A Bayesian Treatment of Linear Gaussian Regression

Frank Wood

December 3, 2009

Bayesian Approach to Classical Linear Regression

In classical linear regression we have the following model

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

Unfortunately we often don't know the observation error σ^2 and, as well, we don't know the vector of linear weights $\boldsymbol{\beta}$ that relates the input(s) to the output.

In Bayesian regression, we are interested in several inference objectives. One is the posterior distribution of the model parameters, in particular the posterior distribution of the observation error variance given the inputs and the outputs.

$$P(\sigma^2|\mathbf{X}, \mathbf{y})$$

Posterior Distribution of the Error Variance

Of course in order to derive

$$P(\sigma^2|\mathbf{X}, \mathbf{y})$$

We have to treat $\boldsymbol{\beta}$ as a nuisance parameter and integrate it out

$$\begin{aligned} P(\sigma^2|\mathbf{X}, \mathbf{y}) &= \int P(\sigma^2, \boldsymbol{\beta}|\mathbf{X}, \mathbf{y}) d\boldsymbol{\beta} \\ &= \int P(\sigma^2|\boldsymbol{\beta}, \mathbf{X}, \mathbf{y}) P(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y}) d\boldsymbol{\beta} \end{aligned}$$

Predicting a New Output for a (set of) new Input(s)

Of particular interest is the ability to predict the distribution of output values for a new input

$$P(\mathbf{y}_{new}|\mathbf{X}, \mathbf{y}, \mathbf{X}_{new})$$

Here we have to treat both σ^2 and $\boldsymbol{\beta}$ as a nuisance parameters and integrate them out

$$\begin{aligned} & P(\mathbf{y}_{new}|\mathbf{X}, \mathbf{y}, \mathbf{X}_{new}) \\ &= \int\int P(\mathbf{y}_{new}|\boldsymbol{\beta}, \sigma^2) P(\sigma^2|\boldsymbol{\beta}, \mathbf{X}, \mathbf{y}) P(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y}) d\boldsymbol{\beta}, d\sigma^2 \end{aligned}$$

Noninformative Prior for Classical Regression

For both objectives, we need to place a prior on the model parameters σ^2 and $\boldsymbol{\beta}$. We will choose a noninformative prior to demonstrate the connection between the Bayesian approach to multiple regression and the classical approach.

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

Is this a proper prior? What form will the posterior take in this case? Will it be proper?

Clearly other priors can be imposed, priors that are more informative.

Posterior distribution of $\boldsymbol{\beta}$ given σ^2

Sometimes it is the case that σ^2 is known. In such cases the posterior distribution over the model parameters collapses to the posterior over $\boldsymbol{\beta}$ alone. Even when σ^2 is also unknown, the factorization of the posterior distribution

$$P(\sigma^2, \boldsymbol{\beta}|\mathbf{X}, \mathbf{y}) = P(\boldsymbol{\beta}|\sigma^2, \mathbf{X}, \mathbf{y})P(\sigma^2|\mathbf{X}, \mathbf{y})$$

Suggests that determining the posterior distribution $P(\boldsymbol{\beta}|\sigma^2, \mathbf{X}, \mathbf{y})$ will be of use as a step in posterior analyses.

Posterior distribution of β given σ^2

Given our choice of (improper) prior we have

$$P(\beta|\sigma^2, \mathbf{X}, \mathbf{y})P(\sigma^2|\mathbf{X}, \mathbf{y}) \propto \mathrm{N}(\mathbf{y}|\mathbf{X}\beta, \sigma^2\mathbf{I})\sigma^{-2}$$

Which, plugging in the normal likelihood and ignoring terms that are not a function of β we have

$$P(\beta|\sigma^2, \mathbf{X}, \mathbf{y}) \quad \propto \quad \exp(-\frac{1}{2}(\mathbf{y} - \mathbf{X}\beta)^T \frac{1}{\sigma^2}\mathbf{I}(\mathbf{y} - \mathbf{X}\beta)))$$

when we expand out the exponent we get an expression that looks like (again dropping terms that do not involve β)

$$\exp(-\frac{1}{2}(-2\mathbf{y}^T \frac{1}{\sigma^2}\mathbf{I}\mathbf{X}\beta + \beta^T\mathbf{X}^T \frac{1}{\sigma^2}\mathbf{I}\mathbf{X}\beta))$$

