Statistics

March 31, 2021

Bayesian Hypothesis Testing

Consider $H_0 : \theta = \theta_0$

vs. $H_1 : \theta \neq \theta_0$

Technique: Put a prior on θ and H_0, then compute: $\mathbb{P}(H_0 \mid \vec{x})$

this is our observed data.

Ex: Put the prior $\mathbb{P}(H_0) = \frac{1}{2}$, $\mathbb{P}(H_1) = 1/2$.

Then compute $\mathbb{P}(H_0 \mid \vec{x})$

$$\mathbb{P}(H_0 \mid \vec{x}) = \frac{f(\vec{x}, H_0)}{f(\vec{x})}$$

$$= \frac{f(\vec{x} \mid H_0) \cdot \cancel{\mathbb{P}(H_0)}}{f(\vec{x} \mid H_0)\, \cancel{\mathbb{P}(H_0)} + f(\vec{x} \mid H_1)\, \cancel{\mathbb{P}(H_1)}}$$

in our case, $\mathbb{P}(H_0) = \mathbb{P}(H_1)$

$$= \frac{f(\vec{x} \mid \theta_0)}{f(\vec{x} \mid \theta_0) + f(\vec{x} \mid H_1)}$$

$$= \frac{f(\vec{x} \mid \theta_0)}{f(\vec{x} \mid \theta_0) + \int f(\vec{x} \mid \theta)\, f(\theta)\, d\theta}$$

$$= \frac{\mathcal{L}(\theta_0)}{\mathcal{L}(\theta_0) + \int \mathcal{L}(\theta)\, f(\theta)\, d\theta}$$

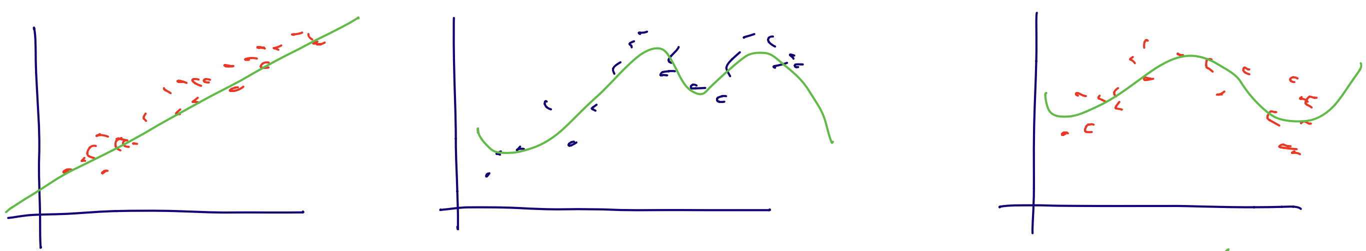
Notes

- prior f can have a large influence on $H_0 | \vec{x}$
- Improper priors are not allowed
- $\mathbb{P}(H_0 | \vec{x})$ is the probability that H_0 is true given $\vec{x}$ → this does not tell us when to reject the null hypothesis. When do we reject? When do we retain? We need more detailed analysis.

Read 11.9 in All of Stats for more strengths/weaknesses.

Regression (standard linear regression)

Goal: Fit noisy data using a curve.



$y = ax + b$ — parametric

$y = a_0 + a_1 x + a_2 x^2 + \dots + a_n x^n$ — parametric

$y | \vec{x}, \vec{y} \sim GP(m, k)$ — non-parametric

Denote: Y ~ random variable, response variable

X ~ covariate, predictor, feature

Regression: $r(x) = \mathbb{E}(Y | \vec{X} = \vec{x}) = \int y \, f(y | \vec{x}) \, dy$

Goal: Given data $(X_1, Y_1) \dots (X_n, Y_n) \sim F_{XY}$, estimate $r(x)$.

model

Basic Linear Regression

Model: $r(x) = \beta_0 + \beta_1 x$ "simple linear regression model"

