The Book of Statistical Proofs

Joram Soch, BCCN Berlin
joram.soch@bccn-berlin.de

https://github.com/JoramSoch/TBSP

Contents

Chapter I

General Theorems

1 Probability theory: Bayes' theorem

Index:
▷ The Book of Statistical Proofs
  ▷ General Theorems
    ▷ Probability theory: Bayes' theorem

Theorem: Let A and B be two arbitrary statements about random variables, such as statements about the presence or absence of an event or about the value of a scalar, vector or matrix. Then, the conditional probability that A is true, given that B is true, is equal to

$$p(A|B) = \frac{p(B|A)\,p(A)}{p(B)} \; . \tag{1}$$

Proof: The conditional probability is defined as the ratio of joint probability, i.e. the probability of both statements being true, and marginal probability, i.e. the probability of only the second one being true:

$$p(A|B) = \frac{p(A,B)}{p(B)} \; . \tag{2}$$

It can also be written down for the reverse situation, i.e. to calculate the probability that B is true, given that A is true:

$$p(B|A) = \frac{p(A,B)}{p(A)} \; . \tag{3}$$

Both equations can be rearranged for the joint probability

$$p(A|B)\,p(B) \overset{(2)}{=} p(A,B) \overset{(3)}{=} p(B|A)\,p(A) \tag{4}$$

from which Bayes' theorem can be directly derived:

$$p(A|B) \overset{(4)}{=} \frac{p(B|A)\,p(A)}{p(B)} \; . \tag{5}$$

■

Dependencies:
- law of conditional probability, also called "product rule of probability"

Source:
- Koch, Karl-Rudolf (2007): *Introduction to Bayesian Statistics*, second edition, Springer, Berlin/Heidelberg, 2007, pp. 6/13, eqs. 2.12/2.38.

Index: number: A004; shortcut: prob-bayes; author: JoramSoch; date: 2019/05/03.

2 Estimation theory: Partition of mean squared error into bias and variance

Index:
▷ The Book of Statistical Proofs
  ▷ General Theorems
    ▷ Estimation theory: Partition of mean squared error into bias and variance

Theorem: The mean squared error can be partitioned into variance and squared bias

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) - \mathrm{Bias}(\hat{\theta}, \theta)^2 \tag{1}$$

where the variance is given by

$$\mathrm{Var}(\hat{\theta}) = \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right)^2\right] \tag{2}$$

and the bias is given by

$$\mathrm{Bias}(\hat{\theta}, \theta) = \left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right) \; . \tag{3}$$

Proof: The mean squared error (MSE) is defined as the expected value of the squared deviation of the estimated value $\hat{\theta}$ from the true value θ of a parameter, over all values $\hat{\theta}$:

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \theta\right)^2\right] \; . \tag{4}$$

This formula can be evaluated in the following way:

$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}) &= \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \theta\right)^2\right] \\
&= \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta}) + \mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)^2\right] \\
&= \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right)^2 + 2\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right)\left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right) + \left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)^2\right] \\
&= \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right)^2\right] + \mathbb{E}_{\hat{\theta}}\left[2\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right)\left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)\right] + \mathbb{E}_{\hat{\theta}}\left[\left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)^2\right] \; .
\end{aligned} \tag{5}$$

Because $\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta$ is constant as a function of $\hat{\theta}$, we have:

$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}) &= \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right)^2\right] + 2\left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)\mathbb{E}_{\hat{\theta}}\left[\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right] + \left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)^2 \\
&= \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right)^2\right] + 2\left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)\left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right) + \left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)^2 \\
&= \mathbb{E}_{\hat{\theta}}\left[\left(\hat{\theta} - \mathbb{E}_{\hat{\theta}}(\hat{\theta})\right)^2\right] + \left(\mathbb{E}_{\hat{\theta}}(\hat{\theta}) - \theta\right)^2
\end{aligned} \tag{6}$$

This proofs the partition given by (1). ■

Dependencies:
- definition of the mean squared error
- expectation with respect to a random variable

Source:
- Wikipedia: "Mean squared error"; in: *Wikipedia, the free encyclopedia*; URL: https://en.wikipedia.org/wiki/Mean_squared_error#Proof_of_variance_and_bias_relationship.

