STAT 4010
Exploration of Bayesian Linear Regression vs Classical Linear Regression

Jishnu Raychaudhuri

Spring 2022

Contents

1 Introduction

My interest in various types of regression models stems from being a Computer Science major and an Applied Mathematics minor here at CU Boulder. Over the course of my university education I have encountered and learned about multiple ways of classifying, predicting, and explaining data, but regression has always stood out to me as it appeared to be the most intuitive, but incredibly powerful at the same time in terms of what it allowed a researcher or data scientist to achieve.

Many different methods of regression exist, but they are of interest to me because even though they all try to achieve the same basic goal of prediction, they approach the problem using different methodologies. Some models may seem suitable for some data sets and contexts while other models might perform better in other situations. If there was a single model that was better than all other models, the other models would be obsolete and would no longer be used. However, this is not the case, suggesting that the differences in models are subtle. I'd like to investigate these subtle differences in models and figure out the situations which suit them the best.

In this paper, I aim to explore the Bayesian Linear Regression model and compare it with classical linear regression. I also plan to explore the origins of linear regression extensions such as Lasso and Ridge Regression via Bayesian Linear Regression.

2 Background

Statistical models try to determine the strength and nature of the relationship between an outcome variable and a set of predictor variables. This can be shown mathematically as follows:

$$Y = f(\mathbf{x}) + \varepsilon$$

where Y is the outcome variable, $f(\mathbf{x})$ is the systematic component which determines the relationship between the outcome and the predictor variables $\mathbf{x}$, and ε is the random component which represents the random variance in the outcome.

Simply observing the outcome is not enough to determine the relationship between it and the predictors due to the random component. Statistical models, such as regression, must be used to estimate the systematic component and thus the actual relationship.

2.1 Linear Regression

In linear regression, the systematic component is assumed to be linear, i.e. the relationship between the outcome and predictor variables can be modelled by a straight line, plane, or an equivalent higher dimensional shape depending on the number of predictors in the model. This assumption is made as it is the simplest relationship between the variables.

Mathematically, the linear regression model for a single outcome with p predictors can be expressed as follows:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

where y_i is the i^{th} outcome, x_{ij} is the j^{th} input variable for the i^{th} outcome, ε_i is random error associated with the i^{th} outcome, and β_j is the j^{th} coefficient for the j^{th} input variable. The ultimate goal is to estimate the coefficient values β_j.

The linear regression model for the all outcomes in the outcome variable can be better represented with matrices as follows:

$$\mathbf{Y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ is a vector of outcomes, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is a vector of coefficients, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$ is a vector of error terms, and $\mathbf{x} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}$ is a matrix containing p predictors for n outcomes with a column of 1s at the start — known as the design matrix.

For linear regression to accurately predict data, certain assumptions about the data must hold. These assumptions are as follows:

1. Linearity - The relationship between $\mathbf{Y}$ and $\boldsymbol{\beta}$ is linear.
2. Independence - Y_i is independent from Y_j, where $i \neq j$.
3. Homoskedasticity - The variance of all the outcomes are the same, $\mathrm{Var}(Y_i) = \sigma^2$ and $\varepsilon_i \sim N(0, \sigma^2)$.
4. Normality - The outcomes are normally distributed with means at their true values and constant variance, $Y_i \sim N(\beta_0 + \cdots + \beta_p x_{ip}, \sigma^2)$

As mentioned before, the main goal is to find estimates for $\boldsymbol{\beta}$ such that $\mathbf{x}\boldsymbol{\beta}$ is as close as possible to $\mathbf{Y}$. This is done via a process known as Least Squares Estimation (LSE). LSE tries to find values for $\boldsymbol{\beta}$ that minimises

$$\sum_{i=1}^{n} \left(Y_i - [\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}]\right)^2$$

or equivalently

$$||\mathbf{Y} - \mathbf{x}\boldsymbol{\beta}||^2$$

which is the sum of squared vertical distances between the observed points and the line of best fit, also called Residual Sum of Squares (RSS). The minimising values of $\boldsymbol{\beta}$ determined by LSE are denoted $\hat{\boldsymbol{\beta}}$.[1] $\hat{\boldsymbol{\beta}}$ is the best unbiased estimator of $\boldsymbol{\beta}$ as $E\left[\hat{\boldsymbol{\beta}}\right] = \boldsymbol{\beta}$.[2] The $\hat{\boldsymbol{\beta}}$ values are then used to predict the outcome variable. These predicted values are called $\hat{\mathbf{Y}}$.

$$\hat{\mathbf{Y}} = \mathbf{x}\hat{\boldsymbol{\beta}}$$

Although this methodology produces coefficients that fit the observed data well, the purpose of a predictive model is to predict unobserved outcomes using newly measured input variables. One of the main problems that linear regression runs into is the problem of over-fitting. Over-fitting is when a model accurately predicts the outcomes for data it has been 'trained' on but is not very accurate when it comes to predicting new outcomes.

To counter the problem of over-fitting to the observed outcomes, the linear regression model can be extended using regularisation methods such as Lasso and Ridge Regression.

2.1.1 Lasso Regression

Lasso is an acronym that stands for Least Absolute Shrinkage and Selection Operator.[1]It's a regularisation method that penalises the Least Square Estimator by adding a penalty term equal to the sum of absolute values of the coefficients.[1] Because of this, Lasso regularisation is also known as L_1 regularisation as the penalty term is equal to the L_1 norm[3] of $\boldsymbol{\beta}$.

Under Lasso regression, LSE must now do

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[\sum_{i=1}^{n} \left(Y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|\right]$$

[1] For derivation, see Appendix B.1

[2] For proof, see Appendix B.2

[3] For definition, see Appendix B.3

where $\mathbf{x}_i$ is the i^{th} row of the input matrix $\mathbf{x}$, and λ is the tuning parameter. λ determines just how much an effect the penalty term has on the LSE. When $\lambda = 0$ the estimated coefficients $\hat{\boldsymbol{\beta}}$ are the same as in linear regression. As λ increases, the coefficients become less sensitive to the observed data and eventually collapse towards zero. Meaning the higher the tuning parameter, the lower the number of non-zero coefficients.

2.1.2 Ridge Regression

Like Lasso regression Ridge Regression is also a regularisation method that penalises the Least Square Estimator by adding a penalty term. However, in this case the penalty term is equal to the sum of the squares of the coefficients.[2] Because of this, Ridge regularisation is also known as L_2 regularisation as the penalty term is equal to square of the L_2 norm[4] of $\boldsymbol{\beta}$.

Under Ridge regression, LSE must now do

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[\sum_{i=1}^{n} \left(Y_i - \mathbf{x}_i^{\text{T}} \boldsymbol{\beta} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right]$$

where λ is once again the tuning parameter and determines how much an effect the penalty term has on the LSE. However, unlike Lasso regression, increasing the tuning parameter doesn't result in the coefficients collapsing to zero. L_2 regularisation causes the coefficients to approach zero much more smoothly but never makes any of them equal to zero.

2.2 Bayesian Linear Regression

The regular linear regression model is a frequentist approach to regression. The outcome variables are treated as point estimates and a relationship is then deduced from these estimates by minimising the RSS value. The result is a single set of estimated coefficients that rely only on the set of outcome variables for which the LSE minimises the RSS. The resulting outputs, $\hat{\mathbf{Y}}$, are also point estimates. Bayesian linear regression (BLR), on the other hand, provides a probability distribution for the coefficients as well as the predicted outcome variables in the form of a posterior predictive distribution.[3]

Originating from Bayes Theorem,[5] the regression equation under BLR looks slightly different when compared to regular linear regression. Instead of having an additive model where the predictors are a linear combination of the outcome variable, the main goal is to compute values for $\boldsymbol{\beta}$ that maximise the posterior distribution with respect to $\boldsymbol{\beta}$ which is given as

$$P(\boldsymbol{\beta}|\mathbf{x}, \mathbf{Y}) = \frac{P(\mathbf{x}, \mathbf{Y}|\boldsymbol{\beta})P(\boldsymbol{\beta})}{P(\mathbf{x}, \mathbf{Y})}$$

where $P(\boldsymbol{\beta}|\mathbf{x}, \mathbf{Y})$ is called the posterior, $P(\mathbf{x}, \mathbf{Y}|\boldsymbol{\beta})$ is called the likelihood, $P(\boldsymbol{\beta})$ is called the prior, and $P(\mathbf{x}, \mathbf{Y})$ is called the evidence. The posterior yields the probability of a certain set of coefficients given a set of predictors and their outcomes. The likelihood attempts to calculate the probability of the set of predictors and their outcomes given a set of coefficients. The prior computes the probability of the coefficients themselves in the absence of observing any data whatsoever, essentially asking the very philosophical question of whether a certain set of $\boldsymbol{\beta}$ values should exist in the world. The calculation of the prior distribution is very important in the BLR process.

As the evidence term, the probability of observing the current set of predictors anf their outcomes, is a rather tricky notion to try and compute, is usually dropped as it doesn't matter when trying to maximise the posterior distribution. The BLR equation to be maximised is therefore usually written as

$$P(\beta|\mathbf{x}, \mathbf{Y}) \propto P(\mathbf{x}, \mathbf{Y}|\boldsymbol{\beta})P(\boldsymbol{\beta})$$

[4] For definition, see Appendix B.4

[5] For definition, see Appendix B.5

where the posterior distribution is proportional to the product of the likelihood and the prior distribution.[3]

As the $\mathbf{x}$ matrix is constant and doesn't on $\boldsymbol{\beta}$, it can be dropped from the likelihood term. The main goal is to therefore find $\boldsymbol{\beta}$ such that

$$\underset{\boldsymbol{\beta}}{\operatorname{argmax}} P(\mathbf{Y}|\boldsymbol{\beta})P(\boldsymbol{\beta}).$$

When maximising functions that are computationally heavy, the natural log of the function is maximised instead. This is because the natural log is a monotonously increasing function and can also help make computations easier. The natural log of the maximising function above is

$$\underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left[\ln P(\mathbf{Y}|\boldsymbol{\beta}) + \ln P(\boldsymbol{\beta})\right].$$

Carrying on from the assumptions made in regular linear regression, the outcomes are normally distributed[6] with distribution $Y_i \sim N(x_i^{\mathrm{T}}\boldsymbol{\beta}, \sigma^2)$. Using this information, the log likelihood term can be expanded as follows:

$$\begin{aligned}
\ln\left[P(\mathbf{Y}|\boldsymbol{\beta})\right] &= \ln\left[\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\left(Y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2}{2\sigma^2}}\right] = \sum_{i=1}^{n} \ln\left[\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\left(Y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2}{2\sigma^2}}\right] \\
&= \sum_{i=1}^{n}\left[\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2\sigma^2}\left(Y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2\right] \\
&= \sum_{i=1}^{n}\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2
\end{aligned}$$

The first term is not dependent on $\boldsymbol{\beta}$ and is therefore a constant which can be pulled out of the maximising function. The maximising function now looks like

$$\underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2 + \ln P(\boldsymbol{\beta})\right].$$

The expansion of the prior is much more complicated. For starters, the distribution of $\boldsymbol{\beta}$ is unknown. The selection of the prior distribution is usually left up to the researcher or data scientist and the choice depends on the domain being predicted. It is a reflection of the beliefs of the researchers as to what distribution the unknown quantities should follow given the area of prediction. However, certain choices for the prior distribution gives rise to special cases.

2.2.1 Special cases of Bayesian Linear Regression

Assume the researchers chose the prior distribution to be a Laplace distribution[7] with mean 0 and diversity b. $\therefore \beta_j \sim \text{Laplace}(0, b)$. The log prior can now be expanded as follows:

$$\begin{aligned}
\ln P(\boldsymbol{\beta}) &= \ln\left[\prod_{j=1}^{n}\frac{1}{2b}e^{-\frac{|\beta_j|}{b}}\right] = \sum_{j=1}^{p}\ln\left[\frac{1}{2b}e^{-\frac{|\beta_j|}{b}}\right] \\
&= \sum_{j=1}^{p}\left[\ln\left(\frac{1}{2b}\right) - \frac{1}{b}|\beta_j|\right] \\
&= \sum_{j=1}^{p}\ln\left(\frac{1}{2b}\right) - \frac{1}{b}\sum_{j=1}^{p}|\beta_j|
\end{aligned}$$

[6] For definition, see Appendix B.6

[7] For