Probit and Logit Regression Analysis

Question

Assignment Question

Write short notes on Probit Regression. Among the two logit models, how do you determine which model is better? Explain in detail, both theoretically and with examples.

Answer

Probit Regression

Probit regression is a type of regression used when the dependent variable is binary or dichotomous (i.e., it has two possible outcomes, such as 0 or 1). It is one of several techniques that belong to the family of generalized linear models.

In probit regression, the probability that an event occurs (i.e., the dependent variable equals 1) is modeled as a function of the cumulative normal distribution. The model is expressed as:

$$\pi = \Phi(\mathbf{x}'\boldsymbol{\beta})$$

Here, Φ denotes the cumulative distribution function (CDF) of the standard normal distribution, $\mathbf{x}'$ represents the vector of independent variables, and $\boldsymbol{\beta}$ denotes the vector of coefficients.

The probit model is particularly useful when the underlying latent variable (the variable we cannot directly observe) is assumed to follow a normal distribution. This makes probit regression an attractive alternative to logistic regression (or logit regression), especially in cases where the assumptions of the logistic model are not appropriate.

Determining the Better Model: Logit vs. Probit

Both the logit and probit models are widely used for binary response data, but they have some differences:

Mathematical Foundation

- **Logit Model**: Uses the logistic function to model the probability that the dependent variable equals 1. The logit model's link function is the log of the odds ratio:

$$\pi = \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}'\boldsymbol{\beta})}$$

- **Probit Model**: Uses the cumulative normal distribution function to model the same probability. The link function in probit regression is based on the inverse of the standard normal CDF:

$$\pi = \Phi(\mathbf{x}'\boldsymbol{\beta})$$

Interpretation of Coefficients

- **Logit Model**: The coefficients represent the change in the log odds of the dependent variable per unit change in the predictor variable.
- **Probit Model**: The coefficients represent the change in the z-score (standard normal) of the latent variable.

When to Use Which

- **Probit**: Often preferred when the assumption of a normally distributed error term in the latent variable is reasonable. This model is commonly used in fields like economics.
- **Logit**: More commonly used in many applied fields such as epidemiology and social sciences. It has easier interpretation, particularly with odds ratios.

Practical Differences

- **Logit and Probit Similarity**: Both models often produce very similar results because the cumulative normal and logistic distributions are close in shape, especially in the center. However, logit models tend to have slightly heavier tails.
- **Extreme Values Sensitivity**: The probit model is more sensitive to extreme values, given its reliance on the normal distribution. In contrast, the logistic model's heavy tails can better handle data with outliers.

Example to Compare the Two Models

Imagine you are analyzing the likelihood of a person buying a product based on their income. Suppose your dataset contains a variable X (income) and a binary outcome variable Y (buy or not buy).

- In a **logit model**, the odds of purchasing the product are modeled as a function of income.
- In a **probit model**, the z-score associated with purchasing the product is modeled instead.

If your data shows that people are either very likely or very unlikely to buy the product (with few cases near the decision boundary), the logit model might offer a better fit due to its broader tails. However, if the decision is more sensitive to smaller changes in income around a critical threshold, the probit model might provide a better fit.

Conclusion

In general, the choice between a logit and probit model often comes down to context and specific requirements of the analysis. Theoretically, both models are robust, and their practical differences are minimal when sample sizes are large. However, in cases with small samples or where the underlying distribution of the data is believed to be normal, the probit model may be preferred. On the other hand, for more general applications, particularly where odds ratios are a desired output, the logit model is often the better choice.

Effects of Multicollinearity on Least Squares Estimators

Answer

Given the regression model $y = \beta_1 x_1 + \beta_2 x_2 + \epsilon$, where x_1, x_2, and y are scaled to unit length, we can analyze the effect of multicollinearity between x_1 and x_2 on the least squares estimators of the regression coefficients.

1. The Normal Equations

The normal equations for the least squares estimator are given by:

$$(\mathbf{X}^\top \mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{y}$$

Here, $\mathbf{X}^\top \mathbf{X}$ is the correlation matrix:

$$\mathbf{X}^\top \mathbf{X} = \begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}$$

where r_{12} is the correlation between x_1 and x_2.

2. Inverse of $\mathbf{X}^\top \mathbf{X}$

The inverse of $\mathbf{X}^\top \mathbf{X}$ is given by:

$$\mathbf{C} = (\mathbf{X}^\top \mathbf{X})^{-1} = \frac{1}{1 - r_{12}^2} \begin{pmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{pmatrix}$$

3. Estimators of the Regression Coefficients

The least squares estimates for the regression coefficients β_1 and β_2 are:

$$\hat{\beta}_1 = \frac{1}{1 - r_{12}^2}(\beta_1 - r_{12}\beta_2)$$

$$\hat{\beta}_2 = \frac{1}{1 - r_{12}^2}(\beta_2 - r_{12}\beta_1)$$

4. Variances and Covariances

The variance of the estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ are:

$$\text{Var}(\hat{\beta}_1) = \sigma^2 \cdot \frac{1}{1 - r_{12}^2}$$

$$\text{Var}(\hat{\beta}_2) = \sigma^2 \cdot \frac{1}{1 - r_{12}^2}$$

The covariance between $\hat{\beta}_1$ and $\hat{\beta}_2$ is:

$$\text{Cov}(\hat{\beta}_1, \hat{\beta}_2) = -\sigma^2 \cdot \frac{r_{12}}{1 - r_{12}^2}$$

5. Effect of Strong Multicollinearity

When there is strong multicollinearity between x_1 and x_2, the correlation coefficient r_{12} is close to 1. As r_{12} approaches 1, the denominator $1 - r_{12}^2$ approaches 0. This causes both the variances $\text{Var}(\hat{\beta}_1)$ and $\text{Var}(\hat{\beta}_2)$ to become very large, leading to high variability in the estimates of β_1 and β_2.

Additionally, the covariance $\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$ also increases in magnitude, which indicates that the estimates are highly correlated.

6. Conclusion

Strong multicollinearity between x_1 and x_2 results in large variances and covariances for the least squares estimators. This leads to instability and unreliability in the estimated coefficients, making it difficult to draw precise conclusions about the relationship between the predictors and the response variable.

a4paper, total=170mm,257mm, left=20mm, top=20mm,

Analysis of Ridge Regression and Pure Shrinkage Estimators Your Name August 17, 2024

Question

The pure shrinkage estimator is defined as $\hat{\beta}_s = c\hat{\beta}$, where $0 \leq c \leq 1$ is a constant chosen by the analyst. $\hat{\beta}$ is the estimate of parameter β under the

usual setup. Describe the kind of shrinkage that this estimator introduces, and compare it with the shrinkage that results from ridge regression. Intuitively, which estimator seems preferable? *(Hint: the shrinkage is on the scale vs the location.)*

Answer

Pure Shrinkage Estimators

The pure shrinkage estimator, as defined by Stein (1960), is given by:

$$\hat{\beta}_s = c\hat{\beta}$$

where:

- $\hat{\beta}$ is the least squares estimate of the parameter β under the usual setup.
- c is a constant chosen by the analyst, with $0 \leq c \leq 1$.