Metadata: ID: A005 | name: mse-bnv | author: JoramSoch | date: 2019-05-06.

Chapter II

Probability Distributions

1 Multivariate normal distribution: Linear transformation theorem

Index:
▷ The Book of Statistical Proofs
  ▷ Probability Distributions
    ▷ Multivariate normal distribution: Linear transformation theorem

Theorem: Let x follow a multivariate normal distribution:

$$x \sim \mathrm{N}(\mu, \Sigma) \; . \tag{1}$$

Then, any linear transformation of x is also multivariate normally distributed:

$$y = Ax + b \sim \mathrm{N}(A\mu + b, A\Sigma A^T) \; . \tag{2}$$

Proof: The moment-generating function of a random vector x is

$$M_x(t) = \mathbb{E}\left(\exp\left[t^{\mathrm{T}}x\right]\right) \tag{3}$$

and therefore the moment-generating function of the random vector y is given by

$$\begin{aligned} M_y(t) &= \mathbb{E}\left(\exp\left[t^{\mathrm{T}}(Ax+b)\right]\right) \\ &= \mathbb{E}\left(\exp\left[t^{\mathrm{T}}Ax\right] \cdot \exp\left[t^{\mathrm{T}}b\right]\right) \\ &= \exp\left[t^{\mathrm{T}}b\right] \cdot \mathbb{E}\left(\exp\left[t^{\mathrm{T}}Ax\right]\right) \\ &= \exp\left[t^{\mathrm{T}}b\right] \cdot M_x(At) \; . \end{aligned} \tag{4}$$

The joint moment-generating function of the multivariate normal distribution is

$$M_x(t) = \exp\left[t^{\mathrm{T}}\mu + \frac{1}{2}t^{\mathrm{T}}\Sigma t\right] \tag{5}$$

and therefore the moment-generating function of the random vector y becomes

$$\begin{aligned} M_y(t) &= \exp\left[t^{\mathrm{T}}b\right] \cdot M_x(At) \\ &= \exp\left[t^{\mathrm{T}}b\right] \cdot \exp\left[t^{\mathrm{T}}A\mu + \frac{1}{2}t^{\mathrm{T}}A\Sigma A^{\mathrm{T}}t\right] \\ &= \exp\left[t^{\mathrm{T}}\left(A\mu + b\right) + \frac{1}{2}t^{\mathrm{T}}A\Sigma A^{\mathrm{T}}t\right] \; . \end{aligned} \tag{6}$$

Because moment-generating function and probability density function of a random variable are equivalent, this demonstrates that y is following a multivariate normal distribution with mean $A\mu + b$ and covariance $A\Sigma A^{\mathrm{T}}$. ■

Dependencies:
- moment-generating function of a random vector
- joint moment-generating function of the multivariate normal distribution

Source:
- Taboga, Marco (2010): "Linear combinations of normal random variables"; in: *Lectures on probability and statistics*; URL: https://www.statlect.com/probability-distributions/normal-distribution-linear-combinations.

Index: number: A001; shortcut: mvn-ltt; author: JoramSoch; date: 2019/05/02.

2 Normal-gamma distribution: Kullback-Leibler divergence

Index:
▷ The Book of Statistical Proofs
  ▷ Probability Distributions
    ▷ Normal-gamma distribution: Kullback-Leibler divergence

Theorem: Let $x \in \mathbb{R}^k$ be a random vector and $y > 0$ be a random variable. Assume two normal-gamma distributions P and Q specifying the joint distribution of x and y as

$$\begin{aligned} P: \; (x,y) &\sim \mathrm{NG}(\mu_1, \Lambda_1^{-1}, a_1, b_1) \\ Q: \; (x,y) &\sim \mathrm{NG}(\mu_2, \Lambda_2^{-1}, a_2, b_2) \; . \end{aligned} \tag{1}$$