Multivariate Quadratic Square Completion

We recognize the familiar form of the exponent of a multivariate Gaussian in this expression and can derive the mean and the variance of the distribution of $\beta|\sigma^2, \ldots$ by noting that

$$(\beta - \mu_\beta)^T \mathbf{\Sigma}_\beta^{-1} (\beta - \mu_\beta) = \beta^T \mathbf{\Sigma}^{-1} \beta - 2\mu_\beta^T \mathbf{\Sigma}_\beta^{-1} \beta + const$$

From this and the result from the previous slide

$$\exp(\frac{1}{2} - 2\mathbf{y}^T \frac{1}{\sigma^2} \mathbf{I}\mathbf{X}\beta + \beta^T \mathbf{X}^T \frac{1}{\sigma^2} \mathbf{I}\mathbf{X}\beta)$$

We can immediately identify $\mathbf{\Sigma}_\beta^{-1} = \mathbf{X}^T \frac{1}{\sigma^2} \mathbf{I}\mathbf{X}$ and thus that $\mathbf{\Sigma}_\beta = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$. Similarly we can solve for μ_β and we find

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

Distribution of β given σ^2

Mirroring the classical approach to matrix regression we have that the distribution of the regression coefficients given the observation noise variance is

$$\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^2 \sim \mathrm{N}(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$and $\boldsymbol{\mu}_{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

Note that $\boldsymbol{\mu}_{\boldsymbol{\beta}}$ is the same as the maximum likelihood or least squares estimate $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ of the regression coefficients.

Of course we don't usually know the observation noise variance σ^2 and have to simultaneously estimate it from the data. To determine the distribution of this quantity we need a few facts.

Scaled inverse-chi-square distribution

If $\theta \sim \text{Inv}-\chi^2(\nu, s^2)$ then the pdf for θ is given by

$$\begin{aligned} P(\theta) &= \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} \theta^{-(\nu/2+1)} e^{(-\nu s^2/(2\theta))} \\ &\propto \theta^{-(\nu/2+1)} e^{(-\nu s^2/(2\theta))} \end{aligned}$$

You can think of the scaled inverse chi squared distribution as the chi squared distribution where the sum of squares is explicit in the parameterization. $\nu > 0$ is the number of "degrees of freedom", $s > 0$ is the scale parameter.

Distribution of σ^2 given observations $\mathbf{y}$ and $\mathbf{X}$

The posterior distribution of the observation noise can be derived by noting that

$$\begin{aligned} P(\sigma^2|\mathbf{y},\mathbf{X}) &= \frac{P(\boldsymbol{\beta},\sigma^2|\mathbf{y},\mathbf{X})}{P(\boldsymbol{\beta}|\sigma^2,\mathbf{y},\mathbf{X})} \\ &\propto \frac{P(\mathbf{y}|\boldsymbol{\beta},\sigma^2,\mathbf{X})P(\boldsymbol{\beta},\sigma^2|\mathbf{X})}{P(\boldsymbol{\beta}|\sigma^2,\mathbf{y},\mathbf{X})} \end{aligned}$$

But we have all of these terms. $P(\mathbf{y}|\boldsymbol{\beta},\sigma^2,\mathbf{X})$ is the standard regression likelihood. We have just solved for the posterior distribution of $\boldsymbol{\beta}$ given σ^2 and the rest, $P(\boldsymbol{\beta}|\sigma^2,\mathbf{y},\mathbf{X})$ and we specified our prior $P(\sigma^2,\boldsymbol{\beta}) \propto \sigma^{-2}$

Distribution of σ^2 given observations $\mathbf{y}$ and $\mathbf{X}$

When we plug all of these known distributions into the

$$\begin{aligned} P(\sigma^2|\mathbf{y},\mathbf{X}) &\propto \frac{P(\mathbf{y}|\boldsymbol{\beta},\sigma^2,\mathbf{X})P(\boldsymbol{\beta},\sigma^2|\mathbf{X})}{P(\boldsymbol{\beta}|\sigma^2,\mathbf{y},\mathbf{X})} \\ &\propto \frac{\sigma^{-n}\exp(-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\frac{1}{\sigma^2}\mathbf{I}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}))\sigma^{-2}}{\sigma^{-p}\exp(-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_{\boldsymbol{\beta}})^T\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_{\boldsymbol{\beta}}))} \end{aligned}$$

which simplifies to

$$\begin{aligned} \propto \sigma^{-n+p-2}\exp(-\frac{1}{2}(& (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\frac{1}{\sigma^2}\mathbf{I}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\ - & (\boldsymbol{\beta}-\boldsymbol{\mu}_{\boldsymbol{\beta}})^T\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_{\boldsymbol{\beta}}))) \end{aligned}$$

Distribution of σ^2 given observations $\mathbf{y}$ and $\mathbf{X}$

With significant algebraic effort one can arrive at

$$P(\sigma^2|\mathbf{y},\mathbf{X}) \quad \propto \quad \sigma^{-n+p-2}\exp(-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\mu}_{\boldsymbol{\beta}})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\mu}_{\boldsymbol{\beta}}))$$

Remembering that $\boldsymbol{\mu}_{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$ we can rewrite this in a more familiar form, namely

$$P(\sigma^2|\mathbf{y},\mathbf{X}) \quad \propto \quad \sigma^{-n+p-2}\exp(-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}))$$

where the exponent is the sum of squared errors SSE.