Shrinkage Characteristics

1. Scale Shrinkage

The pure shrinkage estimator introduces **scale shrinkage**. It reduces the magnitude of all elements of $\hat{\beta}$ uniformly by a factor of c. This means that the estimated parameters are shrunk towards zero but maintain their relative distances and proportions to one another.

2. Ridge Regression Shrinkage

Ridge regression, on the other hand, introduces **location-based shrinkage**. The ridge estimate $\hat{\beta}_r$ is derived from the optimization problem:

$$\text{Minimize}\,(\beta - \hat{\beta})^T X^T X(\beta - \hat{\beta})$$

$$\text{subject to}\,\beta^T\beta \leq s^2$$

where:

- $\hat{\beta}_r = (X^T X + \lambda I)^{-1} X^T y$
- λ is the ridge penalty parameter.
- s^2 is the constraint on the sum of the squared coefficients.

In ridge regression, the amount of shrinkage depends on the collinearity of the predictors, and it introduces bias by pulling the coefficients towards the origin (zero). The shrinkage is more pronounced for coefficients associated with less important predictors.

Pure Shrinkage Estimator

- **Uniform Shrinkage**: All coefficients are shrunk by the same factor c, regardless of their significance or the degree of multicollinearity among predictors.
- **Simpler to Apply**: The shrinkage factor c is straightforward and easy to apply without additional matrix operations or optimization problems.
- **Potential Loss of Information**: Since all coefficients are shrunk uniformly, important predictors might be overly penalized, leading to a potential loss of important information.

Ridge Regression

- **Adaptive Shrinkage**: Ridge regression adaptively shrinks coefficients based on the amount of multicollinearity among predictors. It applies more shrinkage to less important or highly correlated predictors, which helps in maintaining the significance of the key predictors.
- **Regularization Parameter (λ)**: Requires choosing an optimal λ, often through cross-validation or other methods, which can be computationally expensive.
- **Reduction of Variance**: By introducing a bias, ridge regression reduces the variance of the estimates, which can lead to better predictive performance, especially in the presence of multicollinearity.

Intuitive Preference

Intuitively, the choice between the two estimators depends on the context and the specific goals of the analysis:

- **If the goal is to reduce the overall magnitude of the coefficients uniformly and simplicity is a priority**, the **pure shrinkage estimator** may be preferred.
- **If the goal is to handle multicollinearity and retain important predictors' influence more effectively**, **ridge regression** is often the better choice.

Given that multicollinearity is a common issue in regression analysis, **ridge regression is generally more preferable** due to its ability to balance bias and variance and its adaptive shrinkage mechanism, which tends to result in more robust and reliable models.

Conclusion

In conclusion, while both estimators introduce shrinkage to address different challenges in regression analysis, ridge regression offers a more sophisticated approach that takes into account the relationships among predictors, making it a more suitable choice for many practical applications.

article amsmath

Question:2022-6a

When do we use ridge regression? Does the correlation matrix give any indication of multicollinearity? Illustrate your answer.

Answer:

When to Use Ridge Regression:

Ridge regression is used when there is multicollinearity among the predictor variables in a multiple regression model. Multicollinearity occurs when two or more predictors are highly correlated, leading to inflated variances of the coefficient estimates, making them unstable and sensitive to small changes in the model. This instability can make the model less reliable.

High Multicollinearity: Ridge regression is particularly useful when the predictors are highly correlated with each other. It introduces a biasing parameter k (also known as the ridge parameter) to the regression model to shrink the coefficients, thereby reducing their variance.

Reducing Variance: As shown in the image, the mean square error (MSE) of the ridge estimator $\hat{\beta}_R$ consists of both variance and bias terms. By choosing an appropriate value of k, ridge regression reduces the variance more than it increases the bias, potentially leading to a lower MSE compared to ordinary least squares (OLS) estimates.

Correlation Matrix and Multicollinearity:

Indication of Multicollinearity: The correlation matrix can give an indication of multicollinearity. If two or more predictors have a high pairwise correlation (close to +1 or -1), this is a sign of multicollinearity. However, it's important to note that the correlation matrix only shows pairwise correlations and might not capture the full extent of multicollinearity, which could be more complex involving several variables at once.

Example Illustration:

Let's consider a simple example with three predictors X_1, X_2, and X_3. Suppose the correlation matrix is as follows:

$$\begin{pmatrix} 1 & 0.9 & 0.8 \\ 0.9 & 1 & 0.85 \\ 0.8 & 0.85 & 1 \end{pmatrix}$$

Here, the high correlations between X_1 and X_2 (0.9), and X_1 and X_3 (0.8) indicate strong multicollinearity. In this situation, using OLS might lead to unreliable coefficient estimates. Applying ridge regression would introduce a penalty proportional to the size of the coefficients, helping to reduce the impact of multicollinearity by shrinking the coefficients, thereby improving the model's stability.

Conclusion: In summary, ridge regression is particularly useful when you have multicollinearity in your predictors. The correlation matrix can help detect this multicollinearity, but ridge regression addresses it directly by stabilizing the coefficient estimates.

article amsmath

Question: 2022 7(a,b)

1. Write a short note on nonlinear regression models.
2. Illustrate how the linearization can be accomplished by a Taylor series expansion of the nonlinear regression function, followed by an iteration method of parameter estimation.

Answer:

(a) Nonlinear Regression Models

A **nonlinear regression model** is any statistical model where the relationship between the response variable and one or more predictor variables is represented by a nonlinear function in the model parameters. Unlike linear regression models, which assume that the response variable is a linear combination of the predictors, nonlinear regression models can capture more complex relationships where the effect of the predictors on the response variable changes at different levels of the predictors.

Key Features of Nonlinear Regression Models:

- **Nonlinearity in Parameters:** The model is nonlinear in its parameters, meaning at least one of the parameters appears in a nonlinear way in the model equation. For example, the model $y = \theta_1 e^{\theta_2 x} + \epsilon$ is nonlinear in parameters θ_1 and θ_2.
- **Expectation Function:** The expectation of the response variable y is given by $\mathbb{E}(y) = f(x, \theta)$, where $f(x, \theta)$ is a nonlinear function of the parameters θ.
- **Complexity:** Nonlinear models often arise when the underlying relationship is complex and cannot be adequately captured by a linear model. For instance, relationships derived from physical or biological processes governed by differential equations often lead to nonlinear models.