Then, the Kullback-Leibler divergence of P from Q is given by

$$\begin{aligned} \mathrm{KL}[P \,||\, Q] = \frac{1}{2}\frac{a_1}{b_1}\left[(\mu_2-\mu_1)^T\Lambda_2(\mu_2-\mu_1)\right] + \frac{1}{2}\,\mathrm{tr}(\Lambda_2\Lambda_1^{-1}) - \frac{1}{2}\ln\frac{|\Lambda_2|}{|\Lambda_1|} - \frac{k}{2} \\ + a_2\,\ln\frac{b_1}{b_2} - \ln\frac{\Gamma(a_1)}{\Gamma(a_2)} + (a_1-a_2)\,\psi(a_1) - (b_1-b_2)\,\frac{a_1}{b_1} \; . \end{aligned} \tag{2}$$

Proof: The probabibility density function of the normal-gamma (NG) distribution is

$$p(x,y) = p(x|y) \cdot p(y) = \mathcal{N}(x; \mu, (y\Lambda)^{-1}) \cdot \mathrm{Gam}(y; a, b) \tag{3}$$

where $\mathcal{N}(x; \mu, \Sigma)$ is a multivariate normal density with mean μ and covariance Σ (hence, precision Λ) and $\mathrm{Gam}(y; a, b)$ is a univariate gamma density with shape a and rate b. The Kullback-Leibler (KL) divergence of the multivariate normal distribution is

$$\mathrm{KL}[P \,||\, Q] = \frac{1}{2}\left[(\mu_2-\mu_1)^T\Sigma_2^{-1}(\mu_2-\mu_1) + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1) - \ln\frac{|\Sigma_1|}{|\Sigma_2|} - k\right] \tag{4}$$

and the Kullback-Leibler divergence of the univariate gamma distribution is

$$\mathrm{KL}[P \,||\, Q] = a_2\,\ln\frac{b_1}{b_2} - \ln\frac{\Gamma(a_1)}{\Gamma(a_2)} + (a_1-a_2)\,\psi(a_1) - (b_1-b_2)\,\frac{a_1}{b_1} \tag{5}$$

where $\Gamma(x)$ is the gamma function and $\psi(x)$ is the digamma function.

The KL divergence for a continuous random variable is given by

$$\mathrm{KL}[P \,||\, Q] = \int_Z p(z)\, \ln \frac{p(z)}{q(z)}\, \mathrm{d}z \tag{6}$$

which, applied to the normal-gamma distribution over x and y, yields

$$\mathrm{KL}[P \,||\, Q] = \int_0^\infty \int_{\mathbb{R}^k} p(x,y)\, \ln \frac{p(x,y)}{q(x,y)}\, \mathrm{d}x\, \mathrm{d}y\ . \tag{7}$$

Using the law of conditional probability, this can be evaluated as follows:

$$\begin{aligned}
\mathrm{KL}[P \,||\, Q] &= \int_0^\infty \int_{\mathbb{R}^k} p(x|y)\, p(y)\, \ln \frac{p(x|y)\, p(y)}{q(x|y)\, q(y)}\, \mathrm{d}x\, \mathrm{d}y \\
&= \int_0^\infty \int_{\mathbb{R}^k} p(x|y)\, p(y)\, \ln \frac{p(x|y)}{q(x|y)}\, \mathrm{d}x\, \mathrm{d}y + \int_0^\infty \int_{\mathbb{R}^k} p(x|y)\, p(y)\, \ln \frac{p(y)}{q(y)}\, \mathrm{d}x\, \mathrm{d}y \\
&= \int_0^\infty p(y) \int_{\mathbb{R}^k} p(x|y)\, \ln \frac{p(x|y)}{q(x|y)}\, \mathrm{d}x\, \mathrm{d}y + \int_0^\infty p(y)\, \ln \frac{p(y)}{q(y)} \int_{\mathbb{R}^k} p(x|y)\, \mathrm{d}x\, \mathrm{d}y \\
&= \left\langle \mathrm{KL}[p(x|y) \,||\, q(x|y)] \right\rangle_{p(y)} + \mathrm{KL}[p(y) \,||\, q(y)]\ .
\end{aligned} \tag{8}$$

