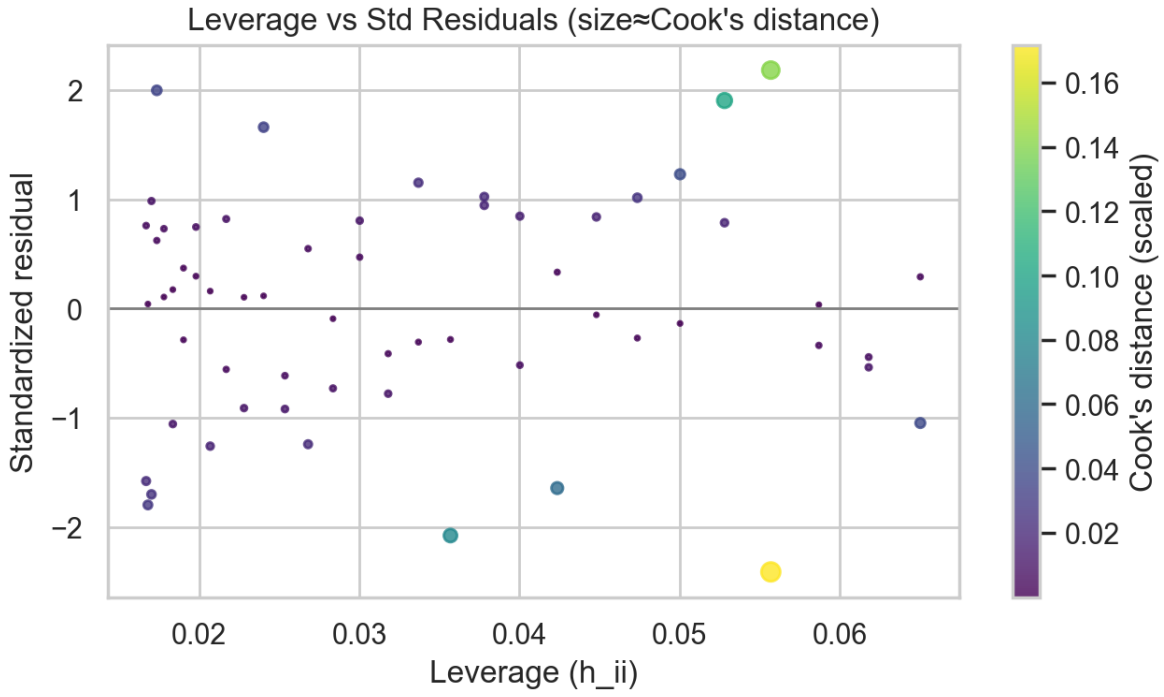

Figure 3: Leverage versus standardized residuals with point size scaled by Cook's distance. High leverage with large residuals indicates influential observations.

6 Regularization and Modern Extensions

6.1 Ridge Regression: Bias-Variance Tradeoff

Ridge regression addresses the problems of multicollinearity and overfitting by adding a penalty term to the least squares objective function. This modification introduces bias but can substantially reduce variance, often leading to improved prediction performance.

Definition 6.1 (Ridge Regression)**.** *Ridge regression minimizes the penalized sum of squares:*

$$RSS(\lambda) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2$$

where $\lambda \geq 0$ is the regularization parameter.

The ridge regression estimator has a closed-form solution:

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

The addition of $\lambda\mathbf{I}$ to $\mathbf{X}^T\mathbf{X}$ ensures that the matrix is invertible even when $\mathbf{X}^T\mathbf{X}$ is singular, addressing multicollinearity problems. The regularization parameter λ controls the amount of shrinkage, with larger values producing more shrinkage toward zero.

The bias-variance decomposition for ridge regression reveals the fundamental tradeoff:

$$\text{Bias}(\hat{\boldsymbol{\beta}}_{\text{ridge}}) = -\lambda(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\boldsymbol{\beta} \tag{11}$$

$$\text{Cov}(\hat{\boldsymbol{\beta}}_{\text{ridge}}) = \sigma^2(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1} \tag{12}$$

The bias increases with λ, while the variance decreases, creating a tradeoff that can be optimized through cross-validation or other model selection techniques.

