UNIVERSITÀ
DEGLI STUDI DI TRIESTE

Dipartimento di scienze economiche,
aziendali, matematiche e statistiche
“Bruno de Finetti”

Bayesian Statistics

Linear regression

Leonardo Egidi

A.A. 2018/19

Indice

Foundations

The problem we will consider in this section is that of using the values of one variable to explain or predict values of another. We shall refer to an *explanatory* (covariate) and a *dependent* (outcome, response) variable.

A common question is: how does one quantity, y, vary as a function of another quantity (or vector of quantities), x?

In general, we are interested in the conditional distribution of y, given x, parametrized as $p(y|\theta, x)$, where n observations (x_i, y_i) are available.

The term *regression* dates back to Francis Galton, namely *regression toward the mean*.

Foundations

The distribution of y given x is typically studied in the context of a set of units or experimental subjects. Regression models the statistical relationship between the variates and the covariates. There are two slightly different situations:

- in the first one the experimenters are free to set the values of x_i: in such a case, x is *deterministic*. We will typically assume that the covariates are independent of the model parameters: $p(x|\theta) = p(x)$.
- in the second one both values are random, although one is thought of as having a causal or explanatory relationship with the other.

The analysis, however, turns out to be the same in both cases.

Foundations

The two approaches differ in the evaluation of the joint distribution of y and x:

- in the first approach the joint distribution of x and y is proportional to the conditional distribution of y given x:

$$p(y, x|\theta) = p(y|x, \theta)p(x)$$

$$p(y, x|\theta) \propto p(y|x, \theta)$$

Ex: studying the height(y) vs the weight (x) of n individuals

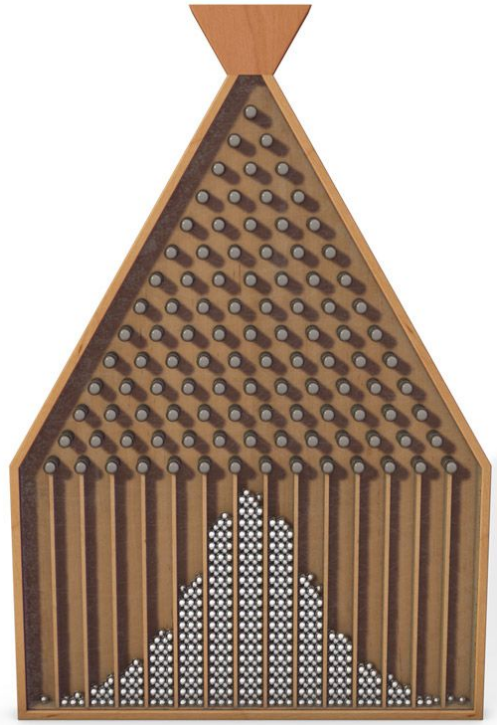
- in the second approach there is a model for the x as well:

$$p(y, x|\theta) = p(y|x, \theta)p(x|\theta)$$

Ex: studying the effect of a noisy protein measurements x on some outcome response.

Foundations

Galton developed the following model: pellets fall through a quincunx to form a normal distribution centered directly under their entrance point. These pellets might then be released down into a second gallery corresponding to a second measurement. Galton then asked the reverse question: "From where did these pellets come?"



The answer was not 'on average directly above'. Rather it was 'on average, more towards the middle', for the simple reason that there were more pellets above it towards the middle that could wander left than there were in the left extreme that could wander to the right, inwards.

Foundations

Goals of a regression analysis

1. understanding the behavior of y, given x
2. predicting y, given x, for future observations
3. causal inference, or predicting how y would change if x were changed in a specified way.

Normal linear model

Describing the pellets fall provoked one of the most powerful scientific revolution, in statistical terms: the normal linear model, in which the distribution of each response measurement $y_i, i = 1, \ldots, n$ given the $n \times p$ matrix of predictors X is normal with a mean that is a linear function of X:

$$E(y_i|\beta, X) = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}, \qquad (1)$$

and the variance for each y_i is σ^2. The parameter vector is then $\theta = (\beta_1, \ldots, \beta_p, \sigma)$.

The conditional mean above may be expressed in matrix terms as $X\beta$, where $\beta = (\beta_1, \ldots, \beta_p)$ is a p-dimensional vector. Then, the likelihood of the model is:

$$p(y|X, \theta) = \mathcal{N}(X\beta, \sigma^2)$$

Normal linear model

Compared to classical inference, Bayesian inference poses a new copernican challenge: the vector parameter θ is not a fixed quantity to be estimated from the data, but is assigned a **prior** distribution reflecting the experimenter's uncertainty and/or his/her substantial knowledge.

Key statistical modelling issue:

- defining y and X in such a way the conditional expectation (1) of y is reasonably linear (the same as in classical inference)
- setting up a prior distribution of the model parameters that accurately reflects substantial knowledge.

Unless otherwise stated, in what follows we assume that X is a $n \times p$ matrix, with the first column of 1's, and $\beta = (\beta_1, \ldots, \beta_p)$, a p-dimensional vector, where β_1 is the intercept.

Normal linear model

Statistical inference

Estimate the parameter θ in terms of a posterior distribution, conditional on the response y and the covariates matrix X.

Under the assumptions that the design matrix X is completely *known*, the posterior distribution for θ is:

$$\pi(\theta|X, y) \propto p(y|\theta, X)\pi(\theta), \tag{2}$$

where $p(y|\theta, X)$ is the likelihood model function and $\pi(\theta)$ the prior distribution for the parameter vector θ.

A Bayesian model is constituted by the *pair prior-likelihood*.

Indice

Normal linear model: priors

Classical linear regression may be thought of as bayesian posterior inference based on a *noninformative* prior distribution for the parameters of the normal linear model. Given $y_i \sim \mathcal{N}(\sum_{j=1}^{p} x_{ij}\beta_j, \sigma^2)$, each β_j may be thought of as drawn from a Uniform prior distribution Unif$(-\infty, +\infty)$.

We need a prior distribution that is sufficiently strong for the model parameters to be accurately estimated from the data at hand, yet not so strong as to dominate the data inappropriately. The prior may then act as a regularization device.

We will briefly outline, from a Bayesian perspective, the choices involved in setting up a regression model, starting from noninformative priors and then moving to more informative priors.

Normal linear model: noninformative prior distribution

A convenient noninformative prior distribution is uniform on $(\beta, \log\sigma)$ or, equivalently,

$$\pi(\beta, \sigma^2|X) \propto \sigma^{-2}. \tag{3}$$

Proof

$p(\beta, \log\sigma) = 1_{(-\infty,+\infty)}$. *If* $\sigma^2 = U$ *and* $\log\sqrt{U} = W$, *then* $\left|\frac{dw}{du}\right| = \frac{1}{\sqrt{u}}\frac{1}{2\sqrt{u}} = \frac{1}{2u} = \frac{1}{2\sigma^2}$. *Thus,*
$p(\beta, \sigma^2) = \left|\frac{d\log\sigma}{d\sigma^2}\right| 1_{(-\infty,+\infty)} = \frac{1}{2\sigma^2} \propto \sigma^{-2}.\square$

When there are many data points and only a few parameters, this prior is useful and takes less effort than specifying prior knowledge in probabilistic form.