In other words, the KL divergence between two normal-gamma distributions over x and y is equal to the sum of a multivariate normal KL divergence regarding x conditional on y, expected over y, and a univariate gamma KL divergence regarding y.
From equations (3) and (4), the first term becomes

$$\begin{aligned}
&\left\langle \mathrm{KL}[p(x|y) \,||\, q(x|y)] \right\rangle_{p(y)} \\
&= \left\langle \frac{1}{2} \left[(\mu_2 - \mu_1)^T (y\Lambda_2)(\mu_2 - \mu_1) + \mathrm{tr}\left((y\Lambda_2)(y\Lambda_1)^{-1} \right) - \ln \frac{|(y\Lambda_1)^{-1}|}{|(y\Lambda_2)^{-1}|} - k \right] \right\rangle_{p(y)} \\
&= \left\langle \frac{y}{2} (\mu_2 - \mu_1)^T \Lambda_2 (\mu_2 - \mu_1) + \frac{1}{2}\, \mathrm{tr}(\Lambda_2 \Lambda_1^{-1}) - \frac{1}{2} \ln \frac{|\Lambda_2|}{|\Lambda_1|} - \frac{k}{2} \right\rangle_{p(y)}
\end{aligned} \tag{9}$$

and using the relation $y \sim \mathrm{Gam}(a,b) \Rightarrow \langle y \rangle = a/b$, we have

$$\left\langle \mathrm{KL}[p(x|y) \,||\, q(x|y)] \right\rangle_{p(y)} = \frac{1}{2} \frac{a_1}{b_1} (\mu_2 - \mu_1)^T \Lambda_2 (\mu_2 - \mu_1) + \frac{1}{2}\, \mathrm{tr}(\Lambda_2 \Lambda_1^{-1}) - \frac{1}{2} \ln \frac{|\Lambda_2|}{|\Lambda_1|} - \frac{k}{2}\ . \tag{10}$$

By plugging (10) and (5) into (8), one arrives at the KL divergence given by (2). ■

Dependencies:

- probability density function of the normal-gamma distribution
- Kullback-Leibler divergence of the multivariate normal distribution
- Kullback-Leibler divergence of the univariate gamma distribution
- Kullback-Leibler divergence for a continuous random variable
- law of conditional probability, also called "product rule of probability"
- expected value of a gamma random variable

Source:

- Soch & Allefeld (2016): "Kullback-Leibler Divergence for the Normal-Gamma Distribution". *arXiv math.ST*, 1611.01437; URL: https://arxiv.org/abs/1611.01437.

Metadata: ID: A006 | name: ng-kl | author: JoramSoch | date: 2019-05-07.

Chapter III

Statistical Models

1 Mutiple linear regression: Ordinary least squares

Index:
▷ The Book of Statistical Proofs
  ▷ Statistical Models
    ▷ Mutiple linear regression: Ordinary least squares

Theorem: Given a linear regression model with independent observations

$$y = X\beta + \varepsilon, \; \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \; , \tag{1}$$

the parameters minimizing the residual sum of squares are given by

$$\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y \; . \tag{2}$$

Proof: Let $\hat{\beta}$ be the ordinary least squares (OLS) solution and let $\hat{\varepsilon} = y - X\hat{\beta}$ be the resulting vector of residuals. Then, this vector must be orthogonal to the design matrix,

$$X^{\mathrm{T}}\hat{\varepsilon} = 0 \; , \tag{3}$$

because if it wasn't, there would be another solution $\tilde{\beta}$ giving another vector $\tilde{\varepsilon}$ with a smaller residual sum of squares. From (3), the OLS formula can be directly derived:

$$\begin{aligned} X^{\mathrm{T}}\hat{\varepsilon} &= 0 \\ X^{\mathrm{T}}\left(y - X\hat{\beta}\right) &= 0 \\ X^{\mathrm{T}}y - X^{\mathrm{T}}X\hat{\beta} &= 0 \\ X^{\mathrm{T}}X\hat{\beta} &= X^{\mathrm{T}}y \\ \hat{\beta} &= (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y \; . \end{aligned} \tag{4}$$

■

Source:
- Stephan, Klaas Enno (2010): "The General Linear Model (GLM)"; in: *Methods and models for fMRI data analysis in neuroeconomics*; URL: http://www.socialbehavior.uzh.ch/teaching/methodsspring10.html

Index: number: A002; shortcut: mlr-ols; author: JoramSoch; date: 2019/05/02.

