Linear Regression

Rebecca C. Steorts

Linear regression

- Assume you have seen this method before and are familiar
- running it on data
- interpreting results
- applying to applications
- estimation and prediction
- We will consider a generalized treatment of linear regression, including statistical rigor
- Why? This is essential for other statistical learning methods

Linear regression as a probabilistic model

Training data $(x_1, y_1), \ldots, (x_n, y_n)$, where $y_i \in \mathbb{R}$ and x_i can be in any space.

- Let $\phi_1, \ldots, \phi_p$ denote basis or feature functions.
- x_i is mapped to $\phi(x_i) = (\phi_1(x_i), \ldots, \phi_p(x_i))^T \in \mathbb{R}^p$

Example: x_3 maps to $\phi(x_3) = (\phi_1(x_3), \ldots, \phi_p(x_3))^T$

Example of basis functions

Response y_i is modeled as random variable (r.v)

$$Y_i = \phi(x_i)^T \beta + \epsilon_i$$

where $\beta \in \mathbb{R}^p$ and $\epsilon_1, \ldots, \epsilon_n \sim N(0, \sigma^2)$, independently.

The "linear" part in regression refers to linearity in the regression parameters β (and not the explantory variables, x_i).

Linear regression

Let $Y = (Y_1, \ldots, Y_n), \epsilon = (\epsilon_1, \ldots, \epsilon_n)$,

and

$$A = \begin{bmatrix} \phi(x_1)^T \\ \phi(x_2)^T \\ \vdots \\ \phi(x_n)^T \end{bmatrix}$$

$$A = \begin{bmatrix} (\phi_1(x_1), \ldots, \phi_p(x_1))^T \\ (\phi_1(x_2), \ldots, \phi_p(x_2))^T \\ \vdots \\ (\phi_1(x_n), \ldots, \phi_p(x_n))^T \end{bmatrix}_{n \times p}$$

Linear regression

$$A = \begin{bmatrix} (\phi_1(x_1), \ldots, \phi_p(x_1))^T \\ (\phi_1(x_2), \ldots, \phi_p(x_2))^T \\ \vdots \\ (\phi_1(x_n), \ldots, \phi_p(x_n))^T \end{bmatrix}_{n \times p}$$

results in

$$A = \begin{bmatrix} \phi_1(x_1) & \phi_1(x_2) & \phi_1(x_3) & \ldots & \phi_1(x_n) \\ \phi_2(x_1) & \phi_2(x_2) & \phi_2(x_3) & \ldots & \phi_2(x_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_p(x_1) & \phi_p(x_2) & \phi_p(x_3) & \ldots & \phi_p(x_n) \end{bmatrix}_{n \times p}$$

Linear regression

Let $I_{p \times p}$ denote the identity matrix.

Then $Y_{n \times 1} = A_{n \times p} \beta_{p \times 1} + \epsilon_{n \times 1}$, where $\epsilon \sim N(0, \sigma^2 I_{p \times p})$.

This implies that $Y \sim N(A\beta, \sigma^2 I)$.

Linear regression

$$Y \sim N(A\beta, \sigma^2 I), \quad \text{where} \quad A = [\phi(x_1), \dots, \phi(x_n)]^T .$$

Let $x = (x_1, \dots, x_n)$. Note that

$$
\begin{aligned}
&p(y \mid \beta, \sigma^2, x) \\
&= N(y \mid A\beta, \sigma^2 I) \\
&= \frac{1}{(2\pi)^{n/2} |det(\sigma^2 I)|^{1/2}} \exp\{-1/2(y - A\beta)^T (\sigma^2 I)^{-1/2} (y - A\beta)\}
\end{aligned}
$$

Observations

Note that

$$|det(\sigma^2 I)|^{1/2} = |(\sigma^2)n|^{1/2} = \sigma^n.$$

$$p(y \mid \beta, \sigma^2, x) = \frac{1}{(2\pi)^{n/2}} \sigma^n \exp\{-1/2(y - A\beta)^T (\sigma^2 I)^{-1/2} (y - A\beta)\}$$

Basis functions

A wide range of input-ouput relationships are tackled via basis functions $\phi_1, \ldots, \phi_p$.

- Can handle non-linear relationships between x_i and y_i.
- Each x_i could be complex. Examples: images of different sizes, natural language text, a collection of records/words.

