STAT 576 Bayesian Analysis

Lecture 12: Bayesian Regression Models

Chencheng Cai

Washington State University

Conditional Modeling

- Traditional regression models are based on the conditional distribution of the response variable given the covariates.

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

  where
  - $\boldsymbol{y}$ is the response variable $(n \times 1)$,
  - $\boldsymbol{X}$ is the design matrix $(n \times p)$,
  - $\boldsymbol{\beta}$ is the regression coefficients $(p \times 1)$,
  - $\boldsymbol{\epsilon}$ is the error term $(n \times 1)$.
- It is often assumed that
  - $\boldsymbol{X}$ is fixed and known,
  - $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$.

Conditional Modeling

- The inference is based on the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{X}$, $\boldsymbol{\beta}$ and σ^2.:

$$\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}).$$

- Frequentists maximize the log-likelihood function:

$$\ell(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y}) = -\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 - \frac{n}{2}\log\sigma^2$$

- The MLE therefore is given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}, \quad \hat{\sigma}^2 = \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2.$$

- However, it is **not** a full probabilistic model.

Bayesian Linear Regression

- In Bayesian regression, we treat $\boldsymbol{\beta}$ and σ^2 as random variables.
- We put priors on $\boldsymbol{\beta}$ and σ^2:

$$\boldsymbol{\beta} \sim \pi(\boldsymbol{\beta}),$$
$$\sigma^2 \sim \pi(\sigma^2).$$

- The joint distribution of $\boldsymbol{y}$, $\boldsymbol{\beta}$ and σ^2 is given by

$$p(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2) = p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\beta})p(\sigma^2).$$

- The posterior distribution of $\boldsymbol{\beta}$ and σ^2 is given by

$$p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\beta})p(\sigma^2).$$

Bayesian Linear Regression

- The noninformative prior for $\boldsymbol{\beta}$ and σ^2 is often taken as

$$\pi(\boldsymbol{\beta}) \propto 1,$$
$$\pi(\sigma^2) \propto \frac{1}{\sigma^2}.$$

  Derivation: (1) Jeffreys prior (2) results for location-scale families.

- We can derive the posteroir distribution of $\boldsymbol{\beta}$ and σ^2 by

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}) &\propto p(\boldsymbol{y} | \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}) p(\sigma^2) \\ &\propto \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2} \|\boldsymbol{y} - \boldsymbol{X}\beta\|^2\right) \times 1 \times \frac{1}{\sigma^2} \\ &\propto \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2} \|\boldsymbol{y} - \boldsymbol{X}\beta\|^2\right) \times \frac{1}{\sigma^2}. \end{aligned}$$

Bayesian Linear Regression

- Notice that:

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \boldsymbol{X}^T \boldsymbol{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \|\boldsymbol{y}\|^2 - \|\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2$$

  where $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$.
- Therefore, the posterior distribution of $\boldsymbol{\beta}$ and σ^2 is given by

$$p(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}) \propto \sigma^{-n-2} \exp\left(-\frac{(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \boldsymbol{X}^T \boldsymbol{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \|\boldsymbol{y}\|^2 - \|\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2}{2\sigma^2} \right).$$

- Compared to Normal-Inverse-Gamma distribution, the normal component is replaced with a multivariate normal distribution.
- Compared to Normal-Inverse-Wishart distribution, the covariance component is replaced with $\sigma^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1}$.

