Lecture 8:

Linear and Logistic Regression

Regression in general

Regression = method of studying the relationship between a **response variable** Y and a **covariate** X. The covariate is also called a **predictor variable** or **feature.** The relationship between X and Y is summarised by the **regression function**

$$r(x) = \mathbb{E}(Y \mid X = x) = \int y f(y \mid x)\, dy$$

Our goal is to estimate the regression function $r(x)$ from data of the form $(Y_1, X_1), \ldots, (Y_n, X_n) \sim F_{X,Y}$

** The term "regression" is due to Francis Galton (1822-1911) – he noticed that tall and short men tend to have sons with heights closer to the mean – he called this "regression towards the mean"*

Simple Linear Regression

Simple Linear Regression

Simplest version of regression is when X_i is simple (one-dimensional) and $r(x)$ is assumed to be linear: $r(x) = \beta_0 + \beta_1 x$

Definition: The **simple linear regression model**:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where $\mathbb{E}(\varepsilon_i \,|\, X_i) = 0$ and $\mathbb{V}(\varepsilon_i \,|\, X_i) = \sigma^2$ – which, we assume, does not depend on x.

The unknown parameters are the intercept β_0, the slope β_1 and the variance σ^2. Denote the estimates of beta-s with $\widehat{\beta}_0$, $\widehat{\beta}_1$. Then the **fitted line** is $\widehat{r}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x$.

Simple Linear Regression

The unknown parameters are the intercept β_0, the slope β_1 and the variance σ^2. Denote the estimates of beta-s with $\widehat{\beta}_0, \widehat{\beta}_1$. Then the **fitted line** is $\widehat{r}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x$.

The **predicted values** or **fitted values** are $\widehat{Y}_i = \widehat{r}(X_i)$ and the **residuals** are $\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i = Y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 X_i\right)$.

The **residual sum of squares** or RSS = $\sum_{i=1}^{n} \widehat{\varepsilon}_i^2$ – measures how well the line fits the data.

Definition: The **least squares estimates** are values $\widehat{\beta}_0$ and $\widehat{\beta}_1$ that **minimize** the RSS.

Simple Linear Regression

Definition: The **least squares estimates** are values $\widehat{\beta}_0$ and $\widehat{\beta}_1$ that **minimize** the RSS.

Theorem: The least squares estimates are given by:

$$\widehat{\beta}_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \frac{\sum_{i=1}^{n} (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sum_{i=1}^{n} (X_i - \overline{X}_n)^2}$$

$$\widehat{\beta}_0 = \overline{Y}_n - \widehat{\beta}_1 \overline{X}_n$$

And an unbiased estimate of σ^2 is $\widehat{\sigma}^2 = \dfrac{1}{n-2} \displaystyle\sum_{i=1}^{n} \widehat{\varepsilon}_i^2$

Least Squares and Maximum Likelihood

Least Squares and Maximum Likelihood

Suppose we add the assumption that $\varepsilon_i \,|\, X_i \sim \mathcal{N}(0,\sigma^2)$, that is, $Y_i \,|\, X_i \sim \mathcal{N}(\mu_i, \sigma^2)$, where $\mu_i = \beta_0 + \beta_1 X_i$. The likelihood is:

$$\prod_{i=1}^{n} f(X_i, Y_i) = \prod_{i=1}^{n} f_X(X_i) f_{Y|X}(Y_i \,|\, X_i) = \prod f_X \cdot \prod f_{Y|X} = \mathcal{L}_1 \cdot \mathcal{L}_2$$

The term $\mathcal{L}_1$ does not involve the parameters β_0, β_1. We'll focus on the second term $\mathcal{L}_2$, called the **conditional likelihood**

$$\mathcal{L}_2 \equiv \mathcal{L}(\beta_0, \beta_1, \sigma) \propto \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2} \sum_i (Y_i - \mu_i)^2 \right)$$

Least Squares and Maximum Likelihood

So the conditional log-likelihood is

$$\ell(\beta_0, \beta_1, \sigma) = -n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(Y_i - (\beta_0 + \beta_1 X_i)\right)^2,$$

so to find the MLE of (β_0, β_1) we **maximize** $\ell(\beta_0, \beta_1, \sigma)$, which is the same as **minimizing** the RSS = $\sum_{i=1}^{n} \left(Y_i - (\beta_0 + \beta_1 X_i)\right)^2$.

Theorem: Under the assumption of Normality, the least squares estimator is also the maximum likelihood estimator.

Maximizing $\ell(\beta_0, \beta_1, \sigma)$ over σ yields $\widehat{\sigma}^2 = \dfrac{1}{n} \sum_i \widehat{\varepsilon}_i^2$

Properties of Least Squares Estimators

Properties of Least Squares Estimators

Let us look at the properties of the estimators conditional on the data, $X^n = (X_1, \ldots, X_n)$

Theorem: Let $\widehat{\beta}^T = (\widehat{\beta}_0, \widehat{\beta}_1)^T$ denote the LSE. Then,

$$\mathbb{E}(\widehat{\beta} \mid X^n) = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad \text{and} \quad \mathbb{V}(\widehat{\beta} \mid X^n) = \frac{\sigma^2}{n\, s_X^2} \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n X_i^2 & -\overline{X}_n \\ -\overline{X}_n & 1 \end{pmatrix}$$

where $s_X^2 = n^{-1} \sum_{i=1}^n (X_i - \overline{X}_n)^2$.