Basis functions transform x_i into a fixed-dimensionality vector of features $(\phi_1(x_i), \ldots, \phi_p(x_i))^T$.

Basis function examples

Linear with intercept

$$\phi(x_i) = (1, x_{i1}, \ldots, x_{id})^T.$$

Quadratic:

$$\phi(x_i) = (1, x_{i1}, \ldots, x_{id}, x_{i1}^2, \ldots, x_{id}^2, x_{i1}x_{i2}, \ldots, x_{i(d-1)x_i d})^T$$

Other basis function examples

- Subset of interaction terms
- Higher order polynomials
- Splines
- Fourier basis (sines and cosines)
- Wavelets

Transformations of basis functions

Suppose our data is categorical (or binary).

- Binary: Use Indicator.

Example: I(subject is hardbook).

- Categorical variable for x_{ij} that can take k values $v_1, \ldots v_k$.
- Transform to $k-1$ dummy variables using:

$$I(x_{ij} = v_1), \ldots, I(x_{ij} = v_{k-1}).$$

Transformations of basis functions

Positive numbers are often transformed using $\log(x)$ as to de-emphasize outliers in the data.

For fractions/proportions, we often use the logit transformation:

$$\text{logit}(x) = \frac{\log(x)}{\log(1-x)}.$$

Controlling flexibility via basis functions

The flexibility of a linear regression model can be controlled via the basis functions.

- We can control the number of variables to use
- We can control which variables to use
- We can control interaction terms
- We can control other types of knobs or tuning parameters that are common in machine learning models

Maximum likelihood estimation for linear regression

Recall that as a function of β and σ^2

$$p(y \mid \beta, \sigma^2, x)$$

is called the likelihood function.

Suppose σ^2 is known.

A common way to estimate the unknown parameters is to maximize the log-likelihood.

Maximum likelihood estimation for linear regression

Recall

$$p(y \mid \beta, \sigma^2, x) = \frac{1}{(2\pi)^{n/2}} \sigma^n \exp\{\frac{-1}{2\sigma^2}(y - A\beta)^T(y - A\beta)\} \implies$$

$$\log p(y \mid \beta, \sigma^2, x) = \text{constant} + \frac{-1}{2\sigma^2}(y - A\beta)^T(y - A\beta)$$

Maximizing the log-likelihood of β is the same as minimizing

$$\begin{aligned} h(\beta) &= (y - A\beta)^T(y - A\beta) \\ &= y^T y - 2\beta^T A^T y + \beta^T A^T A\beta \end{aligned}$$

Maximum likelihood estimation for linear regression

To find the minimizer, set the gradient of $h(\beta)$ to zero:

$$\frac{\partial h(\beta)}{\partial \beta} = -2A^T y + 2A^T A\beta := 0$$

$$\beta = (A^T A)^{-1} A^T y$$

which assumes that $(A^T A)^{-1}$ is invertible. This means that $(A^T A)$ is positive definite.

Is it a minimum (and not just a critical point)?

Verify that the second derivative is > 0.

$$\frac{\partial^2 h(\beta)}{\partial \beta^2} = -2A^T y + 2A^T A\beta = 2A^T A > 0.$$

Thus, our solution is a minimum.

Summary

The maximum likelihood estimator (MLE) is

$$\hat{\beta} = (A^T A)^{-1} A^T y$$

The estimated prediction function is

$$\hat{f}(x_o) = \phi(x_o)^T \hat{\beta}.$$

The MLE for σ^2 is

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

Uncertainty quantification

We can quantify our uncertainty in the estimate $\hat{\beta}$ and in the predictions $\hat{f}(x_o)$ by considering their probability distributions under the assumed model.

We view $\hat{\beta}$ as a random vector, where the randomness comes from the outcomes Y_i in the training data $((x_1, Y_1), \ldots, (x_n, Y_n))$.

The inputs x_i are treated as fixed (non-random).

We can derive the distributions of $\hat{\beta}$, $\hat{f}(x_o)$ and $Y_i - \hat{Y}_i$.

Why are these derivations important?

They are used to construct:

- confidence intervals for the coefficient estimates
- p-values for testing whether coefficients are equal to 0
- confidence intervals for the prediction function
- prediction intervals for future outcomes
- residual diagnostics used in analysis

These distributions are only correct when the linear regression model is correct.