Observe some data: X_i, Y_i

Assumption: $\mathrm{Var}(Y \mid X = x) = \sigma^2$

$\hookrightarrow$ $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

$\mathbb{E}(\epsilon_i \mid X_i) = 0$

$\mathrm{Var}(\epsilon_i \mid X_i) = \sigma^2.$

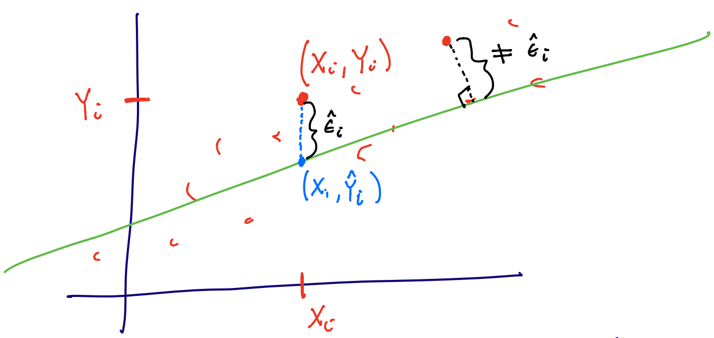
Given this data for the statistical model, find estimators for the unknown coefficients $\beta_0, \beta_1 \to \hat\beta_0, \hat\beta_1$.

Fitted line: $\hat r(x) = \hat\beta_0 + \hat\beta_1 x$

Predicted values: $\hat Y_i = \hat r(x_i)$

Residuals: $\hat\epsilon_i = Y_i - \hat Y_i = Y_i - (\hat\beta_0 + \hat\beta_1 x_i)$

Residual sum of squares: $RSS = \sum \hat\epsilon_i^2$ ← only one such metric to determine how well $\hat r$ fits the data.



Definition:

Least squares estimates: $\hat\beta_0, \hat\beta_1$ minimize $RSS = \sum_{i=1}^{n} \hat\epsilon_i^2$.

$\hat\beta_0, \hat\beta_1$ can be found using calculus, linear algebra, statistics, etc.

<u>Thm</u>: $\hat{\beta}_1 = \left.\dfrac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}\right\} \approx \dfrac{\operatorname{cov}(X,Y)}{\operatorname{var}(X)}$

$= \dfrac{\sigma_X \sigma_Y \rho_{XY}}{\sigma_X^2}$

and $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

$= \dfrac{\sigma_Y}{\sigma_X} \rho_{XY}$

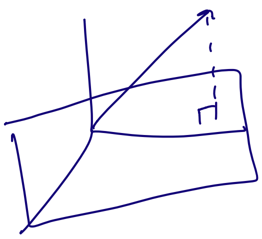
Unbiased estimate of σ^2:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum \hat{\epsilon}_i^2 .$$

To find $\hat{\beta}_0, \hat{\beta}_1$ using calculus: solve $\dfrac{\partial}{\partial \beta_0} RSS = 0$

$\dfrac{\partial}{\partial \beta_1} RSS = 0$

To solve using linear algebra:



compute the orthogonal projection of Y_i onto span$\{\vec{1}, \vec{x}\}$

Least Squares and Maximum Likelihood Estimators

Add assumption that $\epsilon_i | x_i \sim N(0, \sigma^2)$

$$\Rightarrow Y_i | X_i \sim N(\mu_i, \sigma^2)$$

$\hookrightarrow \mu_i = \beta_0 + \beta_1 x_i$

Write down the likelihood function:

$$\mathcal{L} \propto \prod f(X_i, Y_i) = \underbrace{\prod f(X_i)}_{} \cdot \underbrace{f(Y_i | X_i)}_{}$$

$$= \mathcal{L}_1 \times \mathcal{L}_2(\beta_0, \beta_1, \sigma^2)$$

$\mathcal{L}_1$ does not depend on any parameters.