However, for a small sample size or a large number of parameter, the likelihood is less sharply peaked, and (more informative) prior distributions become more important.

Indice

Normal linear model: noninformative analysis

As with the normal distribution with unknown mean and variance, we define the joint posterior distribution for σ^2 and β via the Bayes Theorem. For simplicity, we suppress the dependence on X here:

$$\begin{aligned}\pi(\beta, \sigma^2|y) &\propto p(y|\beta, \sigma^2)\pi(\beta, \sigma^2) \propto \mathcal{N}(X\beta, \sigma^2)\sigma^{-2} \\ &\propto \sigma^{-n-2} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2\right]\right\} \end{aligned} \tag{4}$$

Then, we factor the joint posterior $p(\beta, \sigma^2|y)$ as:

$$\pi(\beta, \sigma^2|y) = \pi(\beta|\sigma^2, y)\pi(\sigma^2|y), \tag{5}$$

where $\pi(\beta|\sigma^2, y)$ is the conditional posterior of β on σ^2, and $\pi(\sigma^2|y)$ is the marginal posterior of σ^2.

Normal linear model: noninformative prior

We rewrite the joint posterior (4) as:

$$
\begin{aligned}
\pi(\beta, \sigma^2|y) &\propto p(y|\beta, \sigma^2)\pi(\beta, \sigma^2) \\
&\propto \sigma^{-n-2} \exp\left\{-\frac{1}{2\sigma^2}\left[(y - X\beta)^T(y - X\beta)\right]\right\} \\
&\propto \sigma^{-n-2} \exp\left\{-\frac{1}{2\sigma^2}\left[SSE + (\beta - \hat{\beta})^T X^T X(\beta - \hat{\beta})\right]\right\} \\
&\propto \sigma^{-n-2} \exp\left\{-\frac{1}{2\sigma^2} SSE\right\} \exp\left\{-\frac{1}{2\sigma^2}\left[(\beta - \hat{\beta})^T X^T X(\beta - \hat{\beta})\right]\right\}, \\
&\propto \underbrace{\sigma^{-\frac{p}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[(\beta - \hat{\beta})^T X^T X(\beta - \hat{\beta})\right]\right\}}_{\beta|\sigma^2, y \sim \mathcal{N}(\hat{\beta}, V_\beta \sigma^2)} \underbrace{\sigma^{-\frac{(n+p)}{2}} \exp\left\{-\frac{1}{2\sigma^2} SSE\right\}}_{\sigma^2|y \sim \text{inv-}\chi^2(n-p, s^2)},
\end{aligned}
$$

where $SSE = (y - X\hat{\beta})^T(y - X\hat{\beta})$ and $\hat{\beta} = (X^T X)^{-1} X^T y$.

Normal linear model: noninformative prior

- To derive the conditional posterior for β, we must extract from the joint posterior only the pieces related to β:

$$\pi(\beta|\sigma^2, y) \propto \sigma^{-p/2} \exp\left\{-\frac{1}{2\sigma^2}\left[(\beta-\hat{\beta})^T X^T X(\beta-\hat{\beta})\right]\right\},$$

thus we may conclude that

$$\beta|\sigma^2, y \sim \mathcal{N}(\hat{\beta}, V_\beta \sigma^2), \tag{6}$$

with $V_\beta = (X^T X)^{-1}$.

- Denoting with $s^2 = \frac{1}{n-p}(y - X\hat{\beta})^T(y - X\hat{\beta})$, after some algebra one may conclude that:

$$\sigma^2|y \sim \text{inv-}\chi^2(n-p, s^2). \tag{7}$$

Normal linear model: noninformative prior

Noninformative posterior

We may factor the joint posterior $\pi(\beta, \sigma^2|y)$ as the product of $p(\beta|\sigma^2, y)$ and $p(\sigma^2|y)$. The joint posterior follows a Normal-Inverse Gamma [a] distribution, $\mathcal{N}\text{IG}(\mu, \lambda, a, b)$, where

$$\mu = \hat{\beta}, \lambda = (X^T X)^{-1}, a = (n-p)/2, b = (n-p)s^2/2$$

.

[a] We use the fact that an invGamma$(n/2, 1/2)$ is an inv$-\chi^2(n, 1)$

- Comparison to classical inference: the classical estimates for β and σ are $\hat{\beta}$ and s.
- Simulation: the factorization (5) is useful in terms of simulation from the model. One can simulate σ^2 from (7) and then β from (6).

Normal linear model: is the posterior distribution proper?

As for any analysis based on an improper prior distribution, we need to check if the posterior distribution is proper. Given a generic parameter vector θ, a prior distribution is said to be *proper* if:

$$\int_{\Theta} \pi(\theta) d\theta < +\infty$$

In our case, it is immediate to check that $\pi(\beta, \sigma^2)$ is not proper:

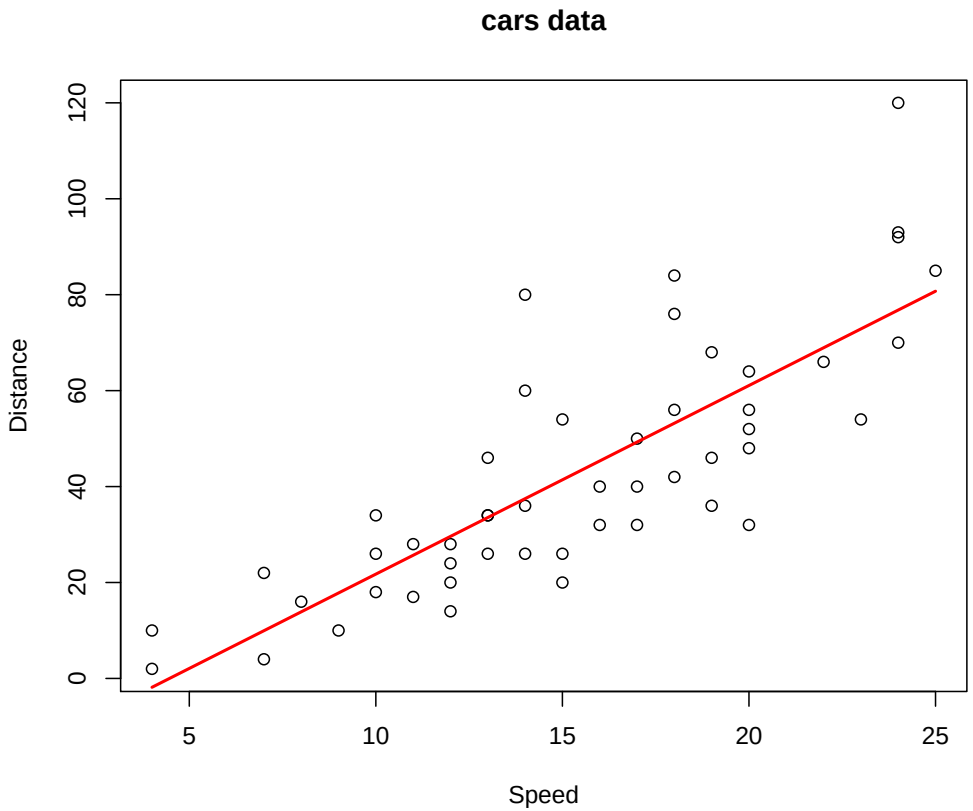
$$\int_{\Theta} \pi(\beta, \sigma^2) d\theta = \int_{\Theta} \sigma^{-2} d\theta = +\infty$$