2 General linear model: Maximum likelihood estimation

Index:
▷ The Book of Statistical Proofs
  ▷ Statistical Models
    ▷ General linear model: Maximum likelihood estimation

Theorem: Given a general linear model with matrix-normally distributed errors

$$Y = XB + E,\ E \sim \mathcal{MN}(0, V, \Sigma)\ , \tag{1}$$

maximum likelihood estimates for the unknown parameters B and Σ are given by

$$\begin{aligned} \hat{B} &= (X^{\mathrm{T}}V^{-1}X)^{-1}X^{\mathrm{T}}V^{-1}Y \\ \hat{\Sigma} &= \frac{1}{n}(Y - X\hat{B})^{\mathrm{T}}V^{-1}(Y - X\hat{B})\ . \end{aligned} \tag{2}$$

Proof: In (1), Y is an $n \times v$ matrix of measurements (n observiations, v dependent variables), X is an $n \times p$ design matrix (n observiations, p independent variables) and V is an $n \times n$ covariance matrix across observations. This multivariate GLM implies the following likelihood function

$$\begin{aligned} p(Y|B,\Sigma) &= \mathcal{MN}(Y; XB, V, \Sigma) \\ &= \sqrt{\frac{1}{(2\pi)^{nv}|\Sigma|^n|V|^v}} \cdot \exp\left[-\frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}(Y - XB)^{\mathrm{T}}V^{-1}(Y - XB)\right)\right] \end{aligned} \tag{3}$$

and the log-likelihood function

$$\begin{aligned} \mathrm{LL}(B,\Sigma) &= \log p(Y|B,\Sigma) \\ &= -\frac{nv}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{v}{2}\log|V| \\ &\quad - \frac{1}{2}\operatorname{tr}\left[\Sigma^{-1}(Y - XB)^{\mathrm{T}}V^{-1}(Y - XB)\right]\ . \end{aligned} \tag{4}$$

Substituting V^{-1} by the precision matrix P to ease notation, we have:

$$\begin{aligned} \mathrm{LL}(B,\Sigma) = &-\frac{nv}{2}\log(2\pi) - \frac{n}{2}\log(|\Sigma|) - \frac{v}{2}\log(|V|) \\ &- \frac{1}{2}\operatorname{tr}\left[\Sigma^{-1}\left(Y^{\mathrm{T}}PY - Y^{\mathrm{T}}PXB - B^{\mathrm{T}}X^{\mathrm{T}}PY + B^{\mathrm{T}}X^{\mathrm{T}}PXB\right)\right]\ . \end{aligned} \tag{5}$$

The derivative of the log-likelihood function (5) with respect to B is

$$\begin{aligned} \frac{\mathrm{dLL}(B,\Sigma)}{\mathrm{d}B} &= \frac{\mathrm{d}}{\mathrm{d}B}\left(-\frac{1}{2}\operatorname{tr}\left[\Sigma^{-1}\left(Y^{\mathrm{T}}PY - Y^{\mathrm{T}}PXB - B^{\mathrm{T}}X^{\mathrm{T}}PY + B^{\mathrm{T}}X^{\mathrm{T}}PXB\right)\right]\right) \\ &= \frac{\mathrm{d}}{\mathrm{d}B}\left(-\frac{1}{2}\operatorname{tr}\left[-2\Sigma^{-1}Y^{\mathrm{T}}PXB\right]\right) + \frac{\mathrm{d}}{\mathrm{d}B}\left(-\frac{1}{2}\operatorname{tr}\left[\Sigma^{-1}B^{\mathrm{T}}X^{\mathrm{T}}PXB\right]\right) \\ &= -\frac{1}{2}\left(-2X^{\mathrm{T}}PY\Sigma^{-1}\right) - \frac{1}{2}\left(X^{\mathrm{T}}PXB\Sigma^{-1} + (X^{\mathrm{T}}PX)^{\mathrm{T}}B(\Sigma^{-1})^{\mathrm{T}}\right) \\ &= X^{\mathrm{T}}PY\Sigma^{-1} - X^{\mathrm{T}}PXB\Sigma^{-1} \end{aligned} \tag{6}$$

and setting this derivative to zero gives the MLE for B:

$$
\begin{aligned}
\frac{\mathrm{dLL}(\hat{B},\Sigma)}{\mathrm{d}B} &= 0 \\
0 &= X^{\mathrm{T}}PY\Sigma^{-1} - X^{\mathrm{T}}PX\hat{B}\Sigma^{-1} \\
0 &= X^{\mathrm{T}}PY - X^{\mathrm{T}}PX\hat{B} \\
X^{\mathrm{T}}PX\hat{B} &= X^{\mathrm{T}}PY \\
\hat{B} &= \left(X^{\mathrm{T}}PX\right)^{-1}X^{\mathrm{T}}PY
\end{aligned}
\tag{7}
$$