Bayesian Linear Regression

$$p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) \propto \sigma^{-n-2} \exp\left(-\frac{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) + \|\boldsymbol{y}\|^2 - \|\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2}{2\sigma^2}\right)$$

- The conditional posterior of $\boldsymbol{\beta}$ given σ^2 and $\boldsymbol{y}$ is given by

$$\boldsymbol{\beta}|\sigma^2, \boldsymbol{y} \sim \mathcal{N}\left(\hat{\boldsymbol{\beta}}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)$$

- The conditional posterior of σ^2 given $\boldsymbol{\beta}$ and $\boldsymbol{y}$ is given by

$$\sigma^2 \mid \boldsymbol{\beta}, \boldsymbol{y} \sim \mathrm{InvGamma}\left(\frac{n}{2},\ \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2\right)$$

- The marginal posterior of σ^2 is given by

$$\sigma^2|\boldsymbol{y} \sim \mathrm{InvGamma}\left(\frac{n-p}{2},\ \frac{\|\boldsymbol{y}\|^2 - \|\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2}{2}\right)$$

Bayesian Linear Regression

$$p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) \propto \sigma^{-n-2} \exp\left(-\frac{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) + \|\boldsymbol{y}\|^2 - \|\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2}{2\sigma^2}\right)$$

- The marginal posterior of $\boldsymbol{\beta}$ can be obtained by

$$\begin{aligned} p(\boldsymbol{\beta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y})}{p(\sigma^2 \mid \boldsymbol{\beta}, \boldsymbol{y})} &\propto \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^{-n} \\ &\propto \left((\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) + \|\boldsymbol{y}\|^2 - \|\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2\right)^{-n/2} \\ &\propto \left(1 + \frac{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})}{\|\boldsymbol{y}\|^2 - \|\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2}\right)^{-n/2} \end{aligned}$$

- This ia a multivariate t distribution with degree $n-p$, mean $\hat{\boldsymbol{\beta}}$ and covariance $\frac{\|\boldsymbol{y}\|^2 - \|\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2}{n-p}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

Sampling from the Posterior

- Easier way:

$$\sigma^2 \mid \boldsymbol{y} \sim \text{InvGamma}\left(\frac{n-p}{2},\ \frac{\|\boldsymbol{y}\|^2 - \|\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2}{2}\right)$$

$$\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{y} \sim \mathcal{N}\left(\hat{\boldsymbol{\beta}}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)$$

- Harder way:

$$\boldsymbol{\beta} \mid \boldsymbol{y} \sim t_{n-p}\left(\hat{\boldsymbol{\beta}}, \frac{\|\boldsymbol{y}\|^2 - \|\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2}{n-p}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)$$

$$\sigma^2 \mid \boldsymbol{\beta}, \boldsymbol{y} \sim \text{InvGamma}\left(\frac{n}{2},\ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2\right)$$

Sampling from the Posterior

$$\sigma^2 \mid \boldsymbol{y} \sim \text{InvGamma}\left(\frac{n-p}{2},\ \frac{\|\boldsymbol{y}\|^2 - \|\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2}{2}\right)$$

$$\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{y} \sim \mathcal{N}\left(\hat{\boldsymbol{\beta}}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)$$

- Sampling from $\text{InvGamma}(\alpha, \beta)$:
  - Generate $x \sim \chi^2_{2\alpha}$,
  - Then $y = \frac{\beta}{2x}$.
- Sampling from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:
  - Cholesky decomposition: $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^T$, where $\boldsymbol{L}$ is lower triangular,
  - Generate $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$,
  - Then $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{z}$.

Predictive Distribution

Suppose σ^2 is known.

- The distribution for new observation $\tilde{\boldsymbol{y}}$ given new covariate $\tilde{\boldsymbol{X}}$ is given by

$$\tilde{\boldsymbol{y}}|\boldsymbol{y}, \sigma^2 \sim \mathcal{N}(\tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}}, \sigma^2 \boldsymbol{I} + \sigma^2 \tilde{\boldsymbol{X}}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\tilde{\boldsymbol{X}}^T).$$

  - The mean is $\tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}}$,
  - The variance is $\sigma^2\left(\boldsymbol{I} + \tilde{\boldsymbol{X}}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\tilde{\boldsymbol{X}}^T\right)$.

Suppose σ^2 is unknown.