6.2 LASSO Regression: Sparsity and Variable Selection

LASSO regression replaces the quadratic penalty of ridge regression with an absolute value penalty, leading to sparse solutions that perform automatic variable selection.

Definition 6.2 (LASSO Regression)**.** *LASSO minimizes the penalized sum of squares:*

$$RSS(\lambda) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1$$

where $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$ is the L_1 norm.

Unlike ridge regression, LASSO does not have a closed-form solution, but can be solved efficiently using coordinate descent or other optimization algorithms. The key property of LASSO is that it can set coefficient estimates exactly to zero, effectively removing variables from the model.

The sparsity property of LASSO arises from the geometry of the L_1 penalty. The constraint region $\|\boldsymbol{\beta}\|_1 \leq t$ forms a diamond (in two dimensions) or hyperoctahedron (in higher dimensions) with sharp corners at the coordinate axes. The least squares solution is most likely to intersect this constraint region at a corner, where some coordinates are zero.

6.3 Elastic Net and Hybrid Approaches

The elastic net combines ridge and LASSO penalties to address limitations of each method when used alone.

Definition 6.3 (Elastic Net)**.** *The elastic net minimizes:*

$$RSS(\lambda_1, \lambda_2) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|^2$$

The elastic net addresses the limitation of LASSO in selecting at most n variables when $p > n$, and its tendency to select only one variable from groups of highly correlated variables. The ridge component encourages the selection of groups of correlated variables, while the LASSO component maintains sparsity.

7 Applications in Modern Data Science

7.1 High-Dimensional Regression

Modern data science applications often involve high-dimensional settings where the number of predictors p is comparable to or larger than the sample size n. In these settings, ordinary least squares may not be feasible or may severely overfit, making regularization essential.

The theoretical properties of regularized regression in high-dimensional settings have been extensively studied. Under appropriate sparsity assumptions, LASSO can achieve near-optimal performance even when $p \gg n$. The key insight is that if only a small number of predictors are truly relevant, LASSO can identify these predictors and estimate their coefficients accurately.

The irrepresentable condition provides a theoretical framework for understanding when LASSO can correctly identify the true model structure. This condition requires that the irrelevant predictors are not too highly correlated with the relevant predictors, ensuring that LASSO will not incorrectly include irrelevant variables.

7.2 Feature Engineering and Polynomial Regression

Linear regression can accommodate non-linear relationships through feature engineering, where new predictors are created as functions of the original predictors. Polynomial regression represents a systematic approach to feature engineering that can capture curved relationships.

For a single predictor, the polynomial regression model of degree d is:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_d x_i^d + \epsilon_i$$

While this model is non-linear in the predictor x, it remains linear in the parameters β_j, allowing the use of standard linear regression techniques. The challenge lies in selecting the appropriate degree d, which can be addressed through cross-validation or information criteria.

Interaction terms provide another important form of feature engineering, allowing the effect of one predictor to depend on the value of another predictor. For two predictors, the interaction model is:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \epsilon_i$$

The interaction coefficient β_3 captures how the effect of x_1 changes with the value of x_2, providing flexibility to model complex relationships.

7.3 Causal Inference and Regression

Linear regression plays a crucial role in causal inference, though the interpretation of regression coefficients as causal effects requires careful attention to confounding and selection bias. In observational studies, regression coefficients represent associations that may or may not reflect causal relationships.

The potential outcomes framework provides a formal approach to causal inference that clarifies when regression coefficients can be interpreted causally. Under the assumption of no unmeasured confounders, regression adjustment can provide unbiased estimates of causal effects.

Instrumental variable regression provides an approach to causal inference when unmeasured confounders are present. The method uses instrumental variables that affect the treatment but not the outcome directly, enabling identification of causal effects even in the presence of confounding.

8 Computational Aspects and Implementation

8.1 Numerical Stability and Matrix Computations

The computational implementation of linear regression requires attention to numerical stability, particularly when the design matrix is ill-conditioned or nearly singular. Direct computation of $(\mathbf{X}^T\mathbf{X})^{-1}$ can be numerically unstable and should be avoided in favor of more robust approaches.