(b) Linearization by Taylor Series Expansion and Iteration for Parameter Estimation

Nonlinear regression models can be challenging to solve because the parameter estimates cannot be obtained using simple algebraic methods like those in linear regression. Instead, we often linearize the nonlinear model around initial parameter estimates using a Taylor series expansion and then apply an iterative method to refine the estimates.

Linearization via Taylor Series Expansion:

Suppose the nonlinear regression model is given by:

$$y = f(x, \theta) + \epsilon$$

where θ is a vector of unknown parameters.

To linearize the model, we expand $f(x, \theta)$ about some initial guess $\theta^{(0)}$ using a first-order Taylor series approximation:

$$f(x, \theta) \approx f(x, \theta^{(0)}) + \sum_{i=1}^{p} \frac{\partial f(x, \theta^{(0)})}{\partial \theta_i} (\theta_i - \theta_i^{(0)})$$

where $\frac{\partial f(x, \theta^{(0)})}{\partial \theta_i}$ is the partial derivative of $f(x, \theta)$ with respect to θ_i evaluated at $\theta^{(0)}$.

This approximation gives a linear model in the deviations $\theta - \theta^{(0)}$ from the initial parameter values:

$$y \approx f(x, \theta^{(0)}) + \sum_{i=1}^{p} J_i (\theta_i - \theta_i^{(0)}) + \epsilon$$

where $J_i = \frac{\partial f(x, \theta^{(0)})}{\partial \theta_i}$ is the Jacobian matrix.

Iterative Method for Parameter Estimation:

1. **Initialization:** Start with an initial guess $\theta^{(0)}$ for the parameters.
2. **Linearization:** Linearize the model around $\theta^{(0)}$ using the Taylor series expansion.
3. **Solve the Linearized System:** Solve the resulting linear equation for the updates $\Delta\theta$ to the parameters:

   $$\Delta\theta = (\mathbf{J}^\top \mathbf{J})^{-1} \mathbf{J}^\top (y - f(x, \theta^{(0)}))$$

   where $\mathbf{J}$ is the Jacobian matrix.
4. **Update the Parameters:** Update the parameter estimates:

   $$\theta^{(1)} = \theta^{(0)} + \Delta\theta$$

5. **Iteration:** Repeat the process until the changes in parameter estimates $\Delta\theta$ become negligibly small, indicating convergence.

This method, known as the **Gauss-Newton algorithm** (or a similar iterative method like **Levenberg-Marquardt**), is widely used for nonlinear regression.

In summary, nonlinear regression models extend the flexibility of regression analysis to capture more complex relationships between variables, but they require more sophisticated methods like Taylor series linearization and iterative algorithms to estimate parameters.

article amsmath

Question:2022 1(a)

1. Show that $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2 \bar{x}}{S_{xx}}$.
2. Show that $\text{Cov}(y, \hat{\beta}_1) = 0$.

Answer:

(a) Show that $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2 \bar{x}}{S_{xx}}$

Context and Definitions:

In simple linear regression, we have:

$$y = \beta_0 + \beta_1 x + \epsilon$$

where:

- $\hat{\beta}_0$ is the estimated intercept.
- $\hat{\beta}_1$ is the estimated slope.
- $\bar{x}$ is the mean of the x values.
- $S_{xx} = \sum(x_i - \bar{x})^2$ is the sum of squares of the deviations of x from its mean.
- σ^2 is the variance of the errors ϵ.

Step-by-Step Explanation:

1. **Understanding $\hat{\beta}_0$ and $\hat{\beta}_1$:**
   - The slope $\hat{\beta}_1$ is given by:

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

   This simplifies to:

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})y_i}{S_{xx}}$$

- The intercept $\hat{\beta}_0$ is given by:

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

where $\bar{y}$ is the mean of the y values.

2. **Covariance between $\hat{\beta}_0$ and $\hat{\beta}_1$:**

- We want to find $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1)$.
- Substituting $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ into the covariance formula:

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \text{Cov}(\bar{y} - \hat{\beta}_1 \bar{x}, \hat{\beta}_1)$$

This can be expanded as:

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \text{Cov}(\bar{y}, \hat{\beta}_1) - \bar{x} \cdot \text{Var}(\hat{\beta}_1)$$

- Since $\bar{y}$ is a constant and not dependent on the sample, $\text{Cov}(\bar{y}, \hat{\beta}_1) = 0$. Therefore:

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\bar{x} \cdot \text{Var}(\hat{\beta}_1)$$

3. **Variance of $\hat{\beta}_1$:**

- The variance of $\hat{\beta}_1$ is given by:

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

- Substituting this into the covariance formula:

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\bar{x} \cdot \frac{\sigma^2}{S_{xx}}$$

- So, the final result is:

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2 \bar{x}}{S_{xx}}$$

(b) Show that $\text{Cov}(y, \hat{\beta}_1) = 0$

Context and Definitions:

We want to find the covariance between the actual response variable y and the slope estimate $\hat{\beta}_1$.

Step-by-Step Explanation:

1. **Expressing y in terms of x and ϵ:**

- We know that $y = \beta_0 + \beta_1 x + \epsilon$.
- For simplicity, let's focus on one observation $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.

2. **Covariance between y and $\hat{\beta}_1$:**

   - Since $\hat{\beta}_1$ is calculated based on the relationship between x and y, and y depends linearly on x, there might be an assumption that there's some correlation.
   - However, $\hat{\beta}_1$ is an estimate of the slope across all data points and is not directly influenced by any single y_i.
   - To calculate $\text{Cov}(y_i, \hat{\beta}_1)$, recognize that:

$$\text{Cov}(y_i, \hat{\beta}_1) = \text{Cov}(\beta_0 + \beta_1 x_i + \epsilon_i, \hat{\beta}_1)$$

   - Since β_0 and $\beta_1 x_i$ are constants when conditioned on the sample, their covariance with $\hat{\beta}_1$ is zero.
   - The only term that could contribute to covariance is ϵ_i. But because ϵ_i is uncorrelated with the predictor x (and hence with $\hat{\beta}_1$), the covariance between y and $\hat{\beta}_1$ is zero.

3. **Conclusion:**

   - Thus, $\text{Cov}(y, \hat{\beta}_1) = 0$.

This means that there is no linear relationship between the actual values of y and the slope $\hat{\beta}_1$ in terms of covariance.

article amsmath

Question:2022 1(b)

1. Consider the multiple regression model $y = X\beta + \epsilon$. Illustrate the significance of the hat matrix in multiple regression.
2. Prove that the matrices H and $I - H$ are idempotent.
3. Show that in the multiple linear regression model, $\text{Var}(\hat{Y}) = \sigma^2 H$.

Answer:

1. Significance of the Hat Matrix in Multiple Regression

In multiple regression, the model is expressed as:

$$y = X\beta + \epsilon$$

where:

- y is the vector of observed values.
- X is the design matrix (containing the values of the independent variables).

- β is the vector of regression coefficients.
- ϵ is the vector of errors.

