

Bayesian Statistics I

Lecture 5 - Regression, Prediction and Decisions

Mattias Villani

Department of Statistics
Stockholm University

Department of Computer and Information Science
Linköping University



mattiasvillani.com



@matvil



[mattiasvillani](https://github.com/mattiasvillani)

Lecture overview

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

Linear regression

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

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

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$
$$X = \begin{pmatrix} x'_1 \\ \vdots \\ x'_n \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

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

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

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 :

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

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

- **Simulate** from the joint posterior by simulating from

- ▶ $p(\sigma^2 | y)$
- ▶ $p(\beta | \sigma^2, y)$

- **Marginal posterior** of β :

$$\beta | y \sim t_{n-k}[\hat{\beta}, s^2(X'X)^{-1}]$$

Linear regression - conjugate prior

■ Joint prior for β and σ^2

$$\begin{aligned}\beta|\sigma^2 &\sim N(\mu_0, \sigma^2 \Omega_0^{-1}) \\ \sigma^2 &\sim \text{Inv} - \chi^2(\nu_0, \sigma_0^2)\end{aligned}$$

■ Posterior

$$\begin{aligned}\beta|\sigma^2, y &\sim N[\mu_n, \sigma^2 \Omega_n^{-1}] \\ \sigma^2|y &\sim \text{Inv} - \chi^2(\nu_n, \sigma_n^2)\end{aligned}$$

$$\begin{aligned}\mu_n &= (X'X + \Omega_0)^{-1} (X'X\hat{\beta} + \Omega_0\mu_0) \\ \Omega_n &= X'X + \Omega_0 \\ \nu_n &= \nu_0 + n \\ \nu_n\sigma_n^2 &= \nu_0\sigma_0^2 + (y'y + \mu_0'\Omega_0\mu_0 - \mu_n'\Omega_n\mu_n)\end{aligned}$$

Polynomial regression

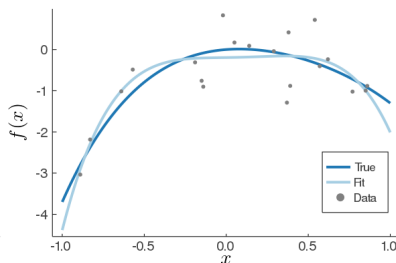
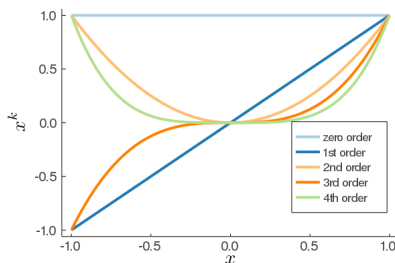
Polynomial regression

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

$$y = X_P \beta + \varepsilon,$$

where

$$X_P = (1, x, x^2, \dots, x^k).$$



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

- **Posterior predictive density** for future \tilde{y} given observed y

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

- IID data:

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

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

Prediction - Normal data, known variance

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

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

$$\theta|y \sim N(\bar{y}, \sigma^2/n)$$

$$\tilde{y}|\theta \sim N(\theta, \sigma^2)$$

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: $\theta^{(1)}, \dots, \theta^{(N)}$
 - ▶ Sequence of predictive draws: $\tilde{y}^{(1)}, \dots, \tilde{y}^{(N)}$.

Predictive distribution - Normal model

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

$$E(\tilde{y}|y) = \bar{y}$$

$$V(\tilde{y}|y) = \frac{\sigma^2}{n} + \sigma^2 = \sigma^2 \left(1 + \frac{1}{n}\right)$$

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

Iteration laws

- Expectation with respect to what? Explicit:

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

- Law of iterated expectation and Law of total variance.

Iteration laws

Law of iterated expectation:

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

Law of total variance:

$$\begin{aligned}\mathbb{V}_X(X) &= \mathbb{E}_Y(\mathbb{V}_{X|Y}(X)) \\ &\quad + \mathbb{V}_Y(\mathbb{E}_{X|Y}(X))\end{aligned}$$

Iteration laws for Bayes

Marginal posterior mean:

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

Marginal posterior variance:

$$\begin{aligned}\mathbb{V}_{\theta_1}(\theta_1) &= \mathbb{E}_{\theta_2|y}(\mathbb{V}_{\theta_1|\theta_2,y}(\theta_1)) \\ &\quad + \mathbb{V}_{\theta_2|y}(\mathbb{E}_{\theta_1|\theta_2,y}(\theta_1))\end{aligned}$$

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:

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

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

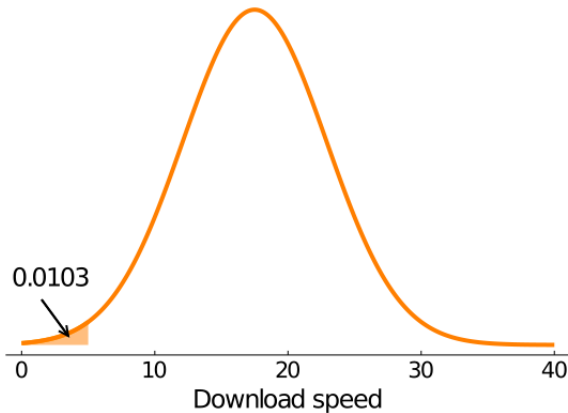
$$\begin{aligned} V(\tilde{y}|y) &= E_{\theta|y}[V_{\tilde{y}|\theta}(\tilde{y})] + V_{\theta|y}[E_{\tilde{y}|\theta}(\tilde{y})] \\ &= E_{\theta|y}(\sigma^2) + V_{\theta|y}(\theta) \\ &= \sigma^2 + \tau_n^2 \end{aligned}$$

- So, predictive distribution is

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

Predictive distribution - Internet speed data

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



Bayesian prediction for time series

■ Autoregressive process

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

Predictive distribution - AR process.