- The distribution for new observation $\tilde{\boldsymbol{y}}$ given new covariate $\tilde{\boldsymbol{X}}$ is a linear transformation of a multivariate t distribution plus a Gaussian noise.
- The mean is $\tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}}$,
- The variance is $\frac{\|\boldsymbol{y}\|^2 - \|\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2}{n-p-2}\tilde{\boldsymbol{X}}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\tilde{\boldsymbol{X}}^T + \sigma^2\boldsymbol{I}$

Example

- Example from textbook Sec. 14.3.
- The data contains the election data for the U.S. House of Representatives in the past century (1900 – 2000).
- We would like to study the relationship between the percentage of votes for the incumbent party and the decision whether the incumbent officeholder runs for reelection.
- Goal: check if there is an advantage for the incumbent officeholder to reelect.
- Some facts of the data:
  - Election every two years.
  - The incumbent party is the party that won the previous election.
  - 435 districts in the U.S. House of Representatives.
  - Roughly 100 - 150 districts are uncontested.

Example

We formulate the problem as a simple linear regression model.

$$y_i = \alpha + \beta R_i + \epsilon_i$$

- y_i: the percentage of votes for the **incumbent party** in district i.
- R_i: a binary variable indicating whether the **incumbent officeholder** runs for reelection.
- α: the expected percentage of votes for the incumbent party when they incumbent officeholder **does not** run for reelection.
- $\alpha + \beta$: the expected percentage of votes for the incumbent party when the incumbent officeholder **does** run for reelection.
- β: **incumbency advantage**.

Example

- The currnet model may have selection bias in the dataset.
- I.e. some variables may affect both the decision of reelection and the percentage of votes.
- We should include those variables in the model as well.

$$y_i = \alpha + \beta R_i + \gamma z_i + \delta P_i + \epsilon_i$$

- z_i: the percentage of votes for the incumbent party in the **previous election**.
- P_i: the indicator for Democratic party (1) or Republican party (0) controlling the seat.

Example

With noninformative priors, the posterior inferences for the year 1988 are displayed below.

Variable	Posterior quantiles				
	2.5%	25%	median	75%	97.5%
Incumbency	0.084	0.103	0.114	0.124	0.144
Vote proportion in 1986	0.576	0.627	0.654	0.680	0.731
Incumbent party	−0.014	−0.009	−0.007	−0.004	0.001
Constant term	0.066	0.106	0.127	0.148	0.188
σ (residual sd)	0.061	0.064	0.066	0.068	0.071

- The incumbency advantage is estimated to be 11.4% and is significant.
- It shows a strong autoregressive effect in the percentage of votes for the incumbent party.
- Party differrence is not significant.

Genearlizations

We consider the following generalizations of the linear regression model in the subsequent slides.

- **Diverse Covariance Structures**: We may consider different covariance structures for the error term.
- **Regularization**: Sometimes we would like to choose a prior that encourages sparsity in the regression coefficients to prevent overfitting.
- **Hierarchical Linear Models**: We assume the regression coefficients are drawn from a common distribution for different subsets of data.

Covariance Structure

In the general case, we may consider the following covariance structures for the error term:

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma}$ is a positive definite matrix, that allows for different variances and correlations between the errors.

In this case, the model is given by

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$$

Covariance Structure — Known Covariance

If $\boldsymbol{\Sigma}$ is known, the posterior distribution of $\boldsymbol{\beta}$ is given by

$$\begin{aligned}
p(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{\Sigma}) &\propto p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\Sigma})p(\boldsymbol{\beta}) \\
&\propto \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right)\times 1 \\
&\propto \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right) \\
&\sim \mathcal{N}\left(\hat{\boldsymbol{\beta}},(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\right)
\end{aligned}$$

with

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

Covariance Structure — Unknown Covariance

If $\boldsymbol{\Sigma}$ is unknown, we may put a prior on $\boldsymbol{\Sigma}$ as well.

$$
\begin{aligned}
p(\boldsymbol{\Sigma} \mid \boldsymbol{y}, \boldsymbol{\beta}) &\propto \frac{p(\boldsymbol{\beta}, \boldsymbol{\Sigma} \mid \boldsymbol{y})}{p(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{\Sigma})} \\
&\propto p(\boldsymbol{\Sigma}) |\boldsymbol{\Sigma}|^{-1/2} |\boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{X}|^{1/2} \exp\left(-\frac{1}{2} (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \right)
\end{aligned}
$$

- It is difficult to set up a prior for $\boldsymbol{\Sigma}$.
- It is difficult to draw from this posterior distribution.
- Therefore, we often need some further simplification on $\boldsymbol{\Sigma}$.

