Bayesian Learning

Lecture 4 - Regression, Prediction and Decisions

Mattias Villani 🧔

Department of Statistics
Stockholm University

mattiasvillani.com @matvil @matvil mattiasvillani

Lecture overview

- **Normal model** with conjugate prior
- The **linear regression** model
- **Prediction**
- **Decision making**

Linear regression

- The linear regression model in **matrix form**

$$\underset{(n\times 1)}{\mathbf{y}} = \underset{(n\times k)(k\times 1)}{\mathbf{X}\beta} + \underset{(n\times 1)}{\varepsilon}$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},\ \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix},\ \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

- Usually $x_{i1} = 1$, for all i. β_1 is the intercept.
- **Likelihood**

$$\mathbf{y}|\beta, \sigma^2, \mathbf{X} \sim N(\mathbf{X}\beta, \sigma^2 I_n)$$

Linear regression - uniform prior

- Standard **non-informative prior**: uniform on $(\beta, \log \sigma^2)$

$$p(\beta, \sigma^2) \propto \sigma^{-2}$$

- **Joint posterior** of β and σ^2:

$$\begin{aligned} \beta|\sigma^2, \mathbf{y} &\sim N\left(\hat{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\right) \\ \sigma^2|\mathbf{y} &\sim Inv\text{-}\chi^2(n-k, s^2) \end{aligned}$$

where $\hat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ and $s^2 = \frac{1}{n-k}(\mathbf{y} - \mathbf{X}\hat{\beta})^\top(\mathbf{y} - \mathbf{X}\hat{\beta})$.

- **Simulate** from the joint posterior by simulating from
  - $p(\sigma^2|\mathbf{y})$
  - $p(\beta|\sigma^2, \mathbf{y})$
- **Marginal posterior** of β :

$$\beta|\mathbf{y} \sim t_{n-k}\left(\hat{\beta}, s^2(X^\top X)^{-1}\right)$$

Linear regression - conjugate prior

- **Joint prior** for β and σ^2

$$\beta|\sigma^2 \sim N\left(\mu_0, \sigma^2\Omega_0^{-1}\right)$$
$$\sigma^2 \sim Inv-\chi^2\left(\nu_0, \sigma_0^2\right)$$

- **Posterior**

$$\beta|\sigma^2, \mathbf{y} \sim N\left(\mu_n, \sigma^2\Omega_n^{-1}\right)$$
$$\sigma^2|\mathbf{y} \sim \mathrm{Inv}{-}\chi^2\left(\nu_n, \sigma_n^2\right)$$

$$\mu_n = \left(\mathbf{X}^\top\mathbf{X} + \Omega_0\right)^{-1}\left(\mathbf{X}^\top\mathbf{X}\hat{\beta} + \Omega_0\mu_0\right)$$
$$\Omega_n = \mathbf{X}^\top\mathbf{X} + \Omega_0$$
$$\nu_n = \nu_0 + n$$
$$\sigma_n^2 = \left(\nu_0\sigma_0^2 + \mathbf{y}^\top\mathbf{y} + \mu_0^\top\Omega_0\mu_0 - \mu_n^\top\Omega_n\mu_n\right)/\nu_n$$

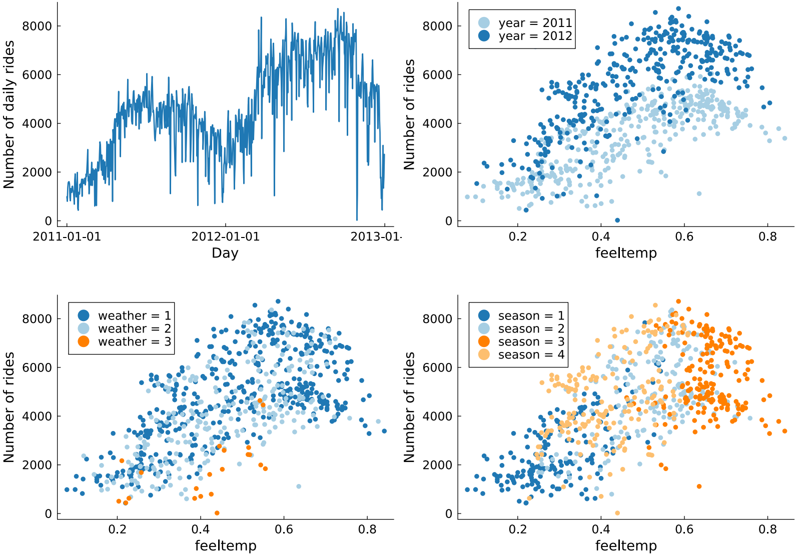
Bike share data

- **Bike share data**. Predict the number of bike rides.
- Response variable: number of rides on 731 days.