However, proper posteriors may come from improper priors. In this case: $\int_{\Theta} \pi(\beta, \sigma^2 | y) d\theta < \infty$ iff:

- $n > p$: there are at least as many data point as parameters;
- $\text{rank}(X) = p$: no collinearity in X.

CARS dataset

We use now the `cars` dataset, consisting of the speed of cars (mph) and the distances taken to stop (ft). We build up a linear model for the distance (Y) explained by the speed (x), with noninformative priors for $\theta = (\alpha, \gamma, \log \sigma)$.



$$y_i \sim \mathcal{N}(\alpha + \gamma x, \sigma^2)$$

$$p(\alpha, \gamma, \log \sigma) = \text{Unif}(-\infty, +\infty)$$

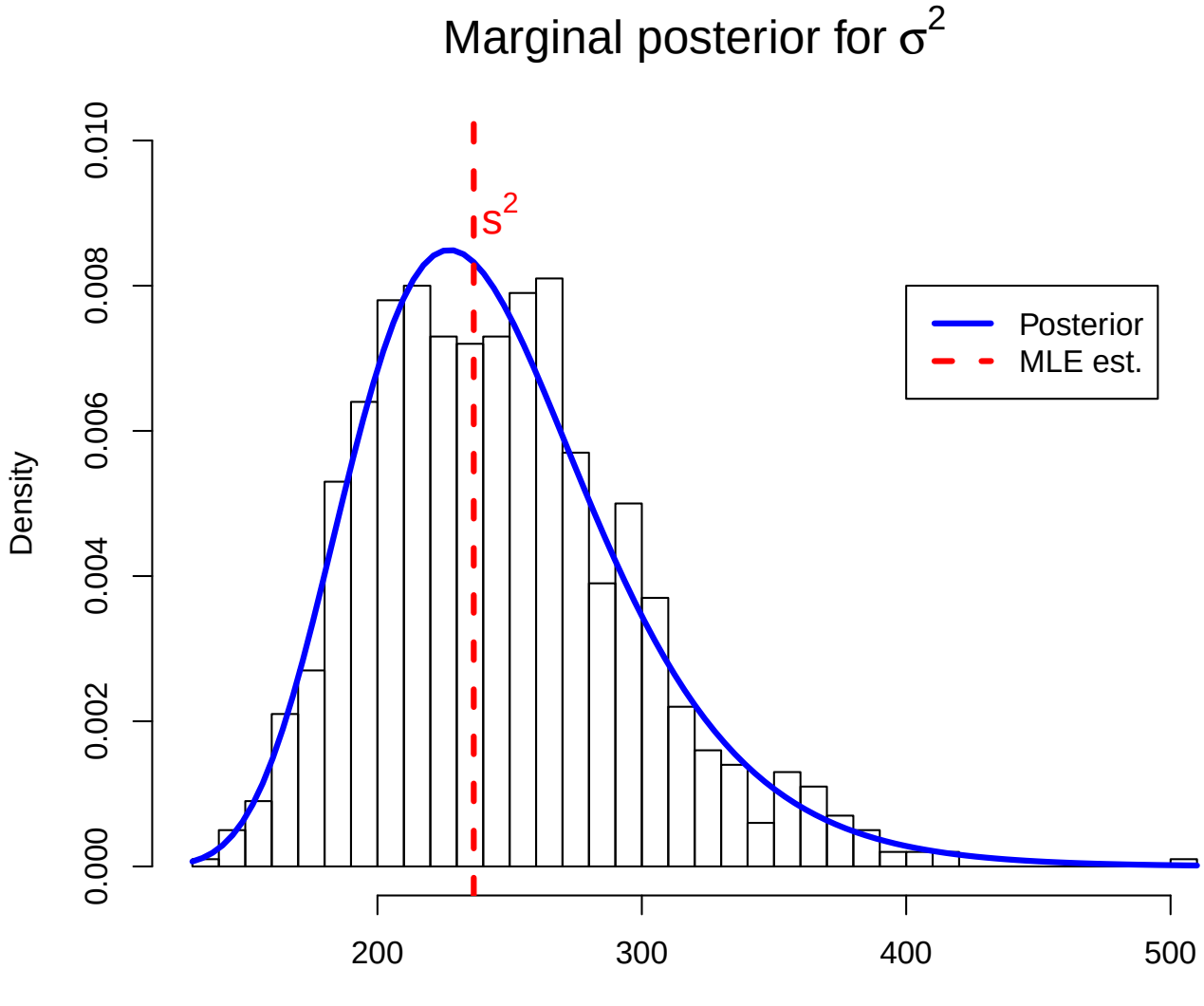
Given $\beta = (\alpha, \gamma)$, and X a $n \times 2$ design matrix (where the first column is a $n \times 1$ vector of 1's), from classical inference we have:

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

$$\hat{\sigma^2} = s^2 = \frac{1}{n-p}(y - X\hat{\beta})^T (y - X\hat{\beta})$$

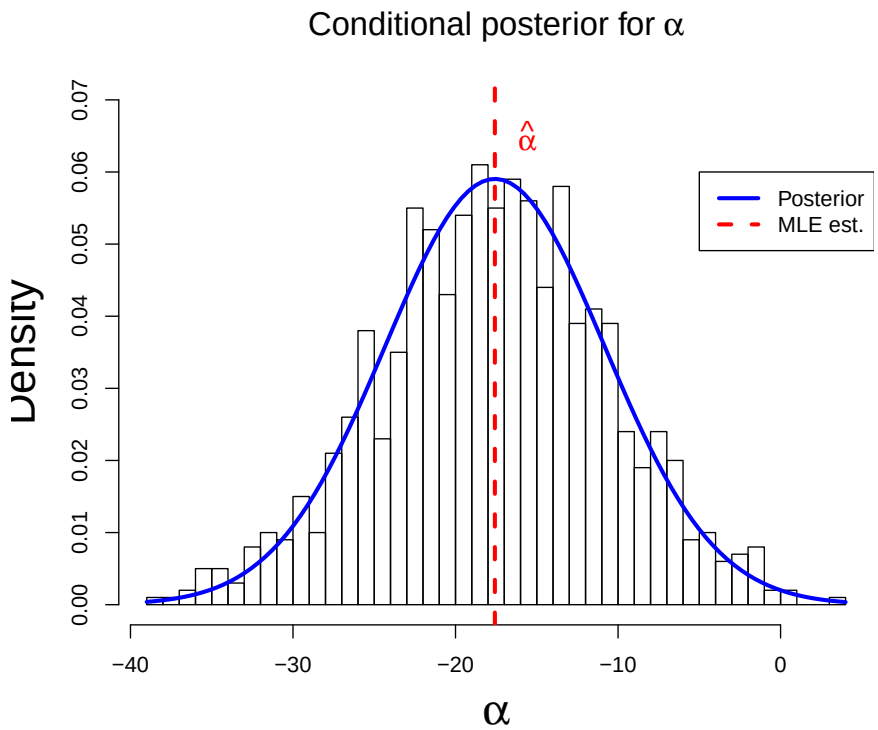
Normal linear model: posterior simulation

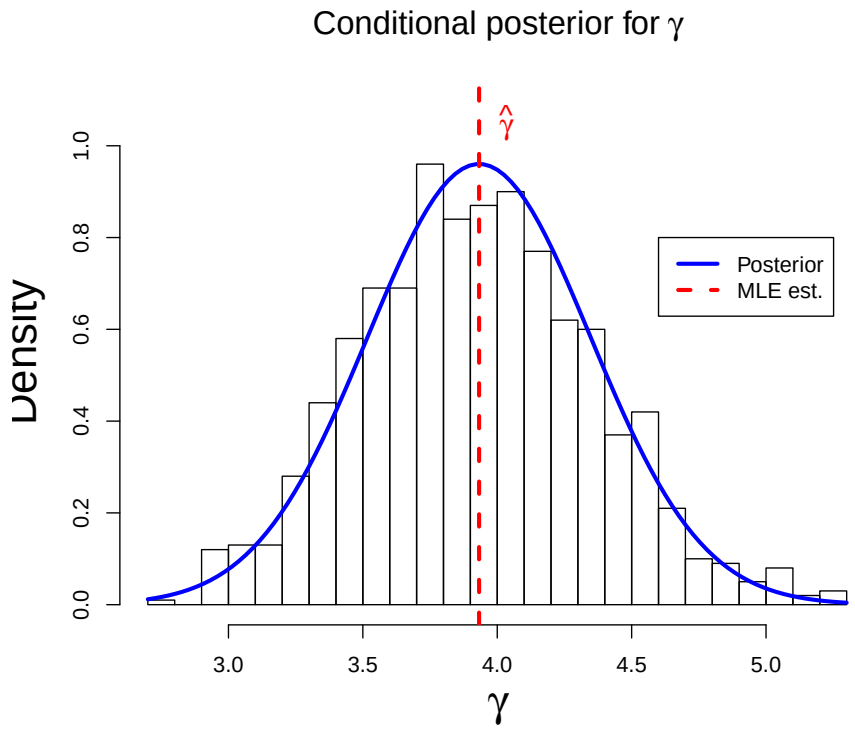
We simulate σ^2 from $\pi(\sigma^2|y) = \text{invGamma}((n-1)/2, (n-1)s^2/2)$:



Normal linear model: posterior simulation

We simulate α and γ from $\mathcal{N}(\hat{\beta}, (X^T X)^{-1}\sigma^2)$:





Normal linear model: comparison

Clearly, the estimates under the two approaches are quite similar:

Par.	MLE	Posterior mean	95 % Conf. Int.	95 % Credible Int.
σ^2	236.4	246.7	–	(165.13, 364.5)
α	-17.6	-17.4	(-31.2, -4)	(-31.12, -3.33)
γ	3.9	3.92	(3, 4.77)	(3.06, 4.77)

Indice

Normal linear model: predictions

Suppose now that we have observed new data, a new matrix $\tilde{X}$ of explanatory variables, and we wish to predict the outcomes $\tilde{y}$. Our current knowledge of β and σ is summarized by our posterior distribution.

We need the posterior predictive distribution of future data $\tilde{y}$, given the current data y, $p(\tilde{y}|y)$.

This distribution has two components of uncertainty:

1. the fundamental variability of the model, represented by the variance σ^2 in y
2. the posterior uncertainty in β, σ

Normal linear model: predictions

We retrieve the formulation of the posterior predictive distribution in the normal linear model (here, we still suppress the dependence on X and $\tilde{X}$)

$$\begin{aligned} p(\tilde{y}|y) &= \int p(\tilde{y}, \beta, \sigma^2|y) d\beta d\sigma^2 \\ &= \int p(\tilde{y}|\beta, \sigma^2) \pi(\beta, \sigma^2|y) d\beta d\sigma^2 \\ &= \int p(\tilde{y}|\theta) \pi(\theta|y) d\theta, \end{aligned} \tag{8}$$

where $\theta = (\beta, \sigma^2)$, and the *conditional independence* between future observations $\tilde{y}$ and current observations y given θ is assumed, meaning that $p(\tilde{y}|\theta, y) = p(\tilde{y}|\theta)$.

Normal linear model: predictions

For the majority of the models, the posterior predictive distribution is usually not available in analytical form. For such a reason, we may obtain it through a two-steps simulation:

- generate β, σ^2 from the posterior $\pi(\beta, \sigma^2|y)$;
- generate $\tilde{y}$ from the likelihood for future values $p(\tilde{y}|\beta, \sigma^2)$.

However, in the normal linear model there is an analytical form of the posterior predictive distribution.

Normal linear model: predictions

We start by definition of the posterior predictive distribution, using the factorization (5):

$$p(\tilde{y}|y) = \int p(\tilde{y}|\beta,\sigma^2)\pi(\beta|\sigma^2,y)\pi(\sigma^2|y)d\beta d\sigma^2$$
$$= \int \mathcal{N}(\tilde{X}\beta,\sigma^2)\mathcal{N}(\hat{\beta},(X^TX)^{-1}\sigma^2)\pi(\sigma^2|y)d\beta d\sigma^2.$$

The product $\mathcal{N}(\tilde{X}\beta,\sigma^2)\mathcal{N}(\hat{\beta},(X^TX)^{-1}\sigma^2)$ is a quadratic form in $(\tilde{y},\beta)$, thus is a Normal distribution.

