Linear Regression

Data 100: Principles and Techniques of Data Science

Sandrine Dudoit

Department of Statistics and Division of Biostatistics, UC Berkeley

Spring 2019

Outline

Version: 12/03/2019, 17:57

mpg Dataset

- We examined the mpg dataset with the goal of relating a car's fuel consumption, i.e., mileage per gallon (mpg), to features such as the number of cylinders and horsepower.
- Let Y denote the random variable for mpg and $X = (X_1, \ldots, X_7)$ the random variables for the other 7 features (numbered in the order "cylinders", "displacement", "horsepower", "weight", "acceleration", "model year", "origin").
- The data for the ith car are (X_i, Y_i), $i = 1, \ldots, n$, $n = 392$.
- A natural function for relating mpg to the 7 features is the regression function, i.e., the conditional expected value of mpg given the 7 variables: $\mathrm{E}[Y|X]$.

mpg Dataset

- There are a variety of models for the regression function, ranging from trivial (e.g., constant model) to complex or non-parametric models.
- Constant regression model.

$$E[Y|X] = \beta_0.$$

  This model completely ignores the obvious association of mpg with features such as horsepower.
- Univariate linear regression model.

$$E[Y|X] = \beta_0 + \beta_4 X_4.$$

  This model is more informative than the constant model, but doesn't account for the association of mpg with the other 6 covariates or potential non-linear effects of horsepower on mpg (cf. higher-order polynomial).

mpg Dataset

- Multiple linear regression model.

$$
\begin{aligned}
E[Y|X] \quad = \quad & \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \\
& + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 \\
& + \beta_{7,Japan} \, \mathrm{I}(X_7 = Japan) + \beta_{7,USA} \, \mathrm{I}(X_7 = USA),
\end{aligned}
$$

  where I() denotes the indicator function, equal to one if its argument is true and zero otherwise. This model accounts for all 7 covariates, but could miss possible non-linear dependencies of mpg on the 7 features as well as interactions between these features.

- Note that we treat the qualitative variable “origin” differently than the other 6 variables that are quantitative.

mpg Dataset

- In general, we rely on dummy variables, a.k.a., indicator variables or one-hot encoding, to indicate whether or not a qualitative covariate takes on a particular value.
- A qualitative variable taking on K values is typically represented by $K-1$ dummy variables, each corresponding to a particular value ($K-1$ out of K values). As discussed later on, using K dummy variables along with an intercept would lead to an ill-defined least squares estimator (design matrix with non-full column rank).
- For the mpg dataset, the "origin" variable takes on three values, "europe", "japan", and "usa", and is thus represented by two dummy variables.

mpg Dataset

- Regression tree model.

$$E[Y|X] = \beta_0 + \sum_{k=1}^{K} \beta_k \, I(X \in A_k),$$

where the sets A_k form a partition of the covariate space. Regression trees are well-suited for covariates of different types and measured on different scales, as well as for interactions, but could lead to unstable estimators.

mpg Dataset

- Non-parametric model.

$$\mathrm{E}[Y|X] = \theta(X),$$

where $\theta : \mathbb{R}^J \rightarrow \mathbb{R}$ denotes an arbitrary function of the covariates X. This general model could be fit by, e.g., robust local regression (e.g., loess), which does not provide a simple interpretable regression function θ. The function could be very data-adaptive at the risk of overfitting, i.e., fit the sample data very closely but not additional data from the same population.

Regression Models and Risk Optimization

- In many inference settings (e.g., mpg dataset, Craigslist rental dataset), the parameter of interest is a regression function, i.e., the conditional expected value of an outcome $Y \in \mathbb{R}$ given covariates $X = (X_j : j = 1, \ldots, J) \in \mathbb{R}^J$

$$\theta(X) = \mathrm{E}[Y|X].$$

- In general, there is an infinite number of regression functions $\theta : \mathbb{R}^J \to \mathbb{R}$, ranging from trivial (e.g., constant) to complex or non-parametric (e.g., robust local regression, ensemble methods).
- Additionally, regression models are typically fit on data from a sample drawn from a population, i.e., a learning set $\mathcal{L}_n = \{(X_i, Y_i) : i = 1, \ldots, n\}$.