$\mathcal{L}_2$ is known as the conditional likelihood and contains all the parameters.

$$\mathcal{L}_2(\beta_0, \beta_1, \sigma^2) \propto \frac{1}{\sigma^n} e^{-\frac{1}{2\sigma^2} \sum_i (Y_i - \beta_0 - \beta_1 X_i)^2}$$

$$\Rightarrow \log \mathcal{L}_2 = l_2 = -n \log \sigma - \frac{1}{2\sigma^2} \underbrace{\sum (Y_i - \beta_0 - \beta_1 X_i)^2}$$

Thm If $\epsilon_i | X_i \sim N(0, \sigma^2)$, then the MLE for β_0, β_1 is the same as the least squares estimate, and $\hat{\sigma}^2_{MLE} = \underbrace{\frac{1}{n} \sum \hat{\epsilon}_i^2}_{\text{biased estimator.}}$

Properties of these least squares estimators

Thm: Define $\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}$

then $\mathbb{E}(\hat{\beta} | \vec{X}) = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$

$$Var(\hat{\beta} | \vec{X}) = \frac{\sigma^2}{n S_X^2} \begin{pmatrix} \frac{1}{n} \sum X_i^2 & -\bar{X} \\ -\bar{X} & 1 \end{pmatrix}$$

where $S_X^2 = \frac{1}{n} \sum (X_i - \bar{X})^2$

The estimated standard errors are (sqrt of diagonals of covariance matrix)

$$\hat{se}(\hat{\beta}_0) = \frac{\hat{\sigma}}{\sqrt{n}\, s_X} \sqrt{\frac{1}{n} \sum X_i^2}$$

$$\hat{se}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{n}\, s_X}$$

Thm These estimators are

- consistent
- asymptotically normal
- and therefore we can apply the Wald test,

e.g. $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$.

Prediction

Setup: Have the data $X_1, Y_1 \ldots X_n, Y_n$ and estimate $\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ using least squares.

Observe (or pick) a new covariate $X = x_*$, and we want to predict Y_*.

Estimate $\hat{Y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_*$,

$$\mathrm{Var}(\hat{Y}_*) = \mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_*)$$

$$= \mathrm{var}(\hat{\beta}_0) + x_*^2 \,\mathrm{var}(\hat{\beta}_1) + 2 x_* \,\mathrm{cov}(\hat{\beta}_0, \hat{\beta}_1)$$

and $\hat{se}(\hat{Y}_*) = \sqrt{\mathrm{var}(\hat{Y}_*)}$ using $\hat{\sigma}^2$.

How about a $(1-\alpha)$ confidence interval for Y_*?

Naively, $\hat{Y}_* \pm z_{\alpha/2}\, \hat{se}(\hat{Y}_*)$ is a $1-\alpha$ confidence interval,

but this is incorrect. (See Thm 13.11, do exercise 10)

The idea behind the mistake:

the above confidence interval is only correct if we never observed the independent noise ϵ_i, i.e. in the real world we observe $Y_* = \beta_0 + \beta_1 X_* + \epsilon$.

Multiple Regression

Setup: $\vec{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$, covariates $X = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1k} \\ X_{21} & & & \vdots \\ \vdots & & & \vdots \\ X_{n1} & & & X_{nk} \end{pmatrix}$

2nd covariate.

$$\vec{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \vec{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Model: $\vec{Y} = X\vec{\beta} + \vec{\epsilon}$

See Thm 13.13 for the least squares solution, same as from linear algebra class.

Logistic Regression

Change the model: Imagine that $Y_i \in \{0,1\}$, i.e. $\mathbb{P}(Y_i \mid X_i) = p_i$

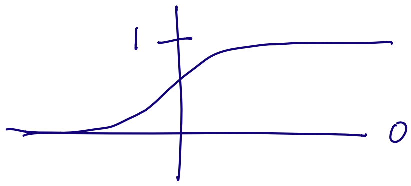
We want to model p_i, not Y_i.

Choose a particular parametric form:

$$p_i = p_i(\beta_0, \ldots, \beta_k) = \mathbb{P}(Y_i = 1 \mid X_i)$$

$$= \frac{e^{\beta_0 + \sum_1^k \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_1^k \beta_j x_{ij}}}$$

The logistic function: $\dfrac{e^x}{1+e^x}$



$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

Since $Y_i \mid X_i \sim \text{Bernoulli}(p_i)$

The conditional likelihood is:

$$\mathcal{L}(\vec{\beta}) = \prod_i p_i(\vec{\beta})^{Y_i} \left(1 - p_i(\vec{\beta})\right)^{1-Y_i}$$

$\mathcal{L}$ must be maximized numerically.