Distribution of σ^2 given observations $\mathbf{y}$ and $\mathbf{X}$

By inspection

$$P(\sigma^2|\mathbf{y},\mathbf{X}) \quad \propto \quad \sigma^{-n+p-2}\exp(-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\hat{\beta})^T(\mathbf{y}-\mathbf{X}\hat{\beta}))$$

follows an scaled inverse χ^2 distribution

$$P(\theta) \quad \propto \quad \theta^{-(\nu/2+1)}e^{(-\nu s^2/(2\theta))}$$

where $\theta = \sigma^2 \implies \nu = n - p$ (i.e. the number of degrees of freedom is the number of observations n minus the number of free parameters in the model p and $s^2 = \frac{1}{n-p}(\mathbf{y}-\mathbf{X}\hat{\beta})^T(\mathbf{y}-\mathbf{X}\hat{\beta})$ is the standard MSE estimate of the sample variance.

Distribution of σ^2 given observations $\mathbf{y}$ and $\mathbf{X}$

Note that this result

$$\sigma^2 \sim \text{Inv}{-}\chi^2(n-p, \frac{1}{n-p}(\mathbf{y}-\mathbf{X}\hat{\beta})^T(\mathbf{y}-\mathbf{X}\hat{\beta})) \tag{1}$$

is exactly analogous to the following result from the classical estimation approach to linear regression.

From Cochran's Theorem we have

$$\frac{SSE}{\sigma^2} = \frac{(\mathbf{y}-\mathbf{X}\hat{\beta})^T(\mathbf{y}-\mathbf{X}\hat{\beta})}{\sigma^2} \sim \chi^2(n-p) \tag{2}$$

To get from (1) to (2) one can use the change of distribution formula with the change of variable $\theta^* = (\mathbf{y}-\mathbf{X}\hat{\beta})^T(\mathbf{y}-\mathbf{X}\hat{\beta})/\sigma^2$.

Distribution of output(s) given new input(s)

Last but not least we will typically be interested in prediction.

$$P(\mathbf{y}_{new}|\mathbf{X}, \mathbf{y}, \mathbf{X}_{new}) = \int\int P(\mathbf{y}_{new}|\boldsymbol{\beta}, \sigma^2)P(\sigma^2|\boldsymbol{\beta}, \mathbf{X}, \mathbf{y})P(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y})d\boldsymbol{\beta}, d\sigma^2$$

we will first assume, as usual that σ^2 is known and proceed with evaluating

$$P(\mathbf{y}_{new}|\mathbf{X}, \mathbf{y}, \mathbf{X}_{new}, \sigma^2) = \int P(\mathbf{y}_{new}|\boldsymbol{\beta}, \sigma^2)P(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y}, \sigma^2)d\boldsymbol{\beta}$$

instead.

Distribution of output(s) given new input(s)

We know the form of each of these expressions, the likelihood is normal as is the distribution of β given the rest

$$P(\mathbf{y}_{new}|\mathbf{X}, \mathbf{y}, \mathbf{X}_{new}, \sigma^2) = \int P(\mathbf{y}_{new}|\boldsymbol{\beta}, \sigma^2) P(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y}, \sigma^2) d\boldsymbol{\beta}$$

In other words

$$P(\mathbf{y}_{new}|\mathbf{X}, \mathbf{y}, \mathbf{X}_{new}, \sigma^2) = \int \mathrm{N}(\mathbf{y}_{new}|\mathbf{X}_{new}\hat{\boldsymbol{\beta}}, \sigma^2)\, \mathrm{N}(\beta|\hat{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}) d\boldsymbol{\beta}$$

Bayes Rule for Gaussians

To solve this integral we will use Bayes' rule for Gaussians (taken from Bishop).