Regression Models and Risk Optimization

- We are therefore faced with the following two key questions.
  - What is an appropriate model for the regression function?
  - How can we use the sample to accurately infer the regression function for an entire population?
    This will depend on how the sample was obtained, i.e., whether it was obtained according to a well-defined probabilistic sampling procedure.
- Model selection, parameter definition and inference, and estimator performance assessment can be handled within the general framework of risk optimization.
- A natural loss function in the context of regression is the squared error/L_2 loss function

$$L_2((X, Y), \theta) \equiv (Y - \theta(X))^2. \tag{1}$$

Regression Models and Risk Optimization

- We demonstrated in a previous lecture that expected values minimize risk for the L_2 loss function, i.e., minimize the mean squared error (MSE).
- The population regression function (an unknown parameter) minimizes MSE computed with respect to the unknown data generating distribution P

$$\theta(X) = \mathrm{E}_P[Y|X] = \mathrm{argmin}_{\theta' \in \Theta} \, \mathrm{E}_P[(Y - \theta'(X))^2], \quad (2)$$

where Θ denote the parameter space.

Regression Models and Risk Optimization

- A natural first-pass at estimating θ is the resubstitution estimator, i.e., the value of θ which minimizes the resubstitution risk estimator or empirical MSE

$$\begin{aligned}\hat{\theta}_n(X) &= \text{argmin}_{\theta \in \Theta} \mathrm{E}_{P_n}[(Y - \theta(X))^2] \\ &= \text{argmin}_{\theta \in \Theta} \sum_{i=1}^{n} (Y_i - \theta(X_i))^2.\end{aligned}$$

- Note that in the above two equations, we did not put any restrictions on the parameter space Θ, i.e., on the regression function θ.
- In principle, one could consider arbitrarily complex functions of X or non-parametric models, that place few, if any, restrictions on the regression function (e.g., continuity).

Regression Models and Risk Optimization

- However, minimizing empirical risk over the resulting large parameter spaces is computationally costly and can lead to overfitting or ill-defined estimators.
- It is customary to consider instead smaller parametric models, such as the well-known linear regression model.

Linear Regression Model

- Consider a data structure $(X, Y) \sim P$, where $Y \in \mathbb{R}$ is a scalar outcome (a.k.a., dependent variable, response) and $X = (X_j : j = 1, \ldots, J) \in \mathbb{R}^J$ is a J-dimensional column vector of covariates (a.k.a., features, explanatory variables, independent variables).
- A common and simple model for regression functions is the linear regression model

$$\mathrm{E}[Y|X] = X^\top \beta = \sum_{j=1}^{J} \beta_j X_j = \beta_1 X_1 + \ldots + \beta_J X_J, \quad (3)$$

where the column vector $\beta = (\beta_j : j = 1, \ldots, J) \in \mathbb{R}^J$ contains the parameters of the model, referred to as regression coefficients.

Linear Regression Model

- The expression "linear regression" refers to linearity in the regression coefficients/parameters β. Covariates X can enter the model via arbitrary functions, e.g., polynomial, logarithm, sine functions.
- In order to accommodate an intercept, one sets $X_1 \equiv 1$.
- Additionally, the X's could correspond to dummy or indicator variables for qualitative covariates.
- Linear regression models are widely used (not always appropriately!) due to their simplicity, mathematical tractability, and optimality properties (provided the assumptions of the model hold!).
- In particular, one can derive a closed-form expression for the LSE of β.

Linear Regression Model

Table 1: *Examples of linear and non-linear regression models.* In order to fit the linear regression models in rows 1–5, we typically redefine the covariates X so that $\mathrm{E}[Y|X] = X^\top \beta$. In particular, in order to accommodate an intercept, we set $X_1 = 1$. For instance, for the model in row 2: $X_1 = 1$, $X_2 \leftarrow X_2$, and $X_3 \leftarrow X_2^2$. For the model in row 5: $X_1 = 1$ and $X_2 \leftarrow \sin(X_2)$.