Multivariate Models

Random vector $\vec{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix}$, mean $\mathbb{E}(\vec{X}) = \begin{pmatrix} \mathbb{E}(X_1) \\ \mathbb{E}(X_2) \\ \vdots \\ \mathbb{E}(X_k) \end{pmatrix} = \vec{\mu}$

Covariance matrix

$$C = \underbrace{\begin{pmatrix} \mathrm{cov}(X_1, X_1) & \mathrm{cov}(X_1, X_2) & \cdots & \mathrm{cov}(X_1, X_k) \\ \mathrm{cov}(X_2, X_1) & & & \vdots \\ \mathrm{cov}(X_3, X_1) & & & \vdots \\ \vdots & & & \vdots \\ \mathrm{cov}(X_k, X_1) & & & \mathrm{cov}(X_k, X_k) \end{pmatrix}}_{k \times k \text{ matrix.}}$$

C^{-1} is known as the <u>precision matrix</u>,

Furthermore: - C is symmetric positive definite

- eigenvalues of C are all positive.

<u>Thm</u> Let $\vec{a} \in \mathbb{R}^k$ be a constant vector,

then ① $\mathbb{E}(\vec{a}^T \vec{X}) = \vec{a}^T \vec{\mu}$

② $\mathrm{Var}(\vec{a}^T \vec{X}) = \vec{a}^T C \vec{a}$

Let $A \in \mathbb{R}^{n \times k}$ constant matrix, then

① $\mathbb{E}(A\vec{X}) = A\vec{\mu}$

② $\mathrm{Var}(A\vec{X}) = ACA^T$ } another covariance matrix

Next consider we have samples

$$\begin{matrix} X_{11}, X_{12}, \ldots, X_{1n} \\ \vdots \\ X_{k1}, \ldots\ldots X_{kn} \end{matrix}$$

The sample mean is then

$$\vec{X} = \begin{pmatrix} \bar{X}_1 = \frac{1}{n} \sum_j X_{1j} \\ \vdots \\ \bar{X}_k = \frac{1}{n} \sum_j X_{kn} \end{pmatrix}$$

The sample variance matrix:

$$S = \begin{pmatrix} S_{11} & S_{12} & \cdots \\ & \ddots & \\ & & S_{kk} \end{pmatrix}$$

where $s_{ij} = \frac{1}{n-1} \sum_{l=1}^{n} (X_{il} - \bar{X}_i)(X_{jl} - \bar{X}_j)$ } unbiased estimate of $\mathrm{cov}(X_i, X_j)$.

$$\mathbb{E}(\bar{X}) = \vec{\mu} \quad , \quad \mathbb{E}(S) = C.$$

And since the correlation of two variables is

$$\rho(X_i, X_j) = \frac{\mathrm{Cov}(X_i, X_j)}{\sqrt{\mathrm{Var}(X_i)\ \mathrm{Var}(X_j)}},$$

the plugin estimator for the correlation is:

$$\hat{\rho}_{ij} = \frac{S_{ij}}{S_{ii}\ S_{jj}}$$

Example Multivariate Normal

$\vec{X} \in \mathbb{R}^k \sim N(\vec{\mu}, C)$ if its density is

$$f(x_1, \dots, x_k ; \vec{\mu}, C) = \frac{1}{(2\pi)^{k/2}} \frac{1}{\sqrt{\det C}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T C^{-1} (\vec{x}-\vec{\mu})}$$

$\vec{\mu} \in \mathbb{R}^k$

$C \in \mathbb{R}^{k \times k}$

$\Rightarrow \mathbb{E}(\vec{X}) = \vec{\mu}$

$\mathrm{Var}(\vec{X}) = C.$

Thm Let $\vec{Z} \sim N(\vec{0}, I)$, and C be a spd matrix.