At the end, the posterior predictive distribution is a multivariate student-t:

$$\tilde{y}|y \sim t_{n-p}(\tilde{X}\hat{\beta}, s^2(I+\tilde{X}(X^TX)^{-1}\tilde{X}^T)),$$

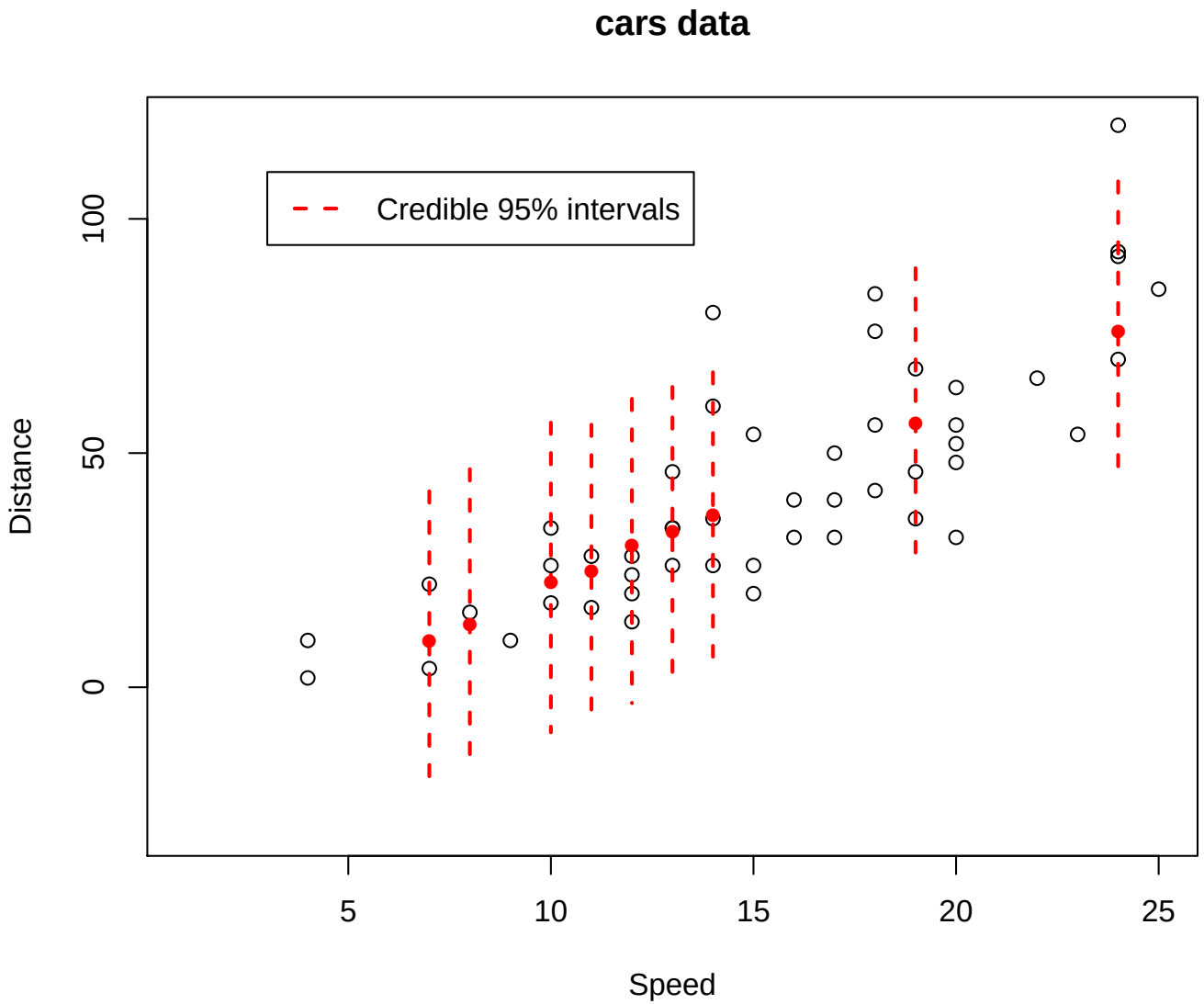
where I is the identity matrix.

Normal linear model: predictions

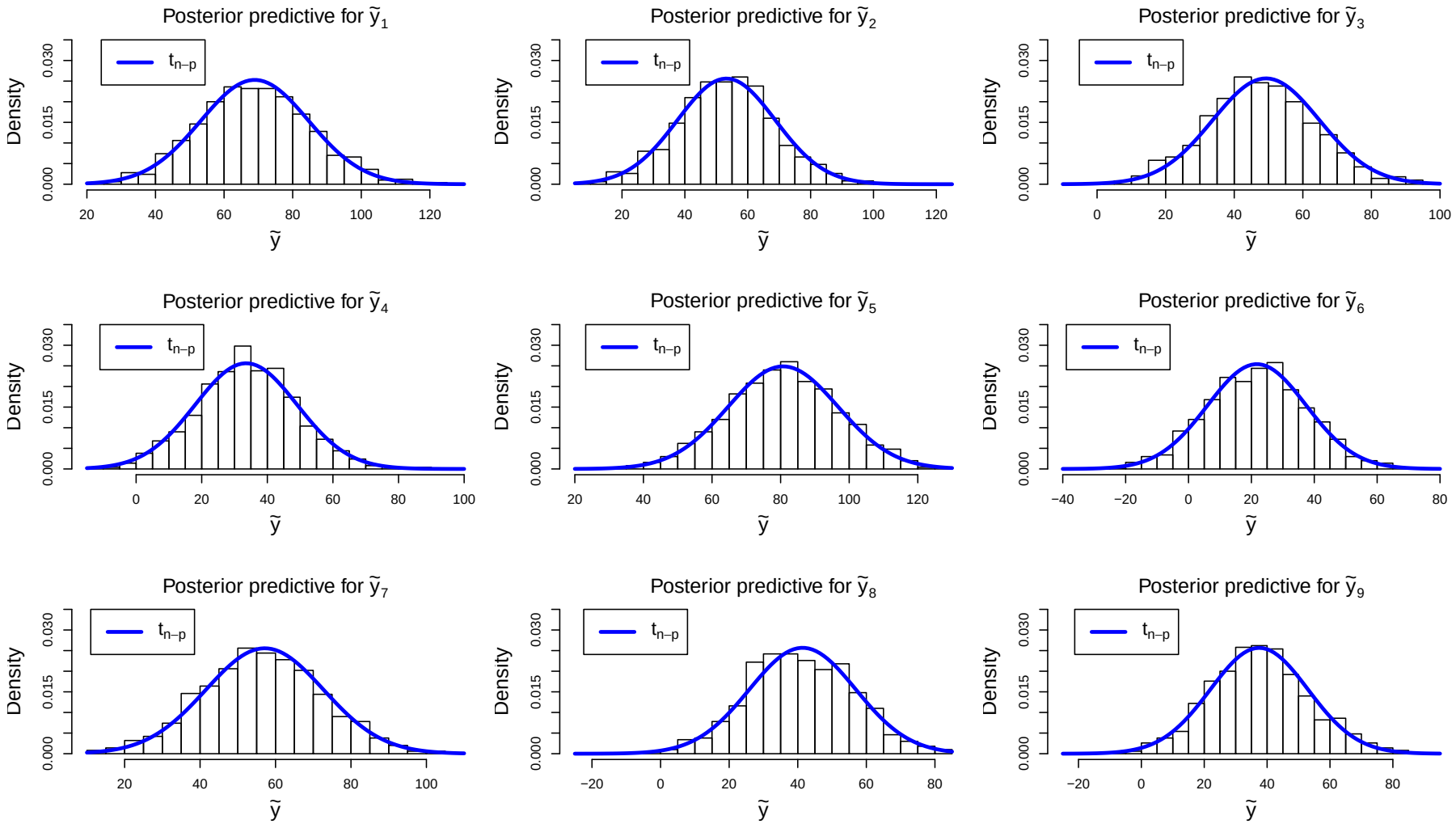
The analytical form of $p(\tilde{y}|y)$ is usually unfeasible for simulation purposes. Thus, we need the two-step simulation involving the posterior $\pi(\beta|\sigma^2, y)$ and the likelihood for future responses $p(\tilde{y}|\beta, \sigma^2)$.