$\theta(X)$	Linear regression model
$\beta_1 + \beta_2 X_2$	Yes
$\beta_1 + \beta_2 X_2 + \beta_3 X_2^2$	Yes
$\beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_2 X_3$	Yes
$\beta_1 + \beta_2 \mathrm{I}(X_2 = 1)$	Yes
$\beta_1 + \beta_2 \sin(X_2)$	Yes
$\beta_1 X_2^{\beta_2}$	No
$\log(\beta_1 + \beta_2 X_2)$	No

Least Squares Estimation

- Suppose one has a random sample $\mathcal{L}_n = \{(X_i, Y_i) : i = 1, \ldots, n\}$ of n covariate/outcome pairs from the population of interest.
- The set $\mathcal{L}_n$ is often referred to as learning set, as it is used to infer/learn the population parameters, here, the regression coefficients β.
- Define the design matrix $\mathbf{X}_n$ as the $n \times J$ matrix with ith row corresponding to the ith covariate vector X_i, $i = 1, \ldots, n$.
- Define the outcome vector $\mathbf{Y}_n$ as an n-dimensional column vector with ith element corresponding to the ith outcome Y_i, $i = 1, \ldots, n$.

Least Squares Estimation

- Then, under the linear regression model

$$
\begin{aligned}
\mathrm{E}[\mathbf{Y}_n|\mathbf{X}_n] &= \mathbf{X}_n\beta \\
&= \begin{bmatrix} X_{1,1} & X_{1,2} & \dots & X_{1,J} \\ X_{2,1} & X_{2,2} & \dots & X_{2,J} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n,1} & X_{n,2} & \dots & X_{n,J} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_J \end{bmatrix} \\
&= \begin{bmatrix} \sum_{j=1}^J \beta_j X_{1,j} \\ \sum_{j=1}^J \beta_j X_{2,j} \\ \vdots \\ \sum_{j=1}^J \beta_j X_{n,j} \end{bmatrix}.
\end{aligned}
$$

Least Squares Estimation

- The least squares estimator (LSE) of the regression coefficients β is the resubstitution estimator, i.e., the value of β that minimizes empirical MSE

$$
\begin{aligned}
\hat{\beta}_n &\equiv \text{argmin}_{\beta \in \mathbb{R}^J} R_2(P_n, \beta) \qquad (4)\\
&= \text{argmin}_{\beta \in \mathbb{R}^J} \mathrm{E}_{P_n}[(Y - X^\top \beta)^2] \\
&= \text{argmin}_{\beta \in \mathbb{R}^J} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{J} \beta_j X_{i,j} \right)^2 \\
&= \text{argmin}_{\beta \in \mathbb{R}^J} (\mathbf{Y}_n - \mathbf{X}_n \beta)^\top (\mathbf{Y}_n - \mathbf{X}_n \beta) .
\end{aligned}
$$

- One can derive a simple closed-form expression for the LSE of the regression coefficients β using calculus.

Least Squares Estimation

- We compute the gradient of the empirical risk with respect to β and define β as the root of this gradient [1]

$$
\begin{aligned}
\nabla_\beta R_2(P_n, \beta) &= \nabla_\beta (\mathbf{Y}_n - \mathbf{X}_n\beta)^\top (\mathbf{Y}_n - \mathbf{X}_n\beta) \\
&= \nabla_\beta \Big(\mathbf{Y}_n^\top \mathbf{Y}_n - \mathbf{Y}_n^\top \mathbf{X}_n\beta \\
&\qquad - \beta^\top \mathbf{X}_n^\top \mathbf{Y}_n + \beta^\top \mathbf{X}_n^\top \mathbf{X}_n\beta \Big) \\
&= \nabla_\beta \left(-2\beta^\top \mathbf{X}_n^\top \mathbf{Y}_n + \beta^\top \mathbf{X}_n^\top \mathbf{X}_n\beta \right) \\
&= -2\mathbf{X}_n^\top \mathbf{Y}_n + 2\mathbf{X}_n^\top \mathbf{X}_n\beta.
\end{aligned}
$$