The **hat matrix** H is defined as:

$$H = X(X'X)^{-1}X'$$

The hat matrix is important because it maps the observed values of the response variable y to the fitted values $\hat{y}$. Specifically:

$$\hat{y} = Hy$$

The name "hat matrix" comes from its role in putting a "hat" on y to produce $\hat{y}$, the predicted values.

Significance:

1. **Projection Operator:** The hat matrix is a projection operator that projects the vector of observed values y onto the column space of X. This means that H transforms the observed data into the best linear approximation given by the model.
2. **Influence of Data Points:** The diagonal elements of H indicate the influence or leverage of each observed data point on its corresponding fitted value $\hat{y}_i$. High leverage points have a large influence on the model's predictions.
3. **Model Diagnostics:** The hat matrix is crucial in diagnostic measures such as identifying outliers and influential data points. It helps in understanding how much each data point contributes to the overall fit of the model.

2. Idempotency of the Hat Matrix H and $I - H$

A matrix A is said to be idempotent if $A^2 = A$.

Idempotency of H:

To prove that H is idempotent, we compute H^2:

$$H^2 = H \cdot H$$

Substitute the expression for H:

$$H^2 = \left[X(X'X)^{-1}X'\right] \cdot \left[X(X'X)^{-1}X'\right]$$

Using matrix properties, particularly the associative property:

$$H^2 = X(X'X)^{-1}(X'X)(X'X)^{-1}X'$$

Since $(X'X)(X'X)^{-1} = I$ (the identity matrix):

$$H^2 = X(X'X)^{-1}X' = H$$

Thus, H is idempotent.

Idempotency of $I - H$:

Next, we prove that $I - H$ is also idempotent.

Let $A = I - H$. We need to show that $A^2 = A$:

$$A^2 = (I - H)(I - H)$$

Expanding the product:

$$A^2 = I - 2H + H^2$$

Since $H^2 = H$:

$$A^2 = I - 2H + H = I - H$$

Thus, $A^2 = I - H$, so $I - H$ is idempotent.

3. Variance of the Predicted Values $\hat{y}$

In multiple regression, the variance of the predicted values $\hat{y}$ is given by:

$$\text{Var}(\hat{y}) = \text{Var}(Hy)$$

Since $y = X\beta + \epsilon$ and $\text{Var}(y) = \text{Var}(\epsilon) = \sigma^2 I$ (assuming homoscedasticity and uncorrelated errors):

$$\text{Var}(\hat{y}) = \text{Var}(Hy) = H \cdot \text{Var}(y) \cdot H'$$

Substitute $\text{Var}(y) = \sigma^2 I$:

$$\text{Var}(\hat{y}) = H \cdot \sigma^2 I \cdot H'$$

Because H is symmetric ($H = H'$):

$$\text{Var}(\hat{y}) = \sigma^2 H^2$$

And since H is idempotent ($H^2 = H$):

$$\text{Var}(\hat{y}) = \sigma^2 H$$

Thus, in the multiple linear regression model:

$$\text{Var}(\hat{y}) = \sigma^2 H$$

This equation shows that the variance of the predicted values depends on the hat matrix H, which reflects the influence of the design matrix X on the fitted

values. The diagonal elements of $\sigma^2 H$ provide the variances of the individual predicted values $\hat{y}_i$.

article amsmath amsfonts

Question 2022-2

Suppose we wish to find the least-square estimator of β in the multiple regression model $y = X\beta + \epsilon$ subject to a set of equality constraints on β, say $T\beta = c$. Show that the estimator is

$$\tilde{\beta} = \hat{\beta} + (X'X)^{-1}T'[T(X'X)^{-1}T']^{-1}(c - T\hat{\beta}),$$

where $\hat{\beta} = (X'X)^{-1}X'y$. Discuss situations in which this constrained estimator might be appropriate.

Answer

To derive the least-squares estimator $\tilde{\beta}$ for β in a multiple regression model $y = X\beta + \epsilon$ subject to the equality constraints $T\beta = c$, follow these steps:

1. Formulate the Constrained Least Squares Problem

The objective is to minimize the residual sum of squares subject to the constraint:

$$\text{Minimize } \|y - X\beta\|^2 \text{ subject to } T\beta = c.$$

2. Lagrange Multiplier Approach

To solve this, we use the method of Lagrange multipliers. Define the Lagrangian function:

$$\mathcal{L}(\beta, \lambda) = \|y - X\beta\|^2 + \lambda'(T\beta - c),$$

where λ is the vector of Lagrange multipliers.

3. Take the Derivatives

First, differentiate $\mathcal{L}$ with respect to β and set it to zero:

$$\frac{\partial \mathcal{L}}{\partial \beta} = -2X'(y - X\beta) + T'\lambda = 0.$$

Rearrange to get:

$$X'(y - X\beta) = \frac{1}{2}T'\lambda.$$

Next, differentiate $\mathcal{L}$ with respect to λ and set it to zero:

$$\frac{\partial \mathcal{L}}{\partial \lambda} = T\beta - c = 0.$$

4. Solve for β and λ

From the derivative with respect to β, we have:

$$X'y - X'X\beta = \frac{1}{2}T'\lambda.$$

Substitute $\frac{1}{2}T'\lambda = X'y - X'X\beta$ into the derivative with respect to λ:

$$T\beta = c.$$

We now have two equations:

1. $X'X\beta + T'\lambda = X'y$
2. $T\beta = c$

First, solve for λ from the first equation:

$$T'\lambda = X'y - X'X\beta.$$

Substitute $T\beta = c$ into this equation to find:

$$T'\lambda = X'y - X'X\beta$$

$$\lambda = [T(X'X)^{-1}T']^{-1}(c - T\beta).$$

Substitute λ back into the original equation for β:

$$X'X\beta = X'y - T[T(X'X)^{-1}T']^{-1}(c - T\beta).$$

Solve for β to obtain:

$$\beta = (X'X)^{-1}X'y + (X'X)^{-1}T'[T(X'X)^{-1}T']^{-1}(c - T(X'X)^{-1}X'y).$$

Let $\hat{\beta} = (X'X)^{-1}X'y$ be the unconstrained estimator. Then:

$$\tilde{\beta} = \hat{\beta} + (X'X)^{-1}T'[T(X'X)^{-1}T']^{-1}(c - T\hat{\beta}).$$

Discussion of Situations for Constrained Estimator

The constrained least-squares estimator $\tilde{\beta}$ is particularly useful in scenarios where:

1. **Prior Knowledge or Restrictions**: There are known restrictions or prior knowledge about the parameters (e.g., some parameters should be zero, sum to a specific value, or lie within certain bounds).
2. **Model Simplification**: The constraints help in simplifying or reducing the complexity of the model by imposing structure.
3. **Overfitting Prevention**: Constraints can help prevent overfitting by limiting the degrees of freedom of the model.