□

The derivative of the log-likelihood function (4) with respect to B is

$$
\begin{aligned}
\frac{\mathrm{dLL}(B,\Sigma)}{\mathrm{d}\Sigma} &= \frac{\mathrm{d}}{\mathrm{d}\Sigma}\left(-\frac{n}{2}\log|\Sigma| - \frac{1}{2}\,\mathrm{tr}\left[\Sigma^{-1}(Y-XB)^{\mathrm{T}}V^{-1}(Y-XB)\right]\right) \\
&= -\frac{n}{2}\left(\Sigma^{-1}\right)^{\mathrm{T}} + \frac{1}{2}\left(\Sigma^{-1}(Y-XB)^{\mathrm{T}}V^{-1}(Y-XB)\,\Sigma^{-1}\right)^{\mathrm{T}} \\
&= -\frac{n}{2}\,\Sigma^{-1} + \frac{1}{2}\,\Sigma^{-1}(Y-XB)^{\mathrm{T}}V^{-1}(Y-XB)\,\Sigma^{-1}
\end{aligned}
\tag{8}
$$

and setting this derivative to zero gives the MLE for Σ:

$$
\begin{aligned}
\frac{\mathrm{dLL}(B,\hat{\Sigma})}{\mathrm{d}\Sigma} &= 0 \\
0 &= -\frac{n}{2}\,\hat{\Sigma}^{-1} + \frac{1}{2}\,\hat{\Sigma}^{-1}(Y-XB)^{\mathrm{T}}V^{-1}(Y-XB)\,\hat{\Sigma}^{-1} \\
\frac{n}{2}\,\hat{\Sigma}^{-1} &= \frac{1}{2}\,\hat{\Sigma}^{-1}(Y-XB)^{\mathrm{T}}V^{-1}(Y-XB)\,\hat{\Sigma}^{-1} \\
\hat{\Sigma}^{-1} &= \frac{1}{n}\,\hat{\Sigma}^{-1}(Y-XB)^{\mathrm{T}}V^{-1}(Y-XB)\,\hat{\Sigma}^{-1} \\
I_v &= \frac{1}{n}\,(Y-XB)^{\mathrm{T}}V^{-1}(Y-XB)\,\hat{\Sigma}^{-1} \\
\hat{\Sigma} &= \frac{1}{n}\,(Y-XB)^{\mathrm{T}}V^{-1}(Y-XB)
\end{aligned}
\tag{9}
$$

□

Together, (7) and (9) constitute the MLE for the GLM. ■

Dependencies:
- probability density function of the matrix-normal distribution
- maximum likelihood estimation for a generative model

Source:
- to be added

Metadata: ID: A007 | name: glm-mle | author: JoramSoch | date: 2019-05-07.

Chapter IV

Model Selection

1 R-squared: Derivation of R^2 and adjusted R^2

Index:

Theorem: Given a linear regression model

$$y = X\beta + \varepsilon, \; \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \tag{1}$$

with n independent observations and p independent variables,
1) the coefficient of determination is given by

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} \tag{2}$$

2) the adjusted coefficient of determination is

$$R^2_{\text{adj}} = 1 - \frac{\text{RSS}/(n-p)}{\text{TSS}/(n-1)} \tag{3}$$

where the residual and total sum of squares are

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \quad \hat{y} = X\hat{\beta} \\ \text{TSS} &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \, , \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \end{aligned} \tag{4}$$

where X is the $n \times p$ design matrix and $\hat{\beta}$ are the ordinary least squares estimates.