(1) Let $C^{1/2}$ be such that $C^{1/2} \cdot C^{1/2} = C.$, then

$$\vec{X} = \vec{\mu} + C^{1/2} \vec{Z} \sim N(\vec{\mu}, C).$$

(2) $C^{-1/2}(\vec{X} - \vec{\mu}) \sim N(\vec{0}, I).$

(3) $\vec{a}^T \vec{X} \sim N(\vec{a}^T \vec{\mu}, \vec{a}^T C \vec{a})$

(4) $V = (\vec{X} - \mu)^T C^{-1} (\vec{X} - \mu)$, then $V \sim \chi^2_k$

$= \vec{Z}^T \vec{Z}.$

Thm For multivariate dat $\begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{k1} \end{pmatrix} \cdots \begin{pmatrix} x_{1n} \\ \vdots \\ x_{kn} \end{pmatrix}$, the log-likelihood

$$\ell(\vec{\mu}, C) \propto \frac{-n}{2} (\bar{x} - \vec{\mu})^T C^{-1} (\bar{x} - \vec{\mu}) - \frac{n}{2} \mathrm{tr}(C^{-1} S) - \frac{n}{2} \log \det C$$

Recall that $tr(A) = \sum A_{ii}$,

S = as before, the sample covariance matrix.

The MLE's are $\hat{\mu} = \overline{X}$, $\hat{C} = \frac{n-1}{n} S$

Gaussian Processes

Simplest interpretation: the extension of random vectors to random functions

Best single source reference:

Rasmussen & Williams, Gaussian Processes for Machine Learning

We say that $f \sim GP(m, k)$ is a Gaussian process with mean m and covariance function k.

$$\Rightarrow \quad \mathbb{E}(f(x)) = m(x)$$

$$\mathrm{Cov}(f(x), f(x')) = k(x, x')$$

$$\Rightarrow \quad \mathbb{E}\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix} = \begin{pmatrix} m(x_1) \\ m(x_2) \\ \vdots \\ m(x_n) \end{pmatrix}$$

$$\mathrm{Cov}\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix} = K \quad \text{with entries } k(x_i, x_j)$$

$$\Leftrightarrow \quad k(x, x') = \mathbb{E}\Big(\big(f(x) - m(x)\big)\big(f(x') - m(x')\big) \Big)$$

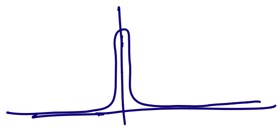
Example covariance functions:

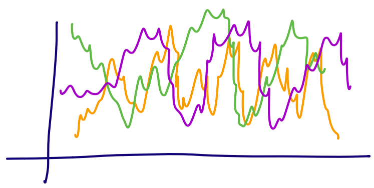
$$k(x,x') = A e^{-(x-x')^2/b}$$

$$k(x,x') = B e^{-|x-x'|/c}$$ } one function from the Matérn family of covariance kernels.

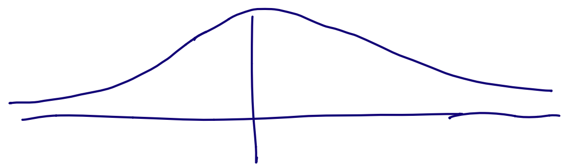
Graphically

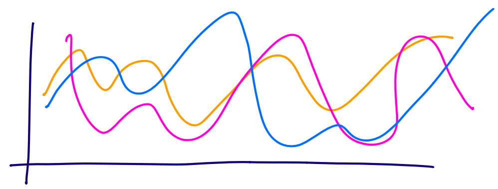
$$k(x,x') = e^{-(x-x')^2/.0001}$$



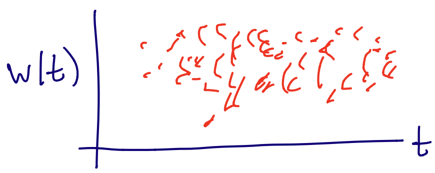


$$k(x,x') = e^{-(x-x')^2/10000}$$





$$k(x,x') = \delta(x-x') = \begin{cases} 1 & \text{if } x=x' \\ 0 & \text{otherwise} \end{cases}$$



Discontinuous everywhere.

Brownian Motion $B(t) = \int_0^t w(\tau)\, d\tau$

$\uparrow$ GP(0, δ)