4. **Economic or Theoretical Considerations**: In some applications, economic or theoretical considerations may dictate specific relationships between parameters.

This constrained approach allows for incorporating additional information or requirements into the regression analysis, making the model more robust and applicable to real-world situations.

article amsmath amsfonts

Question 2023-1-a

Consider the simple linear regression model, $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ with $E(\epsilon_i) = 0$, $\text{var}(\epsilon_i) = \sigma^2$, and ϵ_i's are uncorrelated, $i \in \{1, 2, \ldots, n\}$.

(a) Show that $\text{SSR} = \hat{\beta}_1 S_{xy} = \hat{\beta}_1^2 S_{xx}$.

Answer:

1. **Sum of Squares due to Regression (SSR):**

In a simple linear regression model, $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, the SSR quantifies the variability in y that is explained by the regression on x.

SSR can be expressed as:

$$\text{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

where $\hat{y}_i$ is the predicted value of y_i and $\bar{y}$ is the mean of the observed y.

2. **Estimated Slope $\hat{\beta}_1$:**

The estimated slope $\hat{\beta}_1$ is given by:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

where S_{xy} is the covariance of x and y, and S_{xx} is the variance of x.

3. **Relation between SSR and $\hat{\beta}_1$:**

The SSR can be rewritten using the formula for $\hat{\beta}_1$:

$$\text{SSR} = \hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x})(\hat{y}_i - \bar{y})$$

Since $\hat{y}_i - \bar{y} = \hat{\beta}_1 (x_i - \bar{x})$, we get:

$$\text{SSR} = \hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x})(\hat{\beta}_1 (x_i - \bar{x})) = \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\text{SSR} = \hat{\beta}_1^2 S_{xx}$$

Thus, $\text{SSR} = \hat{\beta}_1 S_{xy}$ and $\text{SSR} = \hat{\beta}_1^2 S_{xx}$ are equivalent.

Question 2023-1-b

(b) Show that $E(\text{SSR}) = \sigma^2 + \beta_1^2 S_{xx}$ and $E(\text{MSE}) = \sigma^2$.
Answer:
1. **Expected Value of SSR:**
Recall that:

$$\text{SSR} = \hat{\beta}_1^2 S_{xx}$$

Since $\hat{\beta}_1$ is an unbiased estimator of β_1, we have:

$$E(\hat{\beta}_1^2) = \text{Var}(\hat{\beta}_1) + \beta_1^2$$

The variance of $\hat{\beta}_1$ is $\frac{\sigma^2}{S_{xx}}$. Therefore:

$$E(\hat{\beta}_1^2) = \frac{\sigma^2}{S_{xx}} + \beta_1^2$$

Substituting this into the expression for SSR:

$$E(\text{SSR}) = E(\hat{\beta}_1^2 S_{xx}) = S_{xx}\left(\frac{\sigma^2}{S_{xx}} + \beta_1^2\right) = \sigma^2 + \beta_1^2 S_{xx}$$

2. **Expected Value of MSE:**
The Mean Squared Error is:

$$\text{MSE} = \frac{\text{SSR}}{n-2}$$

Therefore:

$$E(\text{MSE}) = \frac{E(\text{SSR})}{n-2} = \frac{\sigma^2 + \beta_1^2 S_{xx}}{n-2}$$

Note that the expected value of the MSE should ideally be σ^2, not $\frac{\sigma^2+\beta_1^2 S_{xx}}{n-2}$. In practice, the term $\frac{\beta_1^2 S_{xx}}{n-2}$ represents the bias due to the estimation of β_1, and the MSE is an unbiased estimator of σ^2 when adjusted properly.

Question 2023-1-c

(c) Consider the maximum-likelihood estimator $\tilde{\sigma}^2$ of σ^2. Find the bias in $\tilde{\sigma}^2$.
Answer:
1. **Definition of $\tilde{\sigma}^2$:**
$\tilde{\sigma}^2$ is the maximum-likelihood estimator of σ^2, calculated as:

$$\tilde{\sigma}^2 = \frac{\text{SSR}}{n-2}$$

2. **Bias Calculation:**

Since $E(\text{SSR}) = \sigma^2 + \beta_1^2 S_{xx}$, the bias of $\tilde{\sigma}^2$ is:

$$\text{Bias}(\tilde{\sigma}^2) = E(\tilde{\sigma}^2) - \sigma^2$$

$$\text{Bias}(\tilde{\sigma}^2) = \frac{E(\text{SSR})}{n-2} - \sigma^2$$

$$\text{Bias}(\tilde{\sigma}^2) = \frac{\sigma^2 + \beta_1^2 S_{xx}}{n-2} - \sigma^2$$

$$\text{Bias}(\tilde{\sigma}^2) = \frac{\beta_1^2 S_{xx}}{n-2}$$

Therefore, $\tilde{\sigma}^2$ is biased by $\frac{\beta_1^2 S_{xx}}{n-2}$.

Question 2023-1-d

(d) Prove that the maximum value of R^2 is less than 1 if the data contain repeated (different) observations on y at the same value of x.

Answer:

1. **Formula for R^2:**

R^2 is given by:

$$R^2 = \frac{\text{SSR}}{\text{SST}}$$

where SST is the Total Sum of Squares and SSR is the Sum of Squares due to Regression.

2. **When y values are repeated for the same x:**

If there are repeated y values for the same x, it means that there is no variability in y for those particular x values. In this case, the model will not perfectly explain the variability in y because the regression line cannot account for the variability between the repeated y values.

3. **Implication for R^2:**

Because y values are the same for some x values, the regression line will not be able to fit those data points perfectly. This means that SSR will be less than SST, leading to:

$$R^2 < 1$$

The maximum value of R^2 cannot be 1 because there is unexplained variability in y for those x values where y values are repeated.

In summary, if there are repeated observations of y for the same x, the maximum R^2 will be less than 1 due to the lack of perfect fit in the model for these repeated values.

article amsmath amsfonts

Question 2023-2-a

Consider the multiple regression model $y = X\beta + \epsilon$.

(a) Find the expression for the least-squares estimator $\hat{\beta}$ of β. Show that the least-squares estimator can be written as $\hat{\beta} = \beta + R\epsilon$, where $R = (X^\top X)^{-1} X^\top$.

Answer:

1. **Least-Squares Estimator:**

In the multiple regression model $y = X\beta + \epsilon$, the goal is to find the least-squares estimator $\hat{\beta}$ of β.