The estimated s.e.-s of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are obtained taking sqrt-s of the diag. terms of $\mathbb{V}(\widehat{\beta} \mid X^n)$, and inserting the estimate $\widehat{\sigma}$ for σ, thus:

$$\widehat{\text{se}}(\widehat{\beta}_0 \mid X^n) = \frac{\widehat{\sigma}}{s_X\, n} \sqrt{\sum_{i=1}^n X_i^2} \quad \text{and} \quad \widehat{\text{se}}(\widehat{\beta}_1 \mid X^n) = \widehat{\sigma} / (s_X \sqrt{n}).$$

Properties of Least Squares Estimators

Denote $\hat{\text{se}}(\widehat{\beta}_0 \mid X^n)$ and $\hat{\text{se}}(\widehat{\beta}_1 \mid X^n)$ by $\hat{\text{se}}(\widehat{\beta}_0)$ and $\hat{\text{se}}(\widehat{\beta}_1)$.

Theorem: Under appropriate conditions we have:

1. (Consistency): $\widehat{\beta}_0 \xrightarrow{P} \beta_0$ and $\widehat{\beta}_1 \xrightarrow{P} \beta_1$.

2. (Asympt. Normality): $(\widehat{\beta}_0 - \beta_0)/\hat{\text{se}}(\widehat{\beta}_0) \xrightarrow{d} \mathcal{N}(0,1)$, same for $\widehat{\beta}_1$.

3. Approximate $(1-\alpha)$-confidence intervals for β_0, β_1 thus are
$\widehat{\beta}_0 \pm z_{\alpha/2}\,\hat{\text{se}}(\widehat{\beta}_0)$ and $\widehat{\beta}_1 \pm z_{\alpha/2}\,\hat{\text{se}}(\widehat{\beta}_1)$.

4. The Wald test $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ is – reject H_0 if
$|W| > z_{\alpha/2}$ where $W = \widehat{\beta}_1/\hat{\text{se}}(\widehat{\beta}_1)$.

Multiple Regression

Multiple Regression

Now suppose that the covariate is a vector of length k. The data are $(Y_1, X_1), \ldots, (Y_n, X_n)$ where $X_i = (X_{i1}, \ldots, X_{ik})$ – vector of k covariate values for i-th observation. The linear regression model is $Y_i = \sum_{j=1}^{k} \beta_j X_{ij} + \varepsilon_i$, where $\mathbb{E}(\varepsilon_i \,|\, X_{1i}, \ldots, X_{ki}) = 0$.

Usually we want to include an intercept in the model – which we can do by setting $X_{i1} = 1$ for $i = 1,\ldots, n$.

In matrix notation, $Y = (Y_1, \ldots, Y_n)^T$ and $X = \begin{pmatrix} X_{11} & \ldots & X_{1k} \\ \cdots & \cdots & \cdots \\ X_{n1} & \ldots & X_{nk} \end{pmatrix}$ –

each row is one observarion, columns correspond to k covariates. Let $\beta = (\beta_1, \ldots, \beta_k)^T$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$, then $Y = X\beta + \varepsilon$.

Multiple Regression

Theorem: Assuming that the $(k \times k)$ matrix $X^T X$ is invertible,

$$\widehat{\beta} = (X^T X)^{-1} X^T Y$$

$$\mathbb{V}(\widehat{\beta} \mid X^n) = \sigma^2 (X^T X)^{-1} \quad \text{and} \quad \widehat{\beta} \approx \mathcal{N}(\beta, \sigma^2 (X^T X)^{-1}).$$

The estimate regression function is $\widehat{r}(x) = \sum_{j=1}^{k} \widehat{\beta}_j x_j$. An unbiased estimate of σ^2 is $\widehat{\sigma}^2 = \frac{1}{n-k} \sum_{i=1}^{n} \widehat{\varepsilon}_i^2$ where $\widehat{\varepsilon} = X\widehat{\beta} - Y$ is the vector of residuals.

An approximate $(1-\alpha)$-confidence interval is $\widehat{\beta}_j \pm z_{\alpha/2}\, \widehat{\text{se}}(\widehat{\beta}_j)$ where $\widehat{\text{se}}^2(\widehat{\beta}_j)$ is the j-th diag. element of the matrix $\widehat{\sigma}^2 (X^T X)^{-1}$.