Proof: The coefficient of determination R^2 is defined as the proportion of the variance explained by the independent variables, relative to the total variance in the data.

1) If we define the explained sum of squares as

$$\text{ESS} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \, , \tag{5}$$

then R^2 is given by

$$R^2 = \frac{\text{ESS}}{\text{TSS}} \, . \tag{6}$$

which is equal to

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} \, , \tag{7}$$

because $\text{TSS} = \text{ESS} + \text{RSS}$. □

2) Using (4), the coefficient of determination can be also written as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} \; . \tag{8}$$

If we replace the variance estimates by their unbiased estimators, we obtain

$$R^2_{\text{adj}} = 1 - \frac{\frac{1}{n-p}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\text{RSS}/\text{df}_r}{\text{TSS}/\text{df}_t} \tag{9}$$

where $\text{df}_r = n - p$ and $\text{df}_t = n - 1$ are the residual and total degrees of freedom. □
This gives the adjusted R^2 which adjusts R^2 for the number of explanatory variables. ■

Source:
- Wikipedia: "Coefficient of determination"; in: *Wikipedia, the free encyclopedia*; URL: https://en.wikipedia.org/wiki/Coefficient_of_determination#Adjusted_R2.

Metadata: ID: A008 | name: glm-mle | author: JoramSoch | date: 2019-05-07.

2 Log model evidence: Partition into accuracy and complexity

Index:
▷ The Book of Statistical Proofs
  ▷ Model Selection
    ▷ Log model evidence: Partition into accuracy and complexity

Theorem: The log model evidence can be partitioned into accuracy and complexity

$$\text{LME}(m) = \text{Acc}(m) - \text{Com}(m) \tag{1}$$

where the accuracy term is the posterior expectation of the log-likelihood function

$$\text{Acc}(m) = \left\langle p(y|\theta, m) \right\rangle_{p(\theta|y,m)} \tag{2}$$

and the complexity penalty is the Kullback-Leibler divergence of posterior from prior

$$\text{Com}(m) = \text{KL}\left[p(\theta|y, m) \,||\, p(\theta|m)\right] \; . \tag{3}$$

Proof: We consider Bayesian inference on data y using model m with parameters θ. Then, Bayes' theorem makes a statement about the posterior distribution, i.e. the probability of parameters, given the data and the model:

$$p(\theta|y, m) = \frac{p(y|\theta, m)\, p(\theta|m)}{p(y|m)} \; . \tag{4}$$

Rearranging this for the model evidence, we have:

$$p(y|m) = \frac{p(y|\theta,m)\,p(\theta|m)}{p(\theta|y,m)} \;. \tag{5}$$

Logarthmizing both sides of the equation, we obtain:

$$\log p(y|m) = \log p(y|\theta,m) - \log \frac{p(\theta|y,m)}{p(\theta|m)} \;. \tag{6}$$

Now taking the expectation over the posterior distribution yields:

$$\log p(y|m) = \int p(\theta|y,m) \log p(y|\theta,m)\,\mathrm{d}\theta - \int p(\theta|y,m) \log \frac{p(\theta|y,m)}{p(\theta|m)}\,\mathrm{d}\theta \;. \tag{7}$$

By definition, the left-hand side is the log model evidence and the terms on the right-hand side correspond to the posterior expectation of the log-likelihood function and the Kullback-Leibler divergence of posterior from prior

$$\mathrm{LME}(m) = \left\langle p(y|\theta,m) \right\rangle_{p(\theta|y,m)} - \mathrm{KL}\left[p(\theta|y,m) \,||\, p(\theta|m)\right] \tag{8}$$

which proofs the partition given by (1). ■

Dependencies:

- Bayes' theorem
- derivation of the log model evidence
- expectation with respect to a random variable
- Kullback-Leibler divergence of two random variables

Source:

- Penny et al. (2007): "Bayesian Comparison of Spatially Regularised General Linear Models". *Human Brain Mapping*, vol. 28, pp. 275–293.
- Soch et al. (2016): "How to avoid mismodelling in GLM-based fMRI data analysis: cross-validated Bayesian model selection". *NeuroImage*, vol. 141, pp. 469–489.

Index: number: A003; shortcut: lme-anc; author: JoramSoch; date: 2019/05/02.