The least-squares estimator $\hat{\beta}$ minimizes the sum of squared residuals:

$$\text{SSR} = \|y - X\beta\|^2$$

To find $\hat{\beta}$, take the derivative of SSR with respect to β and set it to zero:

$$\frac{\partial}{\partial \beta} \left((y - X\beta)^\top (y - X\beta) \right) = -2X^\top (y - X\beta)$$

Setting the derivative equal to zero:

$$-2X^\top (y - X\beta) = 0$$
$$X^\top (y - X\hat{\beta}) = 0$$
$$X^\top y = X^\top X\hat{\beta}$$
$$\hat{\beta} = (X^\top X)^{-1} X^\top y$$

So, the least-squares estimator $\hat{\beta}$ is:

$$\hat{\beta} = (X^\top X)^{-1} X^\top y$$

2. **Showing $\hat{\beta} = \beta + R\epsilon$:**

Define R as:

$$R = (X^\top X)^{-1} X^\top$$

Substitute the model $y = X\beta + \epsilon$ into the formula for $\hat{\beta}$:

$$\hat{\beta} = (X^\top X)^{-1} X^\top (X\beta + \epsilon)$$
$$\hat{\beta} = (X^\top X)^{-1} X^\top X\beta + (X^\top X)^{-1} X^\top \epsilon$$
$$\hat{\beta} = \beta + (X^\top X)^{-1} X^\top \epsilon$$

We can see that:

$$\hat{\beta} = \beta + R\epsilon$$

where $R = (X^\top X)^{-1} X^\top$.

Question 2023-2-b

(b) Consider a correctly specified regression model with p terms, including the intercept. Make the usual assumptions about ϵ. Prove that:

$$\sum_{i=1}^{n} \text{Var}(\hat{y}_i) = p\sigma^2$$

Answer:

1. **Variance of $\hat{y}_i$:**

The predicted value $\hat{y}_i$ for each observation i is given by:

$$\hat{y}_i = x_i^\top \hat{\beta}$$

where x_i is the i-th row of the matrix X. Since $\hat{\beta} = \beta + R\epsilon$, we have:

$$\hat{y}_i = x_i^\top (\beta + R\epsilon) = x_i^\top \beta + x_i^\top R\epsilon$$

Because $x_i^\top \beta$ is a constant and does not vary, we focus on $x_i^\top R\epsilon$ for the variance:

$$\text{Var}(\hat{y}_i) = \text{Var}(x_i^\top R\epsilon)$$

Note that:

$$\text{Var}(\hat{y}_i) = x_i^\top R\text{Var}(\epsilon)R^\top x_i$$

Since $\text{Var}(\epsilon) = \sigma^2 I$, where I is the identity matrix:

$$\text{Var}(\hat{y}_i) = x_i^\top R\sigma^2 I R^\top x_i$$

$$\text{Var}(\hat{y}_i) = \sigma^2 x_i^\top R R^\top x_i$$

By the properties of R (i.e., $RR^\top = X(X^\top X)^{-1}X^\top$):

$$x_i^\top RR^\top x_i = x_i^\top (X(X^\top X)^{-1}X^\top)x_i$$

$$\text{Var}(\hat{y}_i) = \sigma^2 \text{Cov}(x_i, x_i)$$

Since the sum of variances of predicted values is the sum of the diagonal elements of the covariance matrix of $\hat{y}$, and there are p terms in the model, each contributing σ^2 to the variance:

$$\sum_{i=1}^{n} \text{Var}(\hat{y}_i) = p\sigma^2$$

This follows from the fact that the trace of $X(X^\top X)^{-1}X^\top$ equals p, where p is the number of parameters in the model.

article amsmath

Question 2023-3-a

Write a short note on PRESS Residual and PRESS statistic. Give mathematical expression if any.

Answer:

PRESS Residuals

Definition: - **PRESS** stands for **Prediction Sum of Squares**. PRESS residuals are a type of residual used to assess the predictive performance of a regression model. They are specifically used in the context of **leave-one-out cross-validation**.

How PRESS Residuals are Computed: 1. **Leave-One-Out Cross-Validation:** - In this method, we fit the regression model on all but one of the data points and then predict the left-out point. This process is repeated for each data point.

2. **PRESS Residual:** - For each data point i, the PRESS residual $e_i^{(i)}$ is defined as:

$$e_i^{(i)} = y_i - \hat{y}_i^{(i)}$$

where: - y_i is the observed value of the response variable for the i-th data point. - $\hat{y}_i^{(i)}$ is the predicted value for the i-th data point when it is left out from the fitting of the model. This prediction is based on the model trained on the remaining $n-1$ data points.

PRESS Statistic

Definition: - The PRESS statistic is a measure used to evaluate the predictive accuracy of a regression model. It summarizes the PRESS residuals across all data points.

Mathematical Expression: - The PRESS statistic is given by the sum of squared PRESS residuals:

$$\text{PRESS} = \sum_{i=1}^{n} (e_i^{(i)})^2$$

where: - $e_i^{(i)}$ is the PRESS residual for the i-th data point, as defined earlier.

Interpretation: - The PRESS statistic provides an indication of how well the regression model predicts new data points. A smaller PRESS statistic suggests that the model has good predictive accuracy because the predictions are close to the actual values when each data point is left out. - Conversely, a larger PRESS statistic indicates that the model may not be performing well in terms of prediction, as the residuals are larger when data points are left out.

Usefulness: - The PRESS statistic is particularly useful in diagnosing model performance and comparing different models. It helps in understanding the model's ability to generalize to new data, which is crucial for assessing the robustness of the model.

In summary, PRESS residuals and the PRESS statistic are important tools for evaluating the predictive performance of a regression model by examining

how well the model can predict data points that were not used in fitting the model.

article amsmath

Question 2023-3-b

Diagnose if the following statement is True/False with suitable explanation(s). A studentized residual r_i is just a deleted residual d_i divided by its estimated standard deviation $s(d_i)$ (first formula). This turns out to be equivalent to the ordinary residual divided by a factor that includes the mean squared error based on the estimated model with the i-th observation deleted, $MSE^{(-i)}$, and the leverage, h_{ii} (second formula). In other words,

$$r_i = \frac{d_i}{s(d_i)} = \frac{e_i}{\sqrt{MSE^{(-i)}(1 - h_{ii})}}$$

Hence (or otherwise), explain that the studentized residual for a given data point depends not only on the ordinary residual but also on the size of the mean squared error (MSE) and the leverage.

Answer:

Studentized Residuals

Definition: - **Studentized residual** is a type of residual that has been scaled to account for the variability of the residual. It helps to standardize residuals so that they can be more easily compared across different observations.

Key Concepts: 1. **Residual (e_i):** The difference between the observed value and the predicted value for the i-th observation, $e_i = y_i - \hat{y}_i$.

2. **Deleted Residual (d_i):** This is the residual obtained when the i-th observation is left out from the regression model fitting. It's the difference between the observed value y_i and the prediction $\hat{y}_i^{(-i)}$, where $\hat{y}_i^{(-i)}$ is the predicted value when the i-th observation is excluded.

3. **Estimated Standard Deviation of the Deleted Residual ($s(d_i)$):** This is the standard error associated with the deleted residual d_i.