Logistic Regression

Logistic Regression

So far we assumed Y_i are real-valued. In **Logistic regression**, $Y_i \in \{0,1\}$ is binary. For a k-dimensional covariate X, the model is

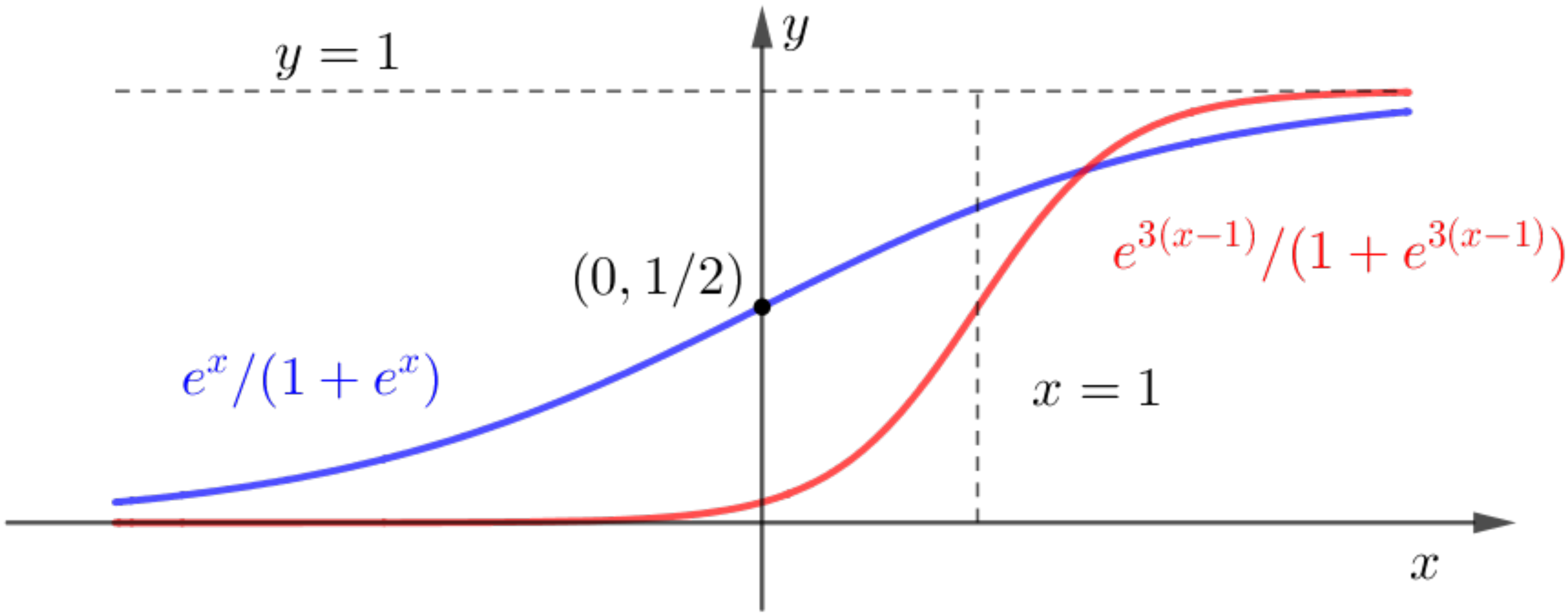
$$p_i(\beta) \equiv \mathbb{P}(Y_i = 1 \,|\, X = x) = \frac{\exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}\right)}{1 + \exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}\right)}$$

or, equiv., logit$(p_i) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$ where logit$(p) = \log \frac{p}{1-p}$

The name "logistic regression" comes from $\dfrac{e^x}{1+e^x}$ – the **logistic function**.

Logistic Regression

$f(x) = e^x/(1 + e^x)$ – the **logistic function** – maps $f : \mathbb{R} \rightarrow (0,1)$ real numbers to probabilities



With a **linear** transform of $x \rightarrow ax + b$, we **adjust** the position of the "decision boundary", and **scale** the "sharpness" of it

Logistic Regression

Because the Y_i are binary, data are $Y_i \,|\, X_i = x_i \sim$ Bernoulli(p_i), so the likelihood function is

$$\mathscr{L}(\beta) = \prod_{i=1}^{n} p_i(\beta)^{Y_i} (1 - p_i(\beta))^{1-Y_i}$$

the MLE is obtained by maximizing $\log \mathscr{L}(\beta)$ numerically.

One way to do so is the Reweighted Least Squares algorithm

Logistic Regression

Reweighted Least Squares algorithm: Choose starting values $\widehat{\beta}^0 = (\widehat{\beta}_0^0, \ldots, \widehat{\beta}_k^0)$ and compute p_i^0 (logistic function). Set $s = 0$ and iterate until convergence:

1. Set $Z_i = \text{logit}(p_i^s) + \dfrac{Y_i - p_i^s}{p_i^s(1 - p_i^s)}, \quad i = 1,\ldots,n$

2. Let W be a diag. matrix with (i, i) element $W_{ii} = p_i^s(1 - p_i^s)$

3. Set $\widehat{\beta}^s = (X^T W X)^{-1} X^T W X$ – weighted linear reg. of Z on Y.

4. Set $s = s + 1$ and back to 1-st step.