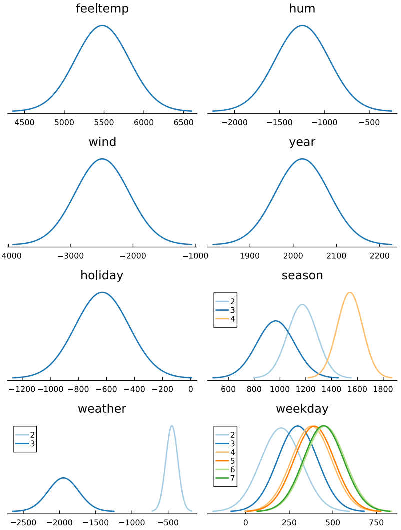
variable	description	type	values	comment
`nrides`	# of rides	counts	$\{0, 1, ...\}$	min$= 22$, max$= 8714$
`feeltemp`	perceived temp	cont.	$[0, 1]$	min$= 0.07$, max$= 0.85$
`hum`	humidity	cont.	$[0, 1]$	min$= 0.00$, max$= 0.98$
`wind`	wind speed	cont.	$[0, 1]$	min$= 0.02$, max$= 0.51$
`year`	year	binary	$\{0, 1\}$	year 2011 $= 0$
`season`	season	cat.	$\{1, 2, 3, 4\}$	winter $\rightarrow$ fall
`weather`	weather	ordinal	$\{1, 2, 3\}$	clear $\rightarrow$ rain/snow
`weekday`	day of week	cat.	$\{0, ..., 6\}$	sunday $\rightarrow$ saturday
`holiday`	holiday	binary	$\{0, 1\}$	holiday $= 1$

- Prior:
  - $\mu_0 = (1000, 0, \ldots, 0)^\top$
  - $\Omega_0 = \frac{\kappa_0}{n} \boldsymbol{X}^\top \boldsymbol{X}$ with $\kappa_0 = 1$ (unit information prior)
  - $\sigma_0^2 = 1000^2$ and $\nu_0 = 5$.

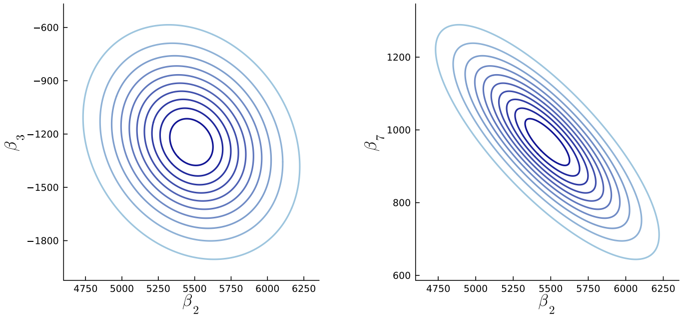
Bike share data



Bike share data - marginal posteriors of β



Bike share data - joint posteriors of β



Interactive - Bayesian regression

Polynomial regression

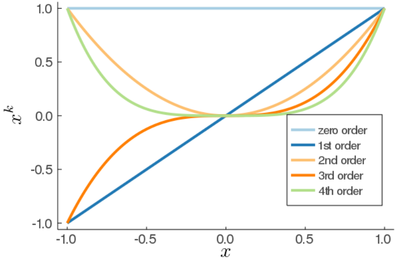
- **Polynomial regression**

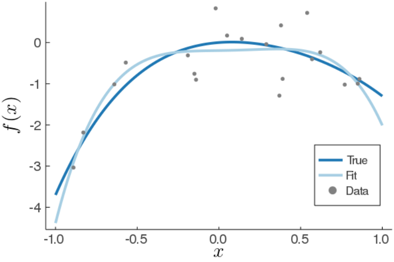
$$f(x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_k x_i^k.$$

$$\mathbf{y} = \mathbf{X}_P \beta + \varepsilon,$$

where

$$\mathbf{X}_P = (1, x, x^2, ..., x^k).$$





- Priors for **regularization** (ridge, lasso etc) in Lecture 6.