4. **Mean Squared Error (MSE):** In the context of regression, MSE is an estimate of the variance of the residuals from the model. When the i-th observation is excluded, we get an MSE value, denoted $MSE^{(-i)}$, based on the model fitted to the remaining observations.

5. **Leverage (h_{ii}):** This is a measure of how far an observation is from the center of the predictor space. It ranges from 0 to 1 and is derived from the hat matrix H, where h_{ii} is the i-th diagonal element of H.

Statement Analysis: The statement suggests that the studentized residual r_i can be expressed in two equivalent ways: 1. **First Formula:**

$$r_i = \frac{d_i}{s(d_i)}$$

where d_i is the deleted residual and $s(d_i)$ is its estimated standard deviation.

2. **Second Formula:**

$$r_i = \frac{e_i}{\sqrt{MSE^{(-i)}(1 - h_{ii})}}$$

where e_i is the ordinary residual, $MSE^{(-i)}$ is the mean squared error when the i-th observation is deleted, and h_{ii} is the leverage for the i-th observation.

Explanation and Diagnosis:

1. **Equivalence of Formulas:** The studentized residual r_i is indeed defined as:

$$r_i = \frac{d_i}{s(d_i)}$$

where d_i is the residual when the i-th observation is omitted, and $s(d_i)$ is its estimated standard deviation.

It can also be expressed using:

$$r_i = \frac{e_i}{\sqrt{MSE^{(-i)}(1 - h_{ii})}}$$

To understand this equivalence, note that the standard deviation of the deleted residual $s(d_i)$ can be expressed in terms of $MSE^{(-i)}$ and leverage h_{ii}.

2. **Mathematical Derivation:** The deleted residual d_i can be related to the ordinary residual e_i and the leverage h_{ii} as follows:

$$d_i = \frac{e_i}{1 - h_{ii}}$$

Therefore, the variance of d_i is:

$$\text{Var}(d_i) = \frac{\text{Var}(e_i)}{(1 - h_{ii})^2}$$

The estimated variance of e_i is $MSE^{(-i)}$, so:

$$s^2(d_i) = \frac{MSE^{(-i)}}{(1 - h_{ii})^2}$$

Thus:

$$s(d_i) = \frac{\sqrt{MSE^{(-i)}}}{1 - h_{ii}}$$

and:

$$r_i = \frac{d_i}{s(d_i)} = \frac{\frac{e_i}{1-h_{ii}}}{\frac{\sqrt{MSE^{(-i)}}}{1-h_{ii}}} = \frac{e_i}{\sqrt{MSE^{(-i)}(1 - h_{ii})}}$$

3. **Conclusion:** The statement is **True**. The studentized residual reflects both the ordinary residual and the effects of MSE and leverage. This ensures that the residuals are standardized in a way that accounts for the variability in the data and the influence of each observation on the model.

article amsmath

Question 2023-4-a(i)

Briefly describe the principle of Logistic Regression. Give mathematical expression if any.

Answer:

Principle of Logistic Regression

Objective: - Logistic regression is a statistical method used to model the probability of a binary outcome based on one or more predictor variables. It is often used when the response variable is categorical with two possible outcomes, such as "yes" or "no," "success" or "failure," etc.

How It Works: 1. **Binary Outcome:** - The response variable in logistic regression is binary, meaning it takes on one of two possible values. For example, it might indicate whether a customer will buy a product (1 for yes, 0 for no).

2. **Predictors:** - Logistic regression uses predictor variables (also known as features or independent variables) to predict the probability of the binary outcome. These predictors can be continuous or categorical.

3. **Probability Modeling:** - The core idea is to model the probability that a given observation falls into one of the two categories. Specifically, logistic regression estimates the probability $P(Y = 1 \mid X)$, where Y is the binary outcome and X represents the predictor variables.

4. **Logit Function:** - Logistic regression uses the logit function to link the probability of the outcome to the predictor variables. The logit function is the natural logarithm of the odds of the outcome occurring. Mathematically, the logit function is defined as:

$$\text{logit}(P) = \ln\left(\frac{P}{1-P}\right)$$

where P is the probability of the outcome being 1.

5. **Model Equation:** - The logistic regression model expresses the logit of the probability as a linear combination of the predictor variables. The model is given by:

$$\text{logit}(P) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

where: - β_0 is the intercept term. - $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients for the predictor variables $X_1, X_2, \ldots, X_p$.

6. **Probability Prediction:** - To obtain the probability P from the logit function, the inverse of the logit function, known as the logistic function, is used. The logistic function is:

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}}$$

Here, e is the base of the natural logarithm, and P represents the probability of the outcome being 1.

Key Points: - **Interpretation:** The coefficients $\beta_1, \beta_2, \ldots, \beta_p$ represent the change in the log odds of the outcome for a one-unit change in the corresponding predictor variable. - **Estimation:** The coefficients are estimated using

maximum likelihood estimation, which finds the values of β that maximize the likelihood of observing the given data. - **Application:** Logistic regression is widely used in various fields such as medicine (e.g., predicting the presence of a disease), finance (e.g., credit scoring), and social sciences (e.g., predicting election outcomes).

In summary, logistic regression models the probability of a binary outcome based on predictor variables by using the logit function to link the predictors to the probability of the outcome occurring. The model is expressed in terms of a logistic function that transforms the linear combination of predictors into a probability.

article amsmath

Question 2023-5-a(i)

Explain how the following problematic scenario can be handled in light of Ridge regression:

(i) The Least Squares (LS) estimate depends upon $(X'X)^{-1}$. We would have problems in computing β_{LS} if $X'X$ were singular or nearly singular.

Answer:

In Least Squares (LS) regression, the coefficient estimates are computed using the formula:

$$\beta_{\text{LS}} = (X'X)^{-1}X'y$$

If the matrix $X'X$ is singular (i.e., not invertible) or nearly singular, then $(X'X)^{-1}$ does not exist or is highly unstable. This issue often arises when there is multicollinearity among the predictor variables, meaning some predictors are highly correlated with each other.

Ridge regression addresses this problem by introducing a regularization term to the LS estimation. The Ridge regression estimator is given by:

$$\beta_{\text{Ridge}} = (X'X + \lambda I)^{-1}X'y$$

where λ is a positive constant (regularization parameter) and I is the identity matrix. By adding λI to $X'X$, Ridge regression ensures that $X'X+\lambda I$ is always invertible, even if $X'X$ is singular or nearly singular. The regularization term stabilizes the inversion process.

Question 2023-5-a(ii)

Explain how the following problematic scenario can be handled in light of Ridge regression:

(ii) In the above case(s), small changes to the elements of X lead to large changes in $(X'X)^{-1}$, and the least squared estimator β_{LS} may provide a good fit to the training data, but it may not fit sufficiently well to the test data.