Suppose we have 10 new speed's measurements in the previous example, say $\tilde{x} = (13, 24, 19, 11, 8, 14, 10, 12, 7)$ and we wish to make some predictions $\tilde{y}$ for these new explanatory variables.

Normal linear model: predictions



Normal linear model: predictions



Indice

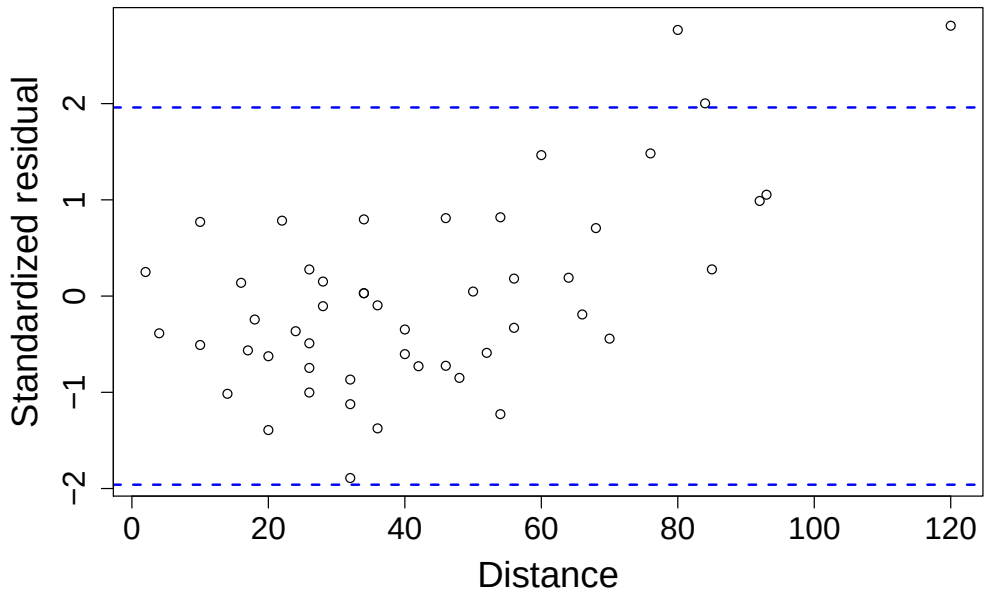
Normal linear model: model checking and residuals

To check the fit of a linear model we may use the standardized residuals:

$$(y_i - X_i\hat{\beta})/s, \tag{9}$$

where s is the estimated standard deviation.

If the normal linear model is correct, the standardized residuals should be approximately distributed as a standard normal distribution.



Normal linear model: model checking and residuals

An advantage of the Bayesian approach is that we can compute, using *simulation*, the posterior predictive distribution for any data summary, so we do not need to put a lot of effort into estimating the sampling distribution of a test statistic.

For example, to assess the statistical and practical significance of patterns in a residual plot, we can obtain the ppd of an appropriate test statistic.

Normal linear model: model checking and residuals

- Draw β, σ^2 from the joint posterior distribution.
- Draw some hypothetical replications, y^{rep}, from the predictive distribution, $y^{rep} \sim \mathcal{N}(X\beta, \sigma^2 I)$.
- Run a regression of y^{rep} on X and save the residuals.

We can then construct the posterior predictive distribution for the proportion of outliers in Bayesian linear regression:

2.5%	Median	97.5%
0	0.002	0.006

Indice

Normal linear model: informative prior

Bayesian analysis is designed to incorporate prior information into statistical inference. Then, we move from formula (3) to:

$$\begin{aligned}\pi(\beta, \sigma^2) =& \pi(\beta|\sigma^2)\pi(\sigma^2) \\ &= \mathcal{N}(\mu_\beta, \sigma^2 W_\beta)\text{invGamma}(a, b) = \mathcal{N}\text{IG}(\mu_\beta, W_\beta, a, b),\end{aligned} \tag{10}$$

where $\mathcal{N}$IG denotes an inverse gamma distribution, $a, b > 0$. The Normal-Inverse Gamma is the *conjugate prior* for the linear model. The posterior is:

$$\pi(\beta, \sigma^2|y) \propto p(y|\beta, \sigma^2)\pi(\beta, \sigma^2) \tag{11}$$

Normal linear model: informative prior

One may prove that the joint posterior is still a $\mathcal{N}\text{IG}(\mu^*, W^*, a^*, b^*)$, where:

$$\begin{aligned}
\mu^* &= (W_\beta^{-1} + X^T X)^{-1}(W_\beta^{-1}\mu_\beta + X^T y) \\
W^* &= (W_\beta^{-1} + X^T X)^{-1} \\
a^* &= a + n/2 \\
b^* &= b + \frac{1}{2}[\mu_\beta^T W_\beta^{-1}\mu_\beta + y^T y - \mu^{*T} W^{*-1}\mu^*].
\end{aligned}$$

The conditional posterior $\beta|\sigma^2, y \sim \mathcal{N}(\mu^*, W^*\sigma^2)$, whereas $\sigma^2|y \sim \text{invGamma}(a^*, b^*)$.