Prediction/Forecasting

- **Posterior predictive density** for future $\tilde{y}$ given observed $\boldsymbol{y} = (y_1, \ldots, y_n)$

$$p(\tilde{y}|\boldsymbol{y}) = \int_\theta p(\tilde{y}|\theta, \boldsymbol{y})p(\theta|\boldsymbol{y})d\theta$$

- IID data:

$$p(\tilde{y}|\boldsymbol{y}) = \int_\theta p(\tilde{y}|\theta)p(\theta|\boldsymbol{y})d\theta$$

- **Parameter uncertainty** in $p(\tilde{y}|\boldsymbol{y})$ by **averaging over** $p(\theta|\boldsymbol{y})$.

Prediction - Normal data, known variance

- Under the uniform prior $p(\theta) \propto c$, then

$$p(\tilde{y}|\mathbf{y}) = \int_{\theta} p(\tilde{y}|\theta)p(\theta|\mathbf{y})d\theta$$
$$\theta|\mathbf{y} \sim N(\bar{y}, \sigma^2/n)$$
$$\tilde{y}|\theta \sim N(\theta, \sigma^2)$$

Simulation algorithm:

1. Generate a **posterior draw** of θ ($\theta^{(1)}$) from $N(\bar{y}, \sigma^2/n)$
2. Generate a **predictive draw** of $\tilde{y}$ ($\tilde{y}^{(1)}$) from $N(\theta^{(1)}, \sigma^2)$
3. Repeat Steps 1 and 2 N times to output:
   - Sequence of posterior draws: $\theta^{(1)}, ..., \theta^{(N)}$
   - Sequence of predictive draws: $\tilde{y}^{(1)}, ..., \tilde{y}^{(N)}$.

Predictive distribution - Normal model

- $\theta^{(1)} = \bar{y} + \varepsilon^{(1)}$, where $\varepsilon^{(1)} \sim N(0, \sigma^2/n)$. (Step 1).
- $\tilde{y}^{(1)} = \theta^{(1)} + \upsilon^{(1)}$, where $\upsilon^{(1)} \sim N(0, \sigma^2)$. (Step 2).
- $\tilde{y}^{(1)} = \bar{y} + \varepsilon^{(1)} + \upsilon^{(1)}$.
- $\varepsilon^{(1)}$ and $\upsilon^{(1)}$ are independent.
- The sum of two normal random variables is normal so

$$\begin{aligned} E(\tilde{y}|\mathbf{y}) &= \bar{y} \\ V(\tilde{y}|\mathbf{y}) &= \frac{\sigma^2}{n} + \sigma^2 = \sigma^2\left(1 + \frac{1}{n}\right) \end{aligned}$$

$$\tilde{y}|\mathbf{y} \sim N\left[\bar{y}, \sigma^2\left(1 + \frac{1}{n}\right)\right]$$

Iteration laws

- Expectation with respect to what? Explicit:

$$\mathbb{E}_{\theta|\mathbf{y}}(\theta) \equiv \int \theta p(\theta|\mathbf{y}) d\theta$$

- **Law of iterated expectation** and **Law of total variance**.

Iteration laws

Law of iterated expectation:

$$\mathbb{E}_X(X) = \mathbb{E}_Y\left(\mathbb{E}_{X|Y}(X)\right)$$

Law of total variance:

$$\mathbb{V}_X(X) = \mathbb{E}_Y\left(\mathbb{V}_{X|Y}(X)\right) + \mathbb{V}_Y\left(\mathbb{E}_{X|Y}(X)\right)$$

Iteration laws for Bayes

Marginal posterior mean:

$$\mathbb{E}_{\theta_1|\mathbf{y}}(\theta_1) = \mathbb{E}_{\theta_2|\mathbf{y}}\left(\mathbb{E}_{\theta_1|\theta_2,\mathbf{y}}(\theta_1)\right)$$

Marginal posterior variance:

$$\mathbb{V}_{\theta_1}(\theta_1) = \mathbb{E}_{\theta_2|\mathbf{y}}\left(\mathbb{V}_{\theta_1|\theta_2,\mathbf{y}}(\theta_1)\right) + \mathbb{V}_{\theta_2|\mathbf{y}}\left(\mathbb{E}_{\theta_1|\theta_2,\mathbf{y}}(\theta_1)\right)$$