Answer:

When $X'X$ is nearly singular, small changes in the data X can cause large changes in the LS estimate β_{LS}. This results in high variance in the estimates, leading to a model that fits the training data well but performs poorly on new, unseen data (test data).

Ridge regression helps to mitigate this issue by adding a penalty term to the loss function. The Ridge regression loss function is:

$$L(\beta) = \|y - X\beta\|^2 + \lambda\|\beta\|^2$$

where $\|y - X\beta\|^2$ is the residual sum of squares (RSS), and $\|\beta\|^2$ is the sum of the squares of the coefficients (L2 norm).

The penalty term $\lambda\|\beta\|^2$ shrinks the coefficients towards zero, which reduces their variance. This regularization controls the sensitivity of the estimates to small changes in X, thereby improving the model's generalization to test data. The Ridge regression model is less likely to overfit the training data, leading to better performance on new data.

In summary: 1. **Singular $X'X$:** Ridge regression stabilizes the coefficient estimation by adding λI to $X'X$, ensuring invertibility and reducing instability. 2. **Sensitivity to Changes in X:** Ridge regression reduces the variance of the estimates by including a regularization term, improving the model's performance on test data.

6. (a) Compare between Least-Squared estimation in linear regression vs. Least-Squared estimation in nonlinear regression.

Introduction:

The method of Least Squares is a common technique in statistics used to find the best-fitting line or curve to a set of data points by minimizing the sum of the squares of the differences between the observed and predicted values. It is applied in both linear and nonlinear regression.

Linear Regression:

- **Model:** The relationship between the dependent variable y and independent variables $x_1, x_2, \ldots, x_n$ is assumed to be linear. This means the model takes the form:

$$y = a_0 + a_1x_1 + a_2x_2 + \cdots + a_nx_n + \epsilon$$

  where $a_0, a_1, \ldots, a_n$ are the coefficients, and ϵ is the error term.

- **Least Squares Estimation:**
  The aim is to find the coefficients $a_0, a_1, \ldots, a_n$ that minimize the sum of

the squared differences between the observed y_i values and the predicted $\hat{y}_i$ values:

$$\text{Sum of Squares Error (SSE)} = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

This equation is quadratic in terms of $a_0, a_1, \ldots, a_n$, making the minimization problem relatively straightforward. The solution to this minimization problem can be obtained by solving a system of linear equations derived from setting the partial derivatives with respect to each coefficient to zero.

Nonlinear Regression:

- **Model:** In nonlinear regression, the relationship between the dependent variable y and independent variables $x_1, x_2, \ldots, x_n$ is not linear. The model could take many forms, such as exponential, logarithmic, or polynomial functions:

  $$y = f(x_1, x_2, \ldots, x_n, \beta) + \epsilon$$

  where f is a nonlinear function, and β represents the parameters to be estimated.

- **Least Squares Estimation:**
  Similar to linear regression, the goal is to minimize the sum of squared errors (SSE):

  $$\text{SSE} = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

  However, because the function f is nonlinear, the SSE equation is often not quadratic and cannot be solved by simple algebraic methods. Instead, iterative numerical methods like the Newton-Raphson method, gradient descent, or others are used to find the minimum SSE. These methods involve guessing initial parameter values and iteratively improving them to reduce the SSE.

Key Differences:

- **Model Type:**
  - **Linear Regression:** Assumes a linear relationship between variables.
  - **Nonlinear Regression:** Assumes a nonlinear relationship.
- **Simplicity of Solution:**
  - **Linear Regression:** The least squares problem results in a set of linear equations that can be solved directly.
  - **Nonlinear Regression:** The least squares problem is more complex and often requires iterative methods to approximate the solution.

- **Computation:**
  - **Linear Regression:** Typically faster and less computationally intensive because it involves solving linear equations.
  - **Nonlinear Regression:** Requires more computational power due to the iterative nature of solving the nonlinear equations.
- **Interpretability:**
  - **Linear Regression:** Coefficients have straightforward interpretations (e.g., how much y changes with a unit change in x).
  - **Nonlinear Regression:** The interpretation of parameters depends on the specific nonlinear function, and can be more complex.

Conclusion:
In summary, while both linear and nonlinear regression use the least squares criterion to estimate model parameters, the complexity, computation, and interpretation of the results differ significantly due to the linear or nonlinear nature of the relationship between the variables

article amsmath

Internal 2, 2024

Consider the polynomial regression model of degree k given by:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + \epsilon_i$$

where $i = 1, 2, \ldots, n$, and $\epsilon_i \sim N(0, \sigma^2)$ are independent and identically distributed (i.i.d.) random errors. Prove that the design matrix X for the polynomial regression model has full column rank if the x_i values are distinct.

Answer:

To prove that the design matrix X for a polynomial regression model of degree k has full column rank when the x_i values are distinct, we can proceed as follows:

1. Understanding the Design Matrix X:

For a polynomial regression model of degree k, the design matrix X is defined as:

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^k \\ 1 & x_2 & x_2^2 & \cdots & x_2^k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^k \end{pmatrix}$$

Here:

- Each row corresponds to one data point x_i.

- Each column corresponds to a term in the polynomial, starting with the constant term (which is 1) and going up to x_i^k.

2. Definition of Full Column Rank:

The design matrix X has full column rank if all its columns are linearly independent. For a matrix with $k+1$ columns, having full column rank means that the rank of the matrix is $k+1$, which is the maximum possible rank for this matrix.

3. Linear Independence of Columns:

To check whether the columns of X are linearly independent, we need to see if the only solution to the equation

$$c_0 + c_1 x_i + c_2 x_i^2 + \cdots + c_k x_i^k = 0 \quad \text{for all } i = 1, 2, \ldots, n$$

is $c_0 = c_1 = \cdots = c_k = 0$.

Assume there is a non-trivial linear combination of the columns that equals the zero vector:

$$c_0 \cdot 1 + c_1 \cdot x_i + c_2 \cdot x_i^2 + \cdots + c_k \cdot x_i^k = 0 \quad \text{for all } i = 1, 2, \ldots, n$$

This equation implies that the polynomial $P(x) = c_0 + c_1 x + c_2 x^2 + \cdots + c_k x^k$ is equal to zero at each x_i.

4. Distinct x_i Values:

Since the x_i values are distinct, the polynomial $P(x)$ has n distinct roots. However, a non-zero polynomial of degree k can have at most k roots.

5. Conclusion:

The only way for $P(x)$ to be zero at more than k distinct points is if all the coefficients $c_0, c_1, \ldots, c_k$ are zero. Hence, $P(x)$ must be the zero polynomial, which means $c_0 = c_1 = \cdots = c_k = 0$.

Since the only solution to this equation is the trivial one (all c values are zero), the columns of X are linearly independent, meaning that X has full column rank.

Therefore, **the design matrix X has full column rank if the x_i values are distinct.**