Covariance Structure — Simplified Covariance

If the covariance matrix $\boldsymbol{\Sigma}$ is proportional to a known matrix $\boldsymbol{Q}$, that is

$$\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{Q}.$$

Then the posterior distribution of $\boldsymbol{\beta}$ is multivariate t and the posterior distribution of σ^2 is inverse gamma.

- One can derive it from the posterior distribution of $\boldsymbol{\beta}$ and σ^2 on the previous few slides.
- Or, it can be seen from the following transformation of data:

$$\boldsymbol{y}^* = \boldsymbol{Q}^{-1/2}\boldsymbol{y},$$
$$\boldsymbol{X}^* = \boldsymbol{Q}^{-1/2}\boldsymbol{X}.$$

  $\boldsymbol{Q}^{-1/2}$ is any matrix such that $(\boldsymbol{Q}^{-1/2})^T \boldsymbol{Q} \boldsymbol{Q}^{-1/2} = \boldsymbol{I}$.
  Then the linear regression problem becomes regress $\boldsymbol{y}^*$ on $\boldsymbol{X}^*$ with i.i.d. noise.
  All previous results apply.

Covariance Structure — Simplified Covariance

In a weighted regression model, we may consider the following covariance structure for the error term:

$$\Sigma_{ii} = \sigma^2 / w_i$$

where w_i is the weight for the ith observation, and Σ_{ii} is the ith diagonal element of $\mathbf{\Sigma}$.

- The model is the same as the previous one, with

$$\boldsymbol{Q} = \mathrm{diag}(w_1, \ldots, w_n)$$

- All previous results apply.

Covariance Structure — Simplified Covariance

The unequal weights can be generalized to a more general setting by introducing the unequalness parameter ϕ such that

$$\Sigma_{ii} = \sigma^2 v(w_i, \phi)$$

where $\phi \in [0, 1]$ controls the unequalness.

- Example: $v(w_i, \phi) = w_i^{-\phi}$. $\phi = 0$ is the equal weight case and $\phi = 1$ is the inverse weight case.
- Example: $v(w_i, \phi) = 1 - \phi + \phi/w_i$. $\phi = 0$ is the equal weight case and $\phi = 1$ is the inverse weight case.
- A naturla noninformative prior for ϕ is the uniform distribution on $[0, 1]$.
- For the posterior and its sampling, please check textbook Eq. (14.21) and (14.22).

Regularization

In linear regression problem, the regularized least squares minimize the following objective function:

$$\min_{\boldsymbol{\beta}} \ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda R(\boldsymbol{\beta}),$$

where $R(\boldsymbol{\beta})$ is a penalty term that penalizes the complexity of the model.

- Ridge regression: $R(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|^2$.
- Lasso regression: $R(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$.
- Elastic net: $R(\boldsymbol{\beta}) = \alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\|\boldsymbol{\beta}\|^2$.

- Notice that the sum of squared errors is equivalent to the negative log-likelihood function.
- The regularized least squares is equivalent to the maximum a posteriori estimation with a prior on $\boldsymbol{\beta}$ that corresponds to the exponential of the negative penalty.

Regularization — Ridge

In Ridge regression, we put a Gaussian prior on $\boldsymbol{\beta}$:

$$p(\boldsymbol{\beta}) \propto \exp\left(-\frac{\lambda}{2\sigma^2}\|\boldsymbol{\beta}\|^2\right)$$

This is a multivariate normal distribution with mean $\mathbf{0}$ and covariance $\frac{\sigma^2}{\lambda}\boldsymbol{I}$.