The QR decomposition provides a numerically stable approach to least squares computation. If $\mathbf{X} = \mathbf{QR}$ where $\mathbf{Q}$ has orthonormal columns and $\mathbf{R}$ is upper triangular, then:

$$\hat{\boldsymbol{\beta}} = \mathbf{R}^{-1}\mathbf{Q}^T\mathbf{y}$$

This approach avoids forming $\mathbf{X}^T\mathbf{X}$ explicitly and provides better numerical properties, particularly when the design matrix is ill-conditioned.

The singular value decomposition (SVD) provides even greater numerical stability and insight into the structure of the regression problem. If $\mathbf{X} = \mathbf{UDV}^T$, then:

$$\hat{\boldsymbol{\beta}} = \mathbf{VD}^{-1}\mathbf{U}^T\mathbf{y}$$

The SVD reveals the effective rank of the design matrix and provides a natural approach to handling rank-deficient problems through truncation of small singular values.

8.2 Cross-Validation and Model Selection

Cross-validation provides a general framework for model selection that is particularly important for regularized regression methods. The basic principle involves splitting the data into training and validation sets, fitting the model on the training set, and evaluating performance on the validation set.

K-fold cross-validation provides a systematic approach that uses all data for both training and validation. The data are divided into K folds, and the model is fit K times, each time using $K-1$ folds for training and one fold for validation. The cross-validation error is the average of the validation errors across all folds.

For regularized regression, cross-validation is typically used to select the optimal value of the regularization parameter. A grid of candidate values is specified, and cross-validation is performed for each value. The value that minimizes the cross-validation error is selected as optimal.

9 Summary

Linear regression represents a fundamental framework in statistical learning that provides both theoretical insights and practical tools essential for modern data science. The mathematical foundations, rooted in linear algebra and statistical theory, provide a solid foundation for understanding more complex methods while offering powerful tools for analyzing real-world data.

The key insights from linear regression include the bias-variance tradeoff, the importance of regularization in high-dimensional settings, and the geometric interpretation of statistical methods. These concepts extend far beyond linear regression itself, providing essential background for understanding machine learning algorithms, statistical inference, and model selection.

The modern extensions of linear regression, including ridge and LASSO regression, demonstrate how classical statistical methods can be adapted to address contemporary challenges in data science. The ability to handle high-dimensional data, perform automatic variable selection, and balance bias and variance makes regularized regression an essential tool in the data scientist's toolkit.

The computational aspects of linear regression, including numerical stability and efficient algorithms, provide important lessons for implementing statistical methods in practice. Understanding these computational considerations is crucial for developing robust and scalable data science applications.

10 Exercises

1. **Least Squares Derivation:** Derive the normal equations for multiple linear regression by minimizing the sum of squared errors. Show that the least squares estimator can be written as $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ and explain the geometric interpretation of this formula.

2. **Gauss-Markov Theorem:** Prove the Gauss-Markov theorem showing that the least squares estimator is BLUE. Explain why this result is important and discuss what happens when the assumptions are violated.

3. **Ridge Regression Bias-Variance:** Derive expressions for the bias and variance of the ridge regression estimator. Show how the bias-variance tradeoff depends on the regularization parameter λ and the eigenvalues of $\mathbf{X}^T\mathbf{X}$.

4. **Hat Matrix Properties:** Prove that the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is idempotent and symmetric. Explain the geometric interpretation of the hat matrix and its role in regression diagnostics.

5. **Residual Analysis:** Explain the difference between raw residuals, standardized residuals, and studentized residuals. Derive the formula for standardized residuals and explain why they are preferred for diagnostic purposes.

6. **LASSO Sparsity:** Explain why the LASSO penalty $\lambda\|\boldsymbol{\beta}\|_1$ leads to sparse solutions while the ridge penalty $\lambda\|\boldsymbol{\beta}\|^2$ does not. Use geometric arguments to illustrate the difference between L_1 and L_2 penalties.

7. **Cross-Validation Theory:** Derive the leave-one-out cross-validation formula for linear regression. Explain why this formula allows LOOCV to be computed from a single model fit and discuss the relationship between LOOCV and AIC.

8. **High-Dimensional Regression:** Discuss the challenges that arise when $p > n$ in linear regression. Explain how regularization addresses these challenges and describe the conditions under which LASSO can successfully recover sparse models in high-dimensional settings.