Line 3 follows by noting that (1) $\mathbf{Y}_n^\top \mathbf{Y}_n$ is a constant in β and thus has gradient zero and (2) $\beta^\top \mathbf{X}_n^\top \mathbf{Y}_n$ is a scalar and thus equal to its transpose $\mathbf{Y}_n^\top \mathbf{X}_n\beta$. Line 4 follows from properties of the gradient (see, for example, https://www.textbook.ds100.org/ch/13/linear_multiple.html).

- A good exercise, to make sure you are comfortable with the matrix representation of the linear regression model and with gradients, is to derive the above result from first principles, i.e., element-wise differentiation of matrices.
- Setting the gradient equal to zero, yields the normal equations

$$\mathbf{X}_n^\top \mathbf{Y}_n = \mathbf{X}_n^\top \mathbf{X}_n \beta. \tag{5}$$

- When the design matrix is of full column rank, i.e., $\mathbf{X}_n^\top \mathbf{X}_n$ is invertible, the normal equations have a unique solution

$$\hat{\beta}_n = (\mathbf{X}_n^\top \mathbf{X}_n)^{-1} \mathbf{X}_n^\top \mathbf{Y}_n. \tag{6}$$

Least Squares Estimation

- In the special case when $J = 2$, $X_1 = 1$, and $X_2 = X \in \mathbb{R}$,

$$\mathbf{X}_n^\top \mathbf{X}_n = \begin{bmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{bmatrix} = \begin{bmatrix} n & n\bar{X} \\ n\bar{X} & \sum_{i=1}^n X_i^2 \end{bmatrix},$$

so that

$$\begin{aligned} (\mathbf{X}_n^\top \mathbf{X}_n)^{-1} &= \frac{1}{n\sum_{i=1}^n X_i^2 - n^2\bar{X}^2} \begin{bmatrix} \sum_{i=1}^n X_i^2 & -n\bar{X} \\ -n\bar{X} & n \end{bmatrix} \\ &= \frac{1}{n^2 s_X^2} \begin{bmatrix} \sum_{i=1}^n X_i^2 & -n\bar{X} \\ -n\bar{X} & n \end{bmatrix}, \end{aligned}$$

where s_X denotes the empirical standard deviation of X

$$s_X^2 \equiv \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n}\left(\sum_{i=1}^n X_i^2 - n\bar{X}^2\right). \qquad (7)$$

Least Squares Estimation

Also,

$$\mathbf{X}_n^\top \mathbf{Y}_n = \begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_i Y_i \end{bmatrix} = \begin{bmatrix} n\bar{Y} \\ \sum_{i=1}^n X_i Y_i \end{bmatrix},$$

thus,

$$\begin{aligned} \hat{\beta}_n &= \frac{1}{n^2 s_X^2} \begin{bmatrix} \sum_{i=1}^n X_i^2 & -n\bar{X} \\ -n\bar{X} & n \end{bmatrix} \begin{bmatrix} n\bar{Y} \\ \sum_{i=1}^n X_i Y_i \end{bmatrix} \\ &= \frac{1}{n^2 s_X^2} \begin{bmatrix} n\bar{Y} \sum_{i=1}^n X_i^2 - n\bar{X} \sum_{i=1}^n X_i Y_i \\ n \sum_{i=1}^n X_i Y_i - n^2 \bar{X}\bar{Y} \end{bmatrix}. \end{aligned}$$

Least Squares Estimation

- Hence,

$$\begin{aligned} \hat{\beta}_{1,n} &= \bar{Y} - \hat{\beta}_{2,n}\bar{X} \\ \hat{\beta}_{2,n} &= \frac{s_Y}{s_X} r_{X,Y}, \end{aligned} \tag{8}$$

where $r_{X,Y}$ denotes the empirical Pearson correlation coefficient between X and Y

$$\begin{aligned} r_{X,Y} &\equiv \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} \\ &= \frac{\sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{(\sum_{i=1}^n X_i^2 - n\bar{X}^2)(\sum_{i=1}^n Y_i^2 - n\bar{Y}^2)}}. \end{aligned} \tag{9}$$

Least Squares Estimation

- The LSE of the regression function is the line that minimizes the sum of the squares of the differences between the observed and fitted responses, i.e., the sum of the squared residuals.