Normal linear model: informative prior

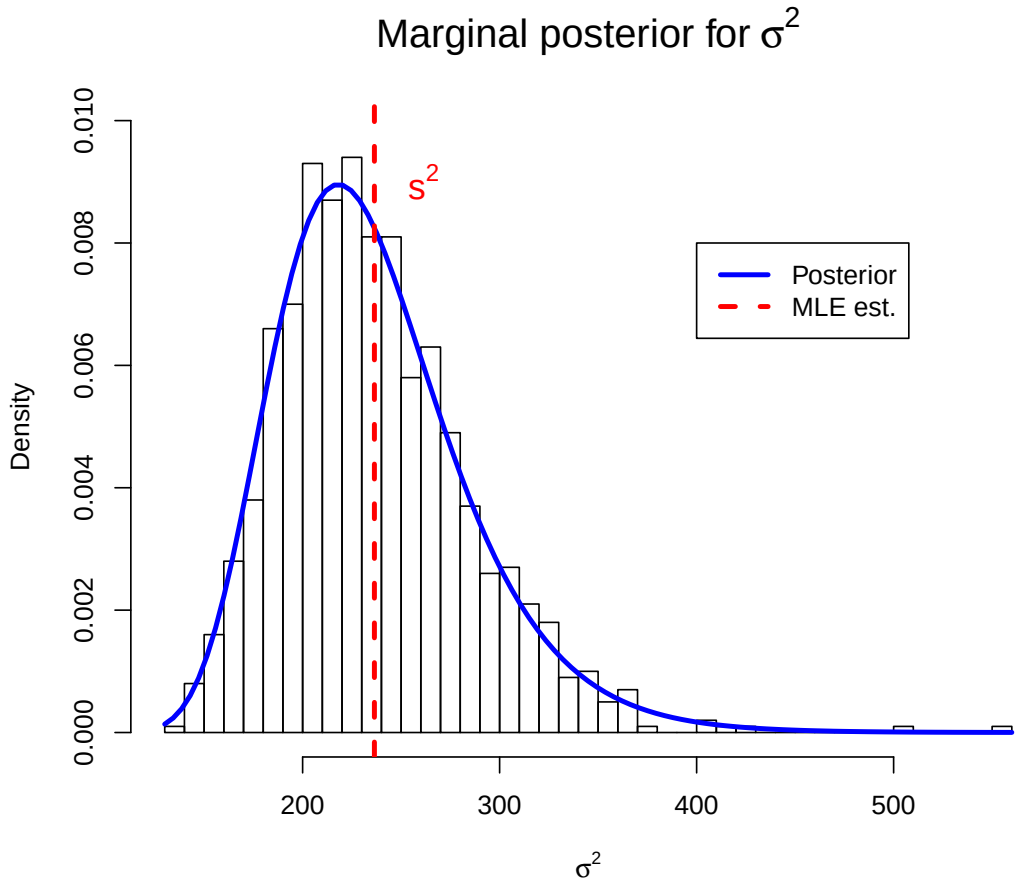


Figure: $a = b = 1$

Normal linear model: informative prior

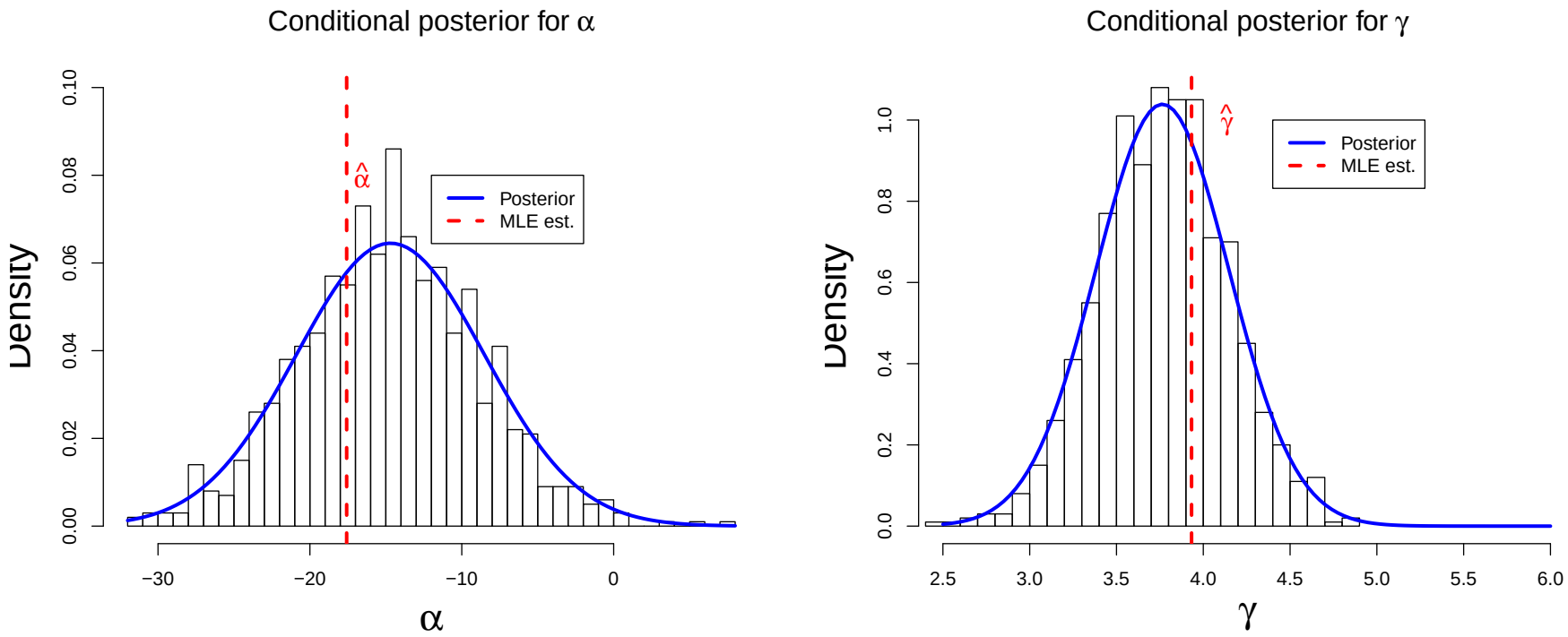


Figure: $W_\beta = diag(2)$, $a = b = 1$

Normal linear model: informative prior

The informative prior is likely to produce less heavy-tailed posterior distribution (see the support in the plot for the intercept α).

However, it is not easy to incorporate *substantial* prior information in a regression model. For such a reason, the noninformative prior may be seen as the standard choice.

Many scholars (Gelman, e.g.) suggest to use *weakly informative priors*.

Every analysis should be checked: is my model robust to the prior choice?

Indice

Limits and extension of the normal linear model

Linearity and normality are rude approximations of the real world. The analyst should always take care about the appropriate of the linear model. Precisely, normal linear models may fail for many reasons:

- non-normality
  Departures from normality assumptions may be revelaed by a QQ plot between theoretical and sample quantiles
- non-linearity
  It often makes sense to transform y and x so that their relation is approximately linear (the logarithm, for instance)
- the range of the response is not $(-\infty, +\infty)$
  If the response y is discrete, it makes sense to use generalized linear models
- heteroschedasticity
  Assuming the same variance for each subject may be restrictive in many cases

Limits and extension of the normal linear model

Bayesian linear models may be extended in many directions:

- Correlation between units.
  Considering n independent units may be difficult in many cases. Thus, we could model correlated errors $y - X\beta$ by allowing a data covariance matrix Σ_y, that is not necessarily proportional to the identity matrix:

$$y \sim \mathcal{N}(X\beta, \Sigma_y) \tag{12}$$

  The covariance matrix may be known, partially known, or unknown.