The posterior is (under noninformative prior for σ^2)

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) &\propto \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 - \lambda\|\boldsymbol{\beta}\|^2\right) \\ &\propto \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right) \\ &\quad \times \exp\left(-\frac{1}{2\sigma^2}\left(\|\boldsymbol{y}\|^2 - \boldsymbol{y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}\right)\right) \end{aligned}$$

with $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$. The conditional/marginal posteriors are the similar as before except that $\boldsymbol{X}^T\boldsymbol{X}$ is replaced with $\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}$.

Regularization — LASSO

In LASSO (Least Absolute Shrinkage and Selection Operator) regression, we put a Laplace prior on $\boldsymbol{\beta}$:

$$p(\boldsymbol{\beta}) \propto \exp\left(-\frac{\lambda}{2\sigma^2}\|\boldsymbol{\beta}\|_1\right)$$

- The posterior distribution is not a standard distribution.
- We usually do not have a closed form for the posterior mode.
- The posterior mode can force some coefficients to be exactly zero, resulting in a sparse model.
- The sparsity is due to the non-differentiability of the prior at 0.
- Or, the sub-derivative of the prior at 0 contains a neighborhood of 0.

Regularization — Spike-and-Slab

Besides the Ridge and LASSO, which "encourage" coefficients to be small through the prior, we may also consider the Spike-and-Slab prior that directly set a probability for the coefficient to be zero.
Specifically, for each coefficient β_j, we set a prior as

$$p(\beta_j) = \theta \underbrace{\delta(\beta_j)}_{spike} + (1-\theta) \underbrace{p_{slab}(\beta_j)}_{slab},$$

- The prior is a mixture of a point mass at 0 and a continuous distribution.
- $\delta(\beta_j)$ is the Dirac delta function at 0 corresponding to the "spike" component.
- $p_{slab}(\beta_j)$ is the continuous distribution corresponding to the "slab" component. p_{slab} can be chosen as uniform, Gaussian, etc..
- θ is the probability of sparsity that controls the mixture rate between the two components.

Regularization — Spike-and-Slab

In practice, several modifications can be used to make inference with the Spike-and-Slab prior:

- It is often more conveinent to introduce a binary variable z_j such that

$$\begin{aligned} z_j &\sim \text{Bernoulli}(\theta), \\ \beta_j \mid z_j = 1 &\sim \delta_0, \\ \beta_j \mid z_j = 0 &\sim p_{slab}. \end{aligned}$$

- It is often more conveient to set the spike component as a Gaussian distribution with a very small variance, and the slab component as a Gaussian distribution with a larger variance.
- Sampling from the posterior distribution is often done by Gibbs sampling for $(\boldsymbol{\beta}, \boldsymbol{z})$.

Hierarchical Linear Models

If we have linear regression models for different subsets of data, we may assume that the regression coefficients are drawn from a common distribution.

$$\boldsymbol{y}_i = \boldsymbol{X}_i \boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i$$

with

$$\boldsymbol{\beta}_i \sim P, i.i.d.$$

where P is common distribution for the linear regression coefficients.

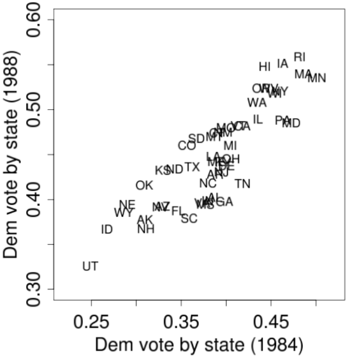
- When P is Gaussian, the model is also called a random effects model.
- Sometimes, only part of the $\boldsymbol{\beta}_i$ are random effects, and the rest are fixed effects (same for all groups).
- If the random effects in above are normal, the model is also called a mixed effects model.

