BIOS 835: Linear Regression and Bayesian Decision Theory

Alexander McLain

August 29, 2023

Outline

Introduction

Least squares solution

Inference in linear regression

Multiple linear regression

Multivariate Regression

Bayesian Decision Theory


my simple linear regression model
outliers
correlated features
non-linearity

Linear Regression

- Linear regression models were developed before computers were around, but are still popular today.
- They are simple and (commonly) adequate and interpretable representations of the real world.
- They commonly work as well as difficult non-linear models for prediction, especially when there is limited data and low signal-to-noise ratio.
- Linear models do have assumptions, but many properties are robust to these assumptions (plus we can transform outcomes).

Linear Regression model

- A linear regression model is defined as

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \\ &= \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon \\ &= f(\boldsymbol{X}) + \epsilon \end{aligned}$$

  - Y is our outcome of interest
  - $(X_1, X_2 \ldots, X_p)$ are our covariates of interest
  - β_0: intercept;
  - $\beta_j, j = 1, \cdots, p$: regression coefficients.
  - $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$.

Linear Regression model

- We can also, define it in matrix form with the following

$$\boldsymbol{X} = \begin{pmatrix} 1 & X_{11} & \dots & X_{1p} \\ 1 & X_{21} & \dots & X_{2p} \\ \vdots & \vdots & & \\ 1 & X_{n1} & \dots & X_{np} \end{pmatrix} \quad \boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

  as the covariate data, observations, error term, and parameters in a linear regression model.

- The model can be expressed succinctly as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Least squares and RSS

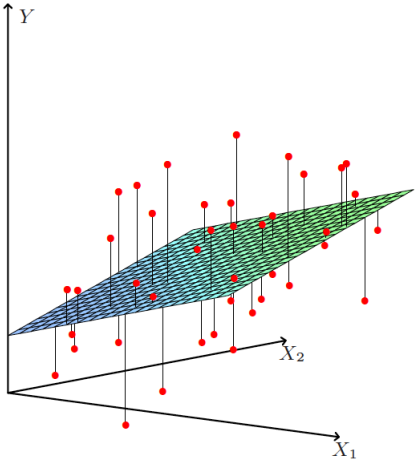
- The residual sum of squares for $\boldsymbol{\beta}$ is defined as

$$\begin{aligned} RSS(\boldsymbol{\beta}) &= \sum_{i=1}^{n}(y_i - \boldsymbol{X}_i'\boldsymbol{\beta})^2 \\ &= \sum_{i=1}^{n}\{y_i - f(\boldsymbol{X}_i)\}^2 \\ &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 \end{aligned}$$

  where $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})'$

- Using least squares criteria the estimate of $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}$, is the value that minimizes the RSS.

Least squares and RSS



The Least Squares Solution

- To minimize $RSS(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ we differentiate with respect to $\boldsymbol{\beta}$ to get

$$\begin{aligned} \frac{\partial RSS(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= -2\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ \frac{\partial^2 RSS(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} &= 2\boldsymbol{X}'\boldsymbol{X}. \end{aligned}$$

- If $\boldsymbol{X}$ has full column rank, setting the first derivative to zero yields,

$$\begin{aligned} 0 &= \boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \quad \text{and} \\ \hat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}X'\boldsymbol{y}. \end{aligned}$$

Predicted values

- Predicted values at a point $\boldsymbol{x}_0$ are given by $\hat{f}(\boldsymbol{x}_0) = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}$.
- The predicted values for the training data are given by

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

- $H = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is referred to as the "Hat Matrix"
- Geometrically, $\hat{\boldsymbol{\beta}}$ is chosen such that $\boldsymbol{y} - \hat{\boldsymbol{y}}$ is orthogonal to the subspace of $\boldsymbol{X}$.

Rank Deficiencies

- Recall, we said *"if* $\boldsymbol{X}$ *has full column rank"*
- When might this not be the case?

"Assumptions" standard regression models

Linear, Independent, Normality, and Equal variance (LINE)

- L: a linear relationship exists between Y and the X's.
- I: the data are independent.
- N: the residuals are normally distributed.
- E: the variance of the residuals is equal for all X's.

Properties of $\hat{\boldsymbol{y}}$, β and σ^2

- The variance of $\hat{\boldsymbol{\beta}}$ is

$$Var(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2$$

  which is estimated using

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

- Neither of these results requires any additional assumptions.