[1]The gradient is the vector of first derivatives with respect to each of the J coefficients β_j.

Sampling Distribution of LSE

- Now that we have an estimator of the regression coefficients β, it is natural to examine its performance, i.e., how accurate it is as an estimator of the population parameter.
- Assuming the linear regression model is true, i.e., $\mathrm{E}[Y|X] = X^\top \beta$, then the LSE of the regression coefficients β is unbiased

$$\mathrm{E}_P[\hat{\beta}_n|\mathbf{X}_n] = \beta.$$

Proof.

$$
\begin{aligned}
E_P[\hat{\beta}_n|\mathbf{X}_n] &= E_P[(\mathbf{X}_n^\top \mathbf{X}_n)^{-1}\mathbf{X}_n^\top \mathbf{Y}_n|\mathbf{X}_n] \\
&= (\mathbf{X}_n^\top \mathbf{X}_n)^{-1}\mathbf{X}_n^\top E_P[\mathbf{Y}_n|\mathbf{X}_n] \\
&= (\mathbf{X}_n^\top \mathbf{X}_n)^{-1}\mathbf{X}_n^\top \mathbf{X}_n \beta \\
&= \beta.
\end{aligned}
$$

Line 2 follows by linearity of expected values, as $\mathbf{X}_n$ is treated as a constant in the conditional expectation. □

- If one further assumes that $\mathrm{Var}[Y|X] = \sigma^2$ (i.e., the outcome has constant variance given the covariates) and that the (X_i, Y_i) are independent, one can show that the conditional covariance matrix of the LSE is

$$
\mathrm{Cov}[\hat{\beta}_n|\mathbf{X}_n] = \sigma^2(\mathbf{X}_n^\top \mathbf{X}_n)^{-1}. \tag{10}
$$

Sampling Distribution of LSE

- Adding yet another assumption, namely that the outcomes have a Gaussian distribution given the covariates, i.e.,

$$\mathbf{Y}_n|\mathbf{X}_n \sim \mathrm{N}\left(\mathbf{X}_n\beta, \sigma^2\mathbf{I}_n\right),$$

where $\mathbf{I}_n$ denotes the $n \times n$ identity matrix, then one can show that the LSE has a Gaussian distribution given the covariates

$$\hat{\beta}_n|\mathbf{X}_n \sim \mathrm{N}\left(\beta, \sigma^2(\mathbf{X}_n^\top\mathbf{X}_n)^{-1}\right). \tag{11}$$

- Assuming that the (X_i, Y_i) are independent, the Central Limit Theorem implies that the LSE is asymptotically (i.e., for large n) normal, irrespective of the distribution of the outcomes.

Sampling Distribution of LSE

- It is important to note that all inference is performed conditional on the design matrix $\mathbf{X}_n$.
- Furthermore, one can compute a $\hat{\beta}_n$ as in Equation (6) given any design matrix $\mathbf{X}_n$ of full column rank and outcome vector $\mathbf{Y}_n$.
- Whether this $\hat{\beta}_n$ is meaningful, unbiased, or has covariance matrix as in Equation (10) is another story, as these properties hold only under certain modeling assumptions about the data generating distribution of the (X_i, Y_i).

mpg Dataset

Univariate linear regression model.