Predictive distribution - Normal model and prior

- Predictive distribution still normal (sum of normals is normal).
- Predictive mean conditional on θ is trivial:

$$E_{\tilde{y}|\theta}(\tilde{y}) = \theta$$

- "Remove the conditioning" on θ by averaging over posterior:

$$E(\tilde{y}|\boldsymbol{y}) = E_{\theta|\boldsymbol{y}}(\theta) = \mu_n \text{ (Posterior mean of } \theta).$$

- The predictive variance of $\tilde{y}$ by **law of total variance**

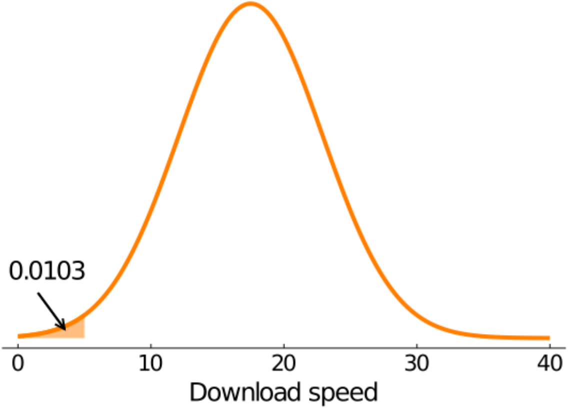
$$\begin{aligned} V(\tilde{y}|\boldsymbol{y}) &= E_{\theta|\boldsymbol{y}}[V_{\tilde{y}|\theta}(\tilde{y})] + V_{\theta|\boldsymbol{y}}[E_{\tilde{y}|\theta}(\tilde{y})] \\ &= E_{\theta|\boldsymbol{y}}(\sigma^2) + V_{\theta|\boldsymbol{y}}(\theta) \\ &= \sigma^2 + \tau_n^2 \end{aligned}$$

- So, predictive distribution is

$$\tilde{y}|\boldsymbol{y} \sim N(\mu_n, \sigma^2 + \tau_n^2).$$

Predictive distribution - Internet speed data

- My Netflix starts to buffer at speeds < 5Mbit. 🤬



Bayesian prediction for time series

Autoregressive process

$$y_t = \mu + \phi_1(y_{t-1} - \mu) + \ldots + \phi_p(y_{t-p} - \mu) + \varepsilon_t, \;\; \varepsilon_t \overset{iid}{\sim} N(0, \sigma^2)$$

Predictive distribution - AR process.

Input: time series $\mathbf{y}_{1:T} = (y_1, \ldots, y_T)$
number of predictive draws m.
forecast horizon h.

for i in $1{:}m$ **do**

$\mu, \phi_1, \ldots, \phi_p, \sigma \leftarrow \text{rPosteriorAR}(\mathbf{y}_{1:T}, \text{PriorSettings})$

$\varepsilon_{T+1} \leftarrow \text{rNorm}(0, \sigma)$

$\tilde{y}_{T+1} \leftarrow \mu + \phi_1(y_T - \mu) + \ldots + \phi_p(y_{T+1-p} - \mu) + \varepsilon_{T+1}$

$\varepsilon_{T+2} \leftarrow \text{rNorm}(0, \sigma)$

$\tilde{y}_{T+2} \leftarrow \mu + \phi_1(\tilde{y}_{T+1} - \mu) + \ldots + \phi_p(y_{T+2-p} - \mu) + \varepsilon_{T+2}$

$\vdots$

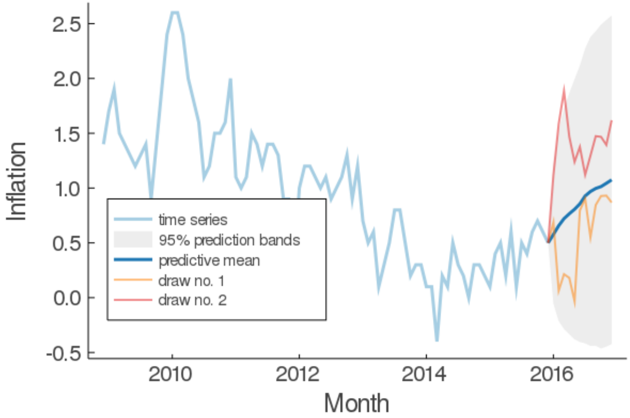
$\varepsilon_{T+h} \leftarrow \text{rNorm}(0, \sigma)$

$\tilde{y}_{T+h} \leftarrow \mu + \phi_1(\tilde{y}_{T+h-1} - \mu) + \ldots + \phi_p(\tilde{y}_{T+h-p} - \mu) + \varepsilon_{T+h}$

end

Output: m draws from the joint predictive density:
$p(\tilde{y}_{T+1}, \ldots, \tilde{y}_{T+h} | \mathbf{y}_{1:T})$.