Properties of $\hat{\boldsymbol{y}}$, β and σ^2

Once we impose the LINE assumptions we get the following properties

- $\hat{\boldsymbol{\beta}} \sim MVN(\boldsymbol{\beta}, (X'X)^{-1}\sigma^2)$, hypothesis tests and confidence intervals for $\hat{\beta}_j$ can be based on

$$z_j = \frac{\hat{\beta}_j}{\hat{\sigma} v_j} \sim t_{n-p-1}$$

  where v_j is the jth diagonal element of $(X'X)^{-1}$
- The predicted value at x_0 is $y \sim N\{f(\boldsymbol{x}_0), \sigma^2(1 + x_0(X'X)^{-1}x_0)\}$

Partial F-test for multiple variables

- Suppose we'd like to test $H_0 : \beta_j = \beta_{j+1} = \ldots = \beta_{j+k-1} = 0$ v.s $H_1 : \beta_l \neq 0$ for some $l = j, \ldots, j+k-1$ given the other $(p-k)$ variables.
- This can be tested via

$$F_0 = \frac{(RSS_0 - RSS_1)/(p1 - p0)}{RSS_1/(N - p_1 - 1)}$$

  where RSS_1 and $p_1 + 1$ are the RSS and number of parameters for the bigger model (similarly for RSS_0 and p_0).
- Under the Gaussian assumption and the null hypotheses $F_0 \sim F_{p_1-p_0, N-p_1-1}$
- Reject if

$$F_0 > F_{p_1-p_0, N-p_1-1}(1-\alpha)$$

Gauss-Markov Theorem

- The Gauss-Markov Theorem is one of the most famous results in statistics.
- It states that the least squares estimate $\hat{\boldsymbol{\beta}}$ has the smallest variance among all linear unbiased estimates.
- Specifically, suppose we wish to estimate $\theta = a'\boldsymbol{\beta}$. The least squares estimate is

$$\hat{\theta} = a'\hat{\boldsymbol{\beta}} = a'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

  which is unbiased when the linear model is correct.
- The Gauss-Markov Theorem shows that another estimator $\tilde{\theta} = \boldsymbol{c}'\boldsymbol{y}$, which is unbiased for $a'\boldsymbol{\beta}$, will have

$$Var(a'\hat{\boldsymbol{\beta}}) \leq Var(\boldsymbol{c}'\boldsymbol{y})$$

Gauss-Markov Theorem

- What does the Gauss-Markov Theorem not say?
- Consider

$$MSE(\hat{\theta}) = E\{(\hat{\theta} - \theta)^2\}$$

Simple (simple) LR and notation

- Consider the following (very) simple LR model with 1 covariate and no intercept

$$Y = X\beta + \epsilon$$

- The LS estimate and residual are

$$\hat{\beta} = \frac{\sum_1^n x_i y_i}{\sum_1^n x_i^2}$$
$$r_i = y_i - x_i\hat{\beta}$$

- It is convenient (and more general) to use the following notation

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{i=1}^{n} x_i y_i = \boldsymbol{x}'\boldsymbol{y}, \quad \text{so} \quad \hat{\beta} = \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\langle \boldsymbol{x}, \boldsymbol{x} \rangle} = \frac{\boldsymbol{x}'\boldsymbol{y}}{\boldsymbol{x}'\boldsymbol{x}}$$

MLR for orthgonal vectors

- Suppose we have $1 = x_0, x_1, \ldots, x_p$ inputs are all orthogonal
  - $\langle \boldsymbol{x}_j, \boldsymbol{x}_k \rangle = 0$ for all $j \neq k$
- Then

$$\hat{\beta}_j = \frac{\langle \boldsymbol{x}_j, \boldsymbol{y} \rangle}{\langle \boldsymbol{x}_j, \boldsymbol{x}_j \rangle}$$

- What are the impacts of this?
- Will this ever happen?