- First, consider fitting a simple univariate linear regression model where mpg is regressed on only horsepower:

$$\mathrm{E}[Y|X] = \beta_0 + \beta_4 X_4.$$

- For this model, the design matrix is $n \times 2$,

$$\mathbf{X}_n = \begin{bmatrix} 1 & 130 \\ 1 & 165 \\ \vdots & \vdots \\ 1 & 82 \end{bmatrix}.$$

mpg Dataset

- One has

$$\mathbf{X}_n^\top \mathbf{X}_n = \begin{bmatrix} 392 & 40,952 \\ 40,952 & 4,857,524 \end{bmatrix}$$
$$\mathbf{X}_n^\top \mathbf{Y}_n = \begin{bmatrix} 9,190.8 \\ 868,718.8 \end{bmatrix},$$

so that the LSE of the two regression coefficients are

$$\hat{\beta}_n = (\mathbf{X}_n^\top \mathbf{X}_n)^{-1} \mathbf{X}_n^\top \mathbf{Y}_n$$
$$= \begin{bmatrix} 392 & 40,952 \\ 40,952 & 4,857,524 \end{bmatrix}^{-1} \begin{bmatrix} 9,190.8 \\ 868,718.8 \end{bmatrix}$$
$$\approxeq \begin{bmatrix} 39.9359 \\ -0.1578 \end{bmatrix}.$$

Linear Regression

Dudoit

Motivation

mpg Dataset

Regression Models and Risk Optimization

Linear Regression Model

Least Squares Estimation

Sampling Distribution of LSE

mpg Dataset

mpg Dataset

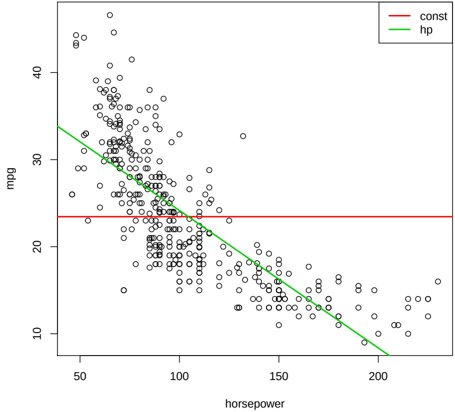

Figure 1: *mpg dataset.* Linear regression of mpg on horsepower.

mpg Dataset

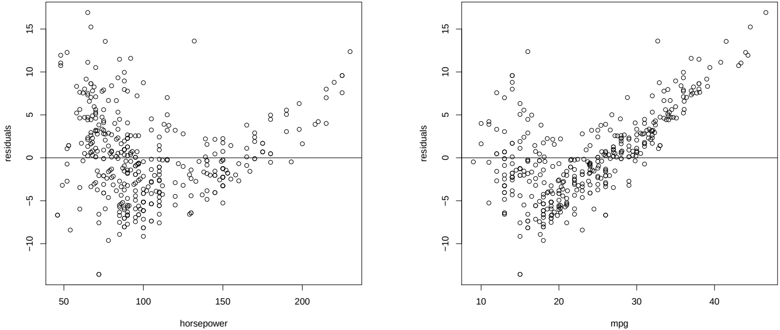

Figure 2: *mpg dataset.* Linear regression of mpg on horsepower, residuals

mpg Dataset

Multiple linear regression model.

- Now, consider fitting a multiple linear regression model where mpg is regressed on all 7 covariates:

$$
\begin{aligned}
E[Y|X] &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \\
&\quad + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 \\
&\quad + \beta_{7,Japan} \mathrm{I}(X_7 = Japan) + \beta_{7,USA} \mathrm{I}(X_7 = USA).
\end{aligned}
$$

- For this model, the design matrix is $n \times 9$, as we use two dummy variables for the qualitative covariate "origin"

$$
\mathbf{X}_n = \begin{bmatrix}
1 & 8 & 307 & 130 & 3,504 & 12.0 & 70 & 0 & 1 \\
1 & 8 & 350 & 165 & 3,693 & 11.5 & 70 & 0 & 1 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & 4 & 119 & 82 & 2,720 & 19.4 & 82 & 0 & 1
\end{bmatrix}.
$$