If

$$\begin{aligned} P(x) &= N(x|\mu, \Lambda^{-1}) \\ P(y|x) &= N(y|Ax + b, L^{-1}) \end{aligned}$$

where x, y, and μ are all vectors and Λ and L are (invertable) matrices of the appropriate size then

$$\begin{aligned} P(y) &= N(y|A\mu + b, \text{Ł}^{-1} + A\Lambda^{-1}A^T) \\ P(x|y) &= N(x|\Sigma(A^T L(y - b) + \Lambda\mu), \Sigma) \end{aligned}$$

where $\Sigma = (\Lambda + A^T LA)^{-1}$

Distribution of output(s) given new input(s)

Since this integral is just an application of Bayes rule for Gaussians we can directly write down the solution

$$
\begin{aligned}
& P(\mathbf{y}_{new}|\mathbf{X}, \mathbf{y}, \mathbf{X}_{new}, \sigma^2) \\
&= \int \mathrm{N}(\mathbf{y}_{new}|\mathbf{X}_{new}\hat{\boldsymbol{\beta}}, \sigma^2)\, \mathrm{N}(\beta|\hat{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}) d\boldsymbol{\beta} \\
&= N(\mathbf{y}_{new}|\mathbf{X}_{new}\hat{\boldsymbol{\beta}}, \sigma^2(\mathbf{I} + \mathbf{X}_{new}\mathbf{V}_{\boldsymbol{\beta}}\mathbf{X}_{new}^T)
\end{aligned}
$$

where $\mathbf{V}_{\boldsymbol{\beta}} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}}/\sigma^2 = (\mathbf{X}^T\mathbf{X})^{-1}$

Distribution of output(s) given new input(s)

This solution

$$\begin{aligned} & P(\mathbf{y}_{new}|\mathbf{X}, \mathbf{y}, \mathbf{X}_{new}, \sigma^2) \\ = \quad & N(\mathbf{y}_{new}|\mathbf{X}_{new}\hat{\boldsymbol{\beta}}, \sigma^2(\mathbf{I} + \mathbf{X}_{new}\mathbf{V}_{\boldsymbol{\beta}}\mathbf{X}_{new}^T) \end{aligned}$$

where $\mathbf{V}_{\boldsymbol{\beta}} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}}/\sigma^2 = (\mathbf{X}^T\mathbf{X})^{-1}$

relies upon σ^2 being known. Our final inference objective is to come up with

$$\begin{aligned} & P(\mathbf{y}_{new}|\mathbf{X}, \mathbf{y}, \mathbf{X}_{new}) \\ = \quad & \int\int P(\mathbf{y}_{new}|\boldsymbol{\beta}, \sigma^2)P(\sigma^2|\boldsymbol{\beta}, \mathbf{X}, \mathbf{y})P(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y})d\boldsymbol{\beta}, d\sigma^2 \\ = \quad & \int P(\mathbf{y}_{new}|\mathbf{X}, \mathbf{y}, \mathbf{X}_{new}, \sigma^2)P(\sigma^2|\mathbf{X}, \mathbf{y}, \mathbf{X}_{new})d\sigma^2 \end{aligned}$$

where we have just derived the first term and the second we known is scaled inverse chi-squared.

Distribution of output(s) given new input(s)

The distributional form of

$$
\begin{aligned}
& P(\mathbf{y}_{new}|\mathbf{X}, \mathbf{y}, \mathbf{X}_{new}) \\
&= \int P(\mathbf{y}_{new}|\mathbf{X}, \mathbf{y}, \mathbf{X}_{new}, \sigma^2) P(\sigma^2|\mathbf{X}, \mathbf{y}, \mathbf{X}_{new}) d\sigma^2
\end{aligned}
$$

is a multivariate Student-t distribution with center $\mathbf{X}_{new}\hat{\boldsymbol{\beta}}$, squared scale marix $s^2(\mathbf{I} + \mathbf{X}_{new}\mathbf{V}_{\boldsymbol{\beta}}\mathbf{X}_{new}^T)$ and $n - p$ degrees of freedom (left as homework).

Again this is the same result as in classical regression analysis – the predictive distribution of a new (set of) points is Student-t when σ^2 is unknown and marginalized out.

Take home

- The Bayesian perspective brings a new analytic perspective to the classical regression setting.
- In classical regression we develop estimators and then determine their distribution under repeated sampling or measurement of the underlying population.
- In Bayesian regression we stick with the single given dataset and calculate the uncertainty in our parameter estimates arising from the fact that we have a finite dataset.
- Given a single choice of prior, namely a particular *improper prior* we see that the posterior uncertainty regarding the model parameters corresponds exactly to the classical sampling distributions for regression estimators.
- Other priors can be utilized.