Simple Gram-Schmidt process

- Now, suppose we have the SLR model

$$Y = \beta_0 + X\beta_1 + \epsilon$$

How can we make 1 *and* X *orthogonal?*

- By orthogonalizing we get the coefficient as

$$\hat{\beta}_1 = \frac{\langle \boldsymbol{x} - \bar{x}\mathbf{1}, \boldsymbol{y} \rangle}{\langle \boldsymbol{x} - \bar{x}\mathbf{1}, \boldsymbol{x} - \bar{x}\mathbf{1} \rangle}$$

which is the coefficient for the model $Y = (x - \bar{x})\beta_1 + \epsilon$

Simple Gram-Schmidt process

- So a way to calculate β_1 is to do the following:
  1. regress $\boldsymbol{x}$ on $\mathbf{1}$, and get the residual $\boldsymbol{z} = \boldsymbol{x} - \bar{x}\mathbf{1}$
  2. regress $\boldsymbol{y}$ on the residual $\boldsymbol{z}$ to give the coefficient $\hat{\beta}_1$

  where regress $\boldsymbol{y}$ on $\boldsymbol{z}$ means to estimate the "no intercept model" we initially started with.

Gram-Schmidt algorithm

We can generalize this for p non-orthogonal inputs $1 = x_0, x_1, \ldots, x_p$

1. Set $\boldsymbol{z}_0 = \boldsymbol{x}_0 = \mathbf{1}$
2. For $j = 1, 2, \ldots, p$:

   regress $\boldsymbol{x}_j$ on $\boldsymbol{z}_0, \ldots, \boldsymbol{z}_{j-1}$, and get the residual $\boldsymbol{z}_j = \boldsymbol{x}_j - \sum_{k=0}^{j-1} \hat{\gamma}_{kj} \boldsymbol{z}_k$ where $\hat{\gamma}_{kj} = \langle \boldsymbol{z}_k, \boldsymbol{x}_j \rangle / \langle \boldsymbol{z}_k, \boldsymbol{z}_k \rangle$.
3. regress $\boldsymbol{y}$ on the residual $\boldsymbol{z}_p$ to give the coefficient

$$\hat{\beta}_p = \frac{\langle \boldsymbol{z}_p, \boldsymbol{y} \rangle}{\langle \boldsymbol{z}_p, \boldsymbol{z}_p \rangle}$$

Implications

- Another way to express the variability in $\hat{\beta}_p$ is

$$Var(\hat{\beta}_p) = \frac{\sigma^2}{\langle \boldsymbol{z}_p, \boldsymbol{z}_p \rangle} = \frac{\sigma^2}{||\boldsymbol{z}_p||^2}$$

- What will $\boldsymbol{z}_p$ look like if $\boldsymbol{x}_p$ is accurately predicted from $x_0, x_1, \ldots, x_{(p-1)}$?
- What about $||\boldsymbol{z}_p||^2$?
- How do we get back to the original $\boldsymbol{x}$ coefficients?
  - Let $\boldsymbol{x}_{0,1,\ldots,(j-1),(j+1),\ldots,p}$ denote the residual after regression $\boldsymbol{x}_j$ on $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{(j-1)}, \boldsymbol{x}_{(j+1)}, \ldots, \boldsymbol{x}_p$
  - The jth multiple regression coefficient is the univariate regression coefficient of $\boldsymbol{y}$ on $\boldsymbol{x}_{0,1,\ldots,(j-1),(j+1),\ldots,p}$.

Matrix GS with QR decomp

- In matrix form, we can represent the second step of the GS algorithm with

$$\boldsymbol{X} = \boldsymbol{Z}\Gamma$$

where $\boldsymbol{Z}$ has columns $\boldsymbol{z}_j$, and Γ is the upper triangular matrix with entries $\hat{\gamma}_{kj}$.

- Let's add in $\boldsymbol{D}$ a diagonal matrix with entries $D_{jj} = ||\boldsymbol{z}_j||$ to get

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{Z}\boldsymbol{D}^{-1}\boldsymbol{D}\Gamma \\ &= \boldsymbol{Q}\boldsymbol{R} \end{aligned}$$

The $\boldsymbol{QR}$ decomposition of $\boldsymbol{X}$

Matrix GS with QR decomp

- Using the $\boldsymbol{QR}$ decomposition of $\boldsymbol{X}$ we get the following form for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{y}}$

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}X'\boldsymbol{y} = \{(\boldsymbol{QR})'\boldsymbol{QR}\}^{-1}(\boldsymbol{QR})'\boldsymbol{y} = \boldsymbol{R}^{-1}\boldsymbol{Q}'\boldsymbol{y}, \\ \hat{\boldsymbol{y}} &= \boldsymbol{QQ}'\boldsymbol{y} \end{aligned}$$

  This form of $\hat{\boldsymbol{\beta}}$ is a lot easier to solve because $\boldsymbol{R}$ is upper triangular.