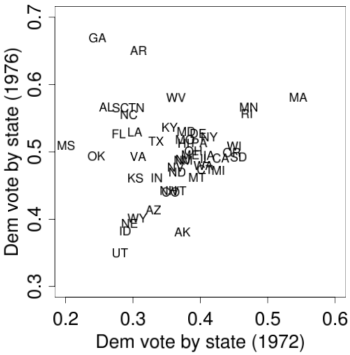
Example: U.S. presidential elections

The data contains results from the U.S. presidential elections for all states from 1948 to 1988.

- 511 records by removing the District of Columbia and all third-party victories.
- The response variable is the percentage of votes for the Democratic party.

Example: U.S. presidential elections





- Previous election results have a strong effect on the current election results.
- Some outiliers from the southern states. (Upper left on the second graph)

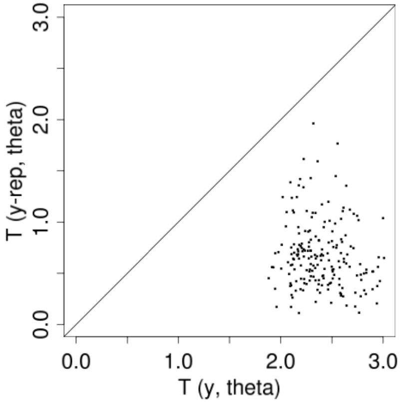
Example: U.S. presidential elections

All covariates used for linear regression:

Description of variable	Sample quantiles		
	min	median	max
Nationwide variables:			
Support for Dem. candidate in Sept. poll	0.37	0.46	0.69
(Presidential approval in July poll) × Inc	−0.69	−0.47	0.74
(Presidential approval in July poll) × Presinc	−0.69	0	0.74
(2nd quarter GNP growth) × Inc	−0.024	−0.005	0.018
Statewide variables:			
Dem. share of state vote in last election	−0.23	−0.02	0.41
Dem. share of state vote two elections ago	−0.48	−0.02	0.41
Home states of presidential candidates	−1	0	1
Home states of vice-presidential candidates	−1	0	1
Democratic majority in the state legislature	−0.49	0.07	0.50
(State economic growth in past year) × Inc	−0.22	−0.00	0.26
Measure of state ideology	−0.78	−0.02	0.69
Ideological compatibility with candidates	−0.32	−0.05	0.32
Proportion Catholic in 1960 (compared to U.S. avg.)	−0.21	0	0.38
Regional/subregional variables:			
South	0	0	1
(South in 1964) × (−1)	−1	0	0
(Deep South in 1964) × (−1)	−1	0	0
New England in 1964	0	0	1
North Central in 1972	0	0	1
(West in 1976) × (−1)	−1	0	0

Example: U.S. presidential elections

We compare the values of the test variable $T(\boldsymbol{y}, \boldsymbol{\theta})$ from the posterior simulations of $\boldsymbol{\beta}$ to the hypothetical replicated valuesunder the model, $T(\boldsymbol{y}^{(rep)}, \boldsymbol{\theta})$.



The performance is not satisfactory.

Example: U.S. presidential elections

Now we consider a hierarchical model for the data.

$$
\begin{aligned}
y_{st} &\sim \mathcal{N}(X_{st}\boldsymbol{\beta} + \gamma_{r(s)t} + \delta_t, \sigma^2), \\
\gamma_{rt} &\sim \begin{cases} \mathcal{N}(0, \tau_{\gamma 1}^2) & for\ r = 1, 2, 3 \text{ (non-south)} \\ \mathcal{N}(0, \tau_{\gamma 2}^2) & for\ r = 4 \text{ (south)} \end{cases} \\
\delta_t &\sim \mathcal{N}(0, \tau_\delta^2)
\end{aligned}
$$

- γ_{rt}: different intercepts for different regions.
- δ_t: different intercepts for different years.
- $\boldsymbol{\beta}$ dependence on other covariates is assumed to be the same for all regions and years.
- Hyperprior for the hyperparameters are set to uniform.

Example: U.S. presidential elections

We conduct the Bayesian predictive checks for the hierarchical model.