In practice, the regression model is not correct, so we must be thoughtful/careful in analysis always and skeptical.

Distribution of β

Recall

$$Y_{n\times 1} = A_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1} \quad \text{where,}$$

$$\epsilon \sim N(0, \sigma^2 I_{p\times p}).$$

- $Y \in \mathbb{R}^n$ is a random vector
- $A \in \mathbb{R}^{n\times p}$ and $\beta \in \mathbb{R}^p$ are fixed.

$$\begin{aligned}
\hat{\beta} &= (A^T A)^{-1} A^T Y \\
&= (A^T A)^{-1} A^T (A\beta + \epsilon) \\
&= \beta + (A^T A)^{-1} A^T \epsilon \\
&\sim N(\beta, \sigma^2 (A^T A)^{-1})
\end{aligned}$$

Verify the any intermediate steps of the distribution of $\hat{\beta}$ on your own.

Distribution of $\hat{\beta}$

Assuming the model is correct,

$$\hat{\beta} \sim N(\beta, \sigma^2(A^T A)^{-1}).$$

- If σ^2 is known, we can construct confidence intervals for the coefficients β_j:

$$\hat{\beta}_j \pm 1.96\sqrt{(}Var(\hat{\beta}_j)).$$

- Typically, we do not know σ^2 and more derivations are needed for such situations. (See Dunn and Smyth (2018) for details.)

Distribution of $\hat{f}(x_o)$

If the linear model is correct, then

$$\hat{f}(x_o) = \phi(x_o)^T\hat{\beta} \sim N(\phi(x_o)^T\hat{\beta}, \sigma^2\phi(x_o)^T(A^TA)^{-1}\phi(x_o))$$

by the affine transformation property.

Can you verify why this is true?

- If σ^2 is known, we use this formula to construct confidence intervals for $f(x_o)$ and prediction intervals for a new outcome $Y_o = f(x_o) + \epsilon$.
- If σ^2 is unknown, then we need to do more work to construct proper confidence and prediction intervals.

Distribution of the residuals

The residuals are the differences between the observed outcomes Y_i and the fitted outcomes $\hat{Y}_i = \phi(x_i)^T \hat{\beta}$

Let $\hat{Y} = (\hat{Y}_1, \ldots, \hat{Y}_n)^T$.

This implies that

$$\hat{Y} = A\hat{\beta} = A(A^T A)^{-1} A^T Y = HY.$$

where $H = A(A^T A)^{-1} A^T$ is called the hat matrix.

Thus, the vector of residuals is

$$Y - \hat{Y} = Y - HY = (I - H)Y$$

$$\sim N((I - H)A\beta, \sigma^2 (I - H)(I - H)^T)$$

by the affine transformation property since $Y \sim N(A\beta, \sigma^2 I)$.

Distribution of the residuals

- $$HA = A \quad \Longrightarrow (I - H)A\beta = A\beta - HA\beta = 0.$$

- $$H = H^T \quad \text{and} \quad HH = H \quad \Longrightarrow (I - H)(I - H)^T = (I - H).$$

Thus,

$$\begin{aligned} Y - \hat{Y} &\sim N((I - H)A\beta, \sigma^2(I - H)(I - H)^T) \\ &\sim N(0, \sigma^2(I - H)) \end{aligned}$$

Distribution of the residuals

Let H_{ii} denote the ith diagonal entry of H.

If σ^2 is known, can calculate the standardized residuals

$$\frac{Y_i - \hat{Y}_i}{\sigma\sqrt{(1 - H_{ii})}}.$$

This result implies they are $N(0, 1)$ but not independent.

If σ^2 is unknown, you can derive the studentized residuals:

$$\frac{Y_i - \hat{Y}_i}{\hat{\sigma}\sqrt{(1 - H_{ii})}}.$$

The definition of both standardized and studentized residuals varies in the literature, so be aware of this and what definition is being used as it can be confusing.

Leverage

The leverage of a point i is defined as H_{ii}.

Then $\hat{Y}_i = \sum_{j=1}^{n} H_{ij} Y_j$ so if H_{ii} is large then Y_i has a large influence on the fitted value of $\hat{Y}_i$

Identifying high leverage points is a useful diagnostic tool that might have an excessive influence and causing strange results.