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

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

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

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

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

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

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

\vdots

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

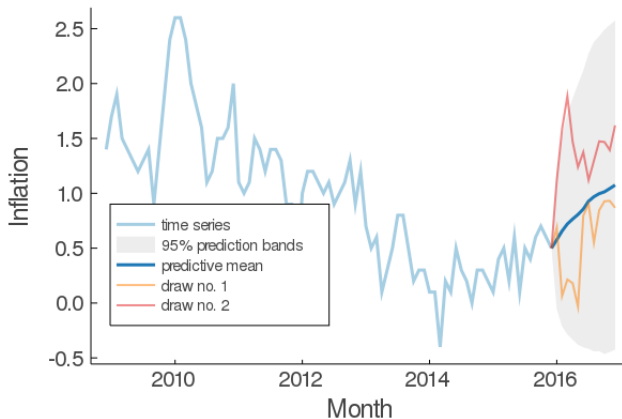
$\tilde{y}_{T+h} \leftarrow \mu + \phi_1(\tilde{y}_{T+h-1} - \mu) + \dots + \phi_p(\tilde{y}_{T+h-p} - \mu) + \varepsilon_{T+h}$

end

Output: m draws from the joint predictive density:

$$p(\tilde{y}_{T+1}, \dots, \tilde{y}_{T+h} | \mathbf{y}_{1:T}).$$

Bayesian prediction of Swedish inflation



Predicting auction prices on eBay

- Problem: **Predicting the final price** in eBay coin auctions.
- **Data**: Bid from 1000 auctions on eBay.
The highest bid is not observed (eBay proxy bidding).
- **Covariates** are auction-specific:
 - ▶ catalog value
 - ▶ seller's **reservation price**
 - ▶ quality
 - ▶ rating of seller etc
- Buyers are **strategic**.
 - ▶ Bid \neq **valuation**.
 - ▶ **Bid function**, $b = \text{BidFunction}(v)$, from **Game theory**.
 - ▶ Very complicated likelihood.

Simulating auction prices on eBay

Predictive distribution - auction price.

Input: training auction bids \mathbf{Y}

training auction covariates \mathbf{X} .

test auction covariates $\tilde{\mathbf{x}}$.

number of predictive draws m .

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

$\mu, \sigma, \lambda \leftarrow \text{RPOSTAUCTION}(\mathbf{Y}, \mathbf{X}, \tilde{\mathbf{x}}, \text{Prior})$ # parameters

$\tilde{n} \leftarrow \text{RPOIS}(\lambda(\tilde{\mathbf{x}}))$ # number of bidders in test auction

$\tilde{\mathbf{v}}_{1:\tilde{n}} \leftarrow \text{RNORM}(\mu(\tilde{\mathbf{x}}), \sigma(\tilde{\mathbf{x}}))$ # valuations for all \tilde{n} bidders

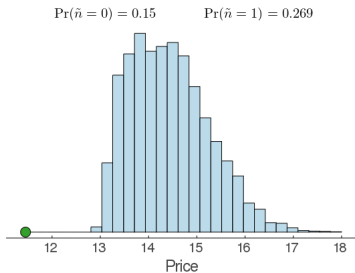
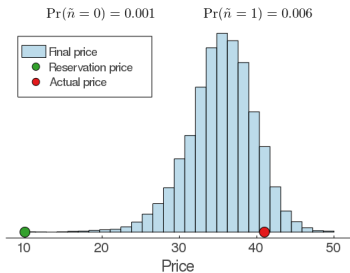
$b_{1:\tilde{n}} \leftarrow \text{BIDFUNCTION}(\tilde{\mathbf{v}}_{1:\tilde{n}}, \tilde{n}, \mu(\tilde{\mathbf{x}}), \sigma(\tilde{\mathbf{x}}))$ # bids

$\tilde{p} \leftarrow \text{SECONDLARGEST}(b_{1:\tilde{n}})$ # final price

end

Output: m predictive draws of the final price \tilde{p} for an auction with covariates $\tilde{\mathbf{x}}$.

Predicting auction prices on eBay



Decision Theory

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

$$U(a, \theta)$$

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

- Loss table:

	θ_1	θ_2
a_1	$L(a_1, \theta_1)$	$L(a_1, \theta_2)$
a_2	$L(a_2, \theta_1)$	$L(a_2, \theta_2)$

- Example:

	Rainy	Sunny
Umbrella	20	10
No umbrella	50	0

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 decision

- Ad hoc decision rules:

- ▶ *Minimax*. Minimizes the maximum loss.
- ▶ *Minimax-regret* ... bla bla bla ...

- **Bayesian theory**: maximize the **posterior expected utility**:

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

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

- Using simulated draws $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(N)}$ 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 $E_{p(\theta|y)}[U(a, \theta)]$.

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** → Posterior median
 - ▶ **Quadratic loss** → Posterior mean
 - ▶ **Zero-one loss** → Posterior mode
 - ▶ **Lin-Lin loss** → $c_2 / (c_1 + c_2)$ quantile of the posterior