- Variable selection
  Ideally, we should include all relevant information in a statistical model, i.e. including all those explanatory variables x that might possibly help predict y. When using noninformative priors, the restriction is that $n > p$: but reasonable prior distribution may help to handle a large number of predictors.

Modelling unequal variances and correlation

Assuming the same variability for all the n independent statistical units is often too restrictive: thus, we need to incorporate correlation and heteroschedasticity.

We consider model extension (12), and we will consider three cases for Σ_y:

1. Σ_y known
2. Σ_y known up to a scalar factor
3. Σ_y unknown

In cases 1 and 2, a symmetric, positive-definite $n \times n$ covariance matrix Σ_y must be entirely specified, and, moreover, in 2 a prior for σ^2 is assigned; in case 3, Σ_y is assigned a prior distribution.

Unequal variances and correlation: (1) Σ_y known

Let $\Sigma_y^{1/2}$ be a Cholesky factor (an upper triangular 'matrix square root') of Σ_y. For simplicity, $\sigma = 1$, thus the parameter vector $\theta = (\beta_1, \ldots, \beta_p)$. Multiplying both sides of the canonical regression equation by $\Sigma_y^{-1/2}$ yields:

$$\Sigma_y^{-1/2} y | \beta, X \sim \mathcal{N}(\Sigma_y^{-1/2} X\beta, I) \tag{13}$$

The objective is always to find and sample from the posterior $\pi(\beta|y)$. One can prove that the posterior for β is now:

$$\beta|y \sim \mathcal{N}(\hat{\beta}, V_\beta), \tag{14}$$

where

$$\hat{\beta} = (X^T \Sigma_y^{-1} X)^{-1} X^T \Sigma_y^{-1} y \tag{15}$$

$$V_\beta = (X^T \Sigma_y^{-1} X)^{-1} \tag{16}$$

Unequal variances and correlation: (1) Σ_y known

Suppose we wish to predict $\tilde{n}$ new values $\tilde{y}$, given a $\tilde{n} \times p$ matrix of explanatory variables, $\tilde{X}$. We must specify the *joint variance matrix* for the old and new data. We use then the following notation for the joint normal distribution of y and $\tilde{y}$, given the predictors and the parameters:

$$y, \tilde{y} | X, \tilde{X}, \beta \sim \mathcal{N}\left(\begin{pmatrix} X\beta \\ \tilde{X}\beta \end{pmatrix}, \begin{pmatrix} \Sigma_y & \Sigma_{y,\tilde{y}} \\ \Sigma_{\tilde{y},y} & \Sigma_{\tilde{y}} \end{pmatrix} \right), \tag{17}$$

with the covariance matrix for $(y, \tilde{y})$ symmetric and positive semidefinite. The mean and the variance are then:

$$\begin{aligned} \mathrm{E}(\tilde{y} | y, \beta, \Sigma_y) &= \tilde{X}\beta + \Sigma_{\tilde{y},y}\Sigma_y^{-1}(y - X\beta) \\ \mathrm{Var}(\tilde{y} | y, \beta, \Sigma_y) &= \Sigma_{\tilde{y}} - \Sigma_{\tilde{y},y}\Sigma_y^{-1}\Sigma_{y,\tilde{y}} \end{aligned}$$

Unequal variances and correlation: (2) Σ_y partially known

Suppose that the covariance matrix is known up to a constant, we can then write Σ_y as:

$$\Sigma_y = Q_y \sigma^2, \tag{18}$$

where the matrix Q_y is known but the scale σ is unknown. Then:

$$Q_y^{-1/2} y | X, \beta, \sigma^2 \sim \mathcal{N}(Q_y^{-1/2} X\beta, \sigma^2 I)$$

Assuming as before $\pi(\beta, \sigma^2) \propto \sigma^{-2}$, we may obtain the conditional posterior for β and the marginal posterior for σ^2, respectively as:

$$\begin{aligned} \beta | y, \sigma^2 \sim & \mathcal{N}(\hat{\beta}, V_\beta) \\ \sigma^2 | y \sim & \text{invGamma}((n-p)/2, (y - X\hat{\beta})^T Q_y^{-1} (y - X\hat{\beta})) \end{aligned} \tag{19}$$

where

$$\hat{\beta} = (X^T Q_y^{-1} X)^{-1} X^T Q_y^{-1} y, \quad V_\beta = (X^T Q_y^{-1} X)^{-1}$$

Unequal variances and correlation: (3) Σ_y unknown

As already seen for multivariate Normal models, if the covariance matrix Σ_y is unknown, we need to specify a prior on Σ_y, which is not an easy task unless Σ_y has a fixed structure. In general, let $\pi(\Sigma_y)$ be such prior and assume $\pi(\beta|\Sigma_y) \propto 1$, then the conditional posterior for β is obtained as:

$$\pi(\beta|y, \Sigma_y) = \mathcal{N}(\hat{\beta}, V_\beta),$$

with $\hat{\beta}$ and V_β defined by (16). The marginal posterior for Σ_y is given by:

$$\begin{aligned} p(\Sigma_y|y) =& \frac{\pi(\beta, \Sigma_y|y)}{\pi(\beta|y, \Sigma_y)} \\ =& \pi(\Sigma_y)\frac{p(y|\beta, \Sigma_y)}{\pi(\beta|y, \Sigma_y)} \\ \propto& \pi(\Sigma_y)|\Sigma_y|^{-1/2}|V_\beta|^{1/2} \exp\left\{-\frac{1}{2}(y - X\hat{\beta})^T\Sigma_y^{-1}(y - X\hat{\beta})\right\} \end{aligned} \tag{20}$$

Unequal variances and correlation: (3) Σ_y unknown

In (20), for convenience and computational stability we set $\beta = \hat{\beta}$. The density (20) is easy to compute, but hard to draw samples from in general, because of the dependence of $\hat{\beta}$ and $|V_\beta|^{1/2}$ on Σ_y. Setting a prior distribution on Σ_y is, in general, a difficult task!

A typical choice for the prior distribution of Σ_y is the Inverse Wishart distribution, $\Sigma_y \sim \text{invWishart}(\Lambda_0, \nu_0)$, where the latter means that

$$\pi(\Sigma_y) \propto |\Sigma_y|^{-(\nu_0+p+1)/2} \exp\left\{-\frac{1}{2}\text{tr}(\Lambda_0 \Sigma_y^{-1})\right\}$$

,
with $\nu_0 > p-1$ degrees of freedom and Λ_0 a $p \times p$ positive definite scale matrix.

Further reading

Further reading for the normal linear model:

- Chapter 14 from *Bayesian Data Analysis*, Gelman et al.
- Chapter 6.3 and 6.4 from *Bayesian Statistics: an introduction*, Lee.