Introduction to Multivariate Regression

- Now let's consider the more classical multivariate case where we have $Y_1, Y_2, \ldots, Y_K$ and $X_0, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$ and the linear model

$$Y_k = \beta_{0k} + \sum_{j=1}^{p} X_j \beta_{jk} + \epsilon_k = f_k(\boldsymbol{X}) + \epsilon_k \quad \text{for } k = 1, 2, \ldots, K$$

or in matrix form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$$

where $\boldsymbol{Y}$ is $n \times K$ are the outcomes, $\boldsymbol{X}$ is $n \times (p+1)$ are the inputs, $\boldsymbol{B}$ is an $(p+1) \times K$ matrix of β coefficients and $\boldsymbol{E}$ is an $n \times K$ matrix of residuals.

RSS and solution

- The RSS for the multivariate case is given by

$$\begin{aligned} RSS(\boldsymbol{B}) &= \sum_{k=1}^{K}\sum_{i=1}^{n}\{y_{ik} - f_k(x_i)\}^2 \\ &= tr\{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})\} \end{aligned}$$

- It's rather straightforward to show that $RSS(\boldsymbol{B})$ is minimized with

$$\hat{\boldsymbol{B}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

Implications

- More specifically, for the kth outcome the coefficient is

$$\hat{\boldsymbol{\beta}}_k = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}_k$$

- If we were to use the general linear model where $Cov(\boldsymbol{\epsilon}) = \Sigma$ and

$$RSS(\boldsymbol{B}, \Sigma) = tr\{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})'\Sigma^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})\},$$

  ..... the solution is still the same.
- What does this imply about doing regression with multiple outcomes?

Introduction

- **Bayesian Decision Theory** is a decision-making framework that combines probability theory and decision theory.
- It is based on the Bayesian approach to probability, where prior beliefs are updated with new evidence in the form of data to produce posterior beliefs.
- Bayesian Decision Theory provides a formal mechanism for making decisions under uncertainty, weighing the risks of different actions against the posterior beliefs about the state of the world.

Probabilistic Setup

- Let θ be a parameter or state of nature we want to learn about (it could represent various things: the true state of a system, a model parameter, etc.)
- $p(\theta)$ is the prior probability distribution over θ which reflects our beliefs about θ before observing data.
- $p(x|\theta)$ is the likelihood function which provides the probability of observing data x given θ.
- $p(\theta|x)$ is the posterior probability distribution over θ given data x, which is computed using Bayes' theorem:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$$

where $p(x)$ is the evidence or marginal likelihood.

Decision Making

- Suppose we have a set of possible actions A we can take.
- A loss function $L(\theta, a)$ represents the cost or loss of taking action a when the true state is θ.
- The loss function quantifies the "penalty" for making decisions; related to the norms we discussed but more general (e.g., for binary data).
- The goal in Bayesian decision theory is to minimize the expected loss. The decision rule is given by:

$$a^* = \arg\min_a \mathbb{E}[L(\theta, a)|x]$$

where the expectation is over the posterior distribution of θ given data x.

Example: Binary Hypothesis Testing

Let's consider a simple case where θ can be either H_0 or H_1.

- Prior Probabilities: $p(H_0)$ and $p(H_1)$.
- Likelihoods: $p(x|H_0)$ and $p(x|H_1)$.
- Actions: here we'll consider $A \in \{a_0, a_1\}$, i.e., a "hard" decision.
- Loss function might be defined such that $L(H_0, a_0) = 0$ (no loss for correctly deciding H_0), $L(H_0, a_1) = 1$ (loss for wrongly deciding H_1 when truth is H_0), and similar definitions for decisions regarding H_1.
  - if $A \in (0, 1)$ the other loss functions would be used.

Example: Binary Hypothesis Testing

- The Bayesian decision rule will decide a_1, i.e., H_1, if:

$$\frac{p(x|H_1)p(H_1)}{p(x)} > \frac{p(x|H_0)p(H_0)}{p(x)}$$

This simplifies to:

$$\frac{p(x|H_1)}{p(x|H_0)} > \frac{p(H_0)}{p(H_1)}$$

Here, $\frac{p(x|H_1)}{p(x|H_0)}$ is the likelihood ratio and $\frac{p(H_0)}{p(H_1)}$ is the prior odds ratio.