Bayesian prediction of Swedish inflation



Decision problems

- Let θ be an **unknown quantity**. **State of nature**.
  - Future inflation
  - Global temperature
  - Disease.
- Let $a \in \mathcal{A}$ be an **action**.
  - Interest rate
  - Energy tax
  - Surgery.
- Choosing action a when state of nature is θ gives **utility**

$$U(a, \theta)$$

- Alternatively **loss** $L(a, \theta) = -U(a, \theta)$.

Decision tables - when both a and θ are discrete

- **Decision table**

	θ_1	θ_2	$\cdots$	θ_K
a_1	$u(a_1, \theta_1)$	$u(a_1, \theta_2)$	$\cdots$	$u(a_1, \theta_K)$
a_2	$u(a_2, \theta_1)$	$u(a_2, \theta_2)$	$\cdots$	$u(a_2, \theta_K)$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
a_J	$u(a_J, \theta_1)$	$u(a_J, \theta_2)$	$\cdots$	$u(a_J, \theta_K)$

- The eternal umbrella decision:

	Rain	Sun
No umbrella	-50	50
Umbrella	10	30

Decision Theory

- Example **loss functions** when both a and θ are continuous:
  - **Linear**: $L(a,\theta) = |a-\theta|$
  - **Quadratic**: $L(a,\theta) = (a-\theta)^2$
  - **Lin-Lin**:
  $$L(a,\theta) = \begin{cases} c_1 \cdot |a-\theta| & \text{if } a \leq \theta \\ c_2 \cdot |a-\theta| & \text{if } a > \theta \end{cases}$$
- Example:
  - θ is the number of items demanded of a product
  - a is the number of items in stock
  - Utility
  $$U(a,\theta) = \begin{cases} p \cdot \theta - c_1(a-\theta) & \text{if } a > \theta \text{ [too much stock]} \\ p \cdot a - c_2(\theta - a)^2 & \text{if } a \leq \theta \text{ [too little stock]} \end{cases}$$

Optimal decisions

- Ad hoc decision rules:
  - *Minimax*. Minimizes the maximum loss.
  - *Minimax-regret*
  - ... 😴
- **Bayesian theory**: maximize **posterior expected utility** 😍

$$a_{bayes} = \mathrm{argmax}_{a \in \mathcal{A}} \; E_{p(\theta|y)}[U(a,\theta)],$$

where $E_{p(\theta|y)}$ denotes the posterior expectation.

- Using simulated draws $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(N)}$ from $p(\theta|y)$:

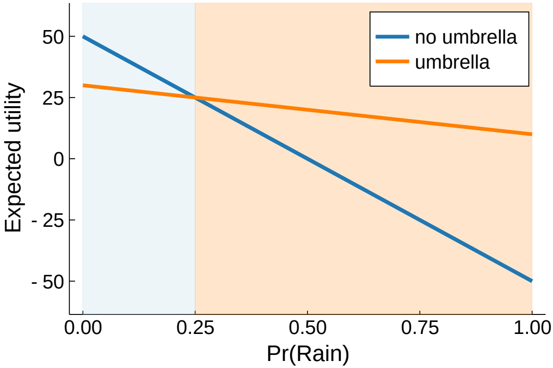
$$E_{p(\theta|y)}[U(a,\theta)] \approx N^{-1} \sum_{i=1}^{N} U(a,\theta^{(i)})$$

- **Separation principle**:

1. First obtain $p(\theta|y)$
2. then form $U(a,\theta)$ and finally
3. choose a that maximizes $E_{p(\theta|y)}[U(a,\theta)]$.

The umbrella decision

	Rain	Sun
No umbrella	-50	50
Umbrella	10	30



Choosing a point estimate is a decision

- Choosing a **point estimator** is a decision problem.
- Which to choose: posterior median, mean or mode?
- It depends on your loss function:
  - **Linear loss** $\rightarrow$ Posterior median
  - **Quadratic loss** $\rightarrow$ Posterior mean
  - **Zero-one loss** $\rightarrow$ Posterior mode
  - **Lin-Lin loss** $\rightarrow$ $c_1/(c_1+c_2)$ quantile of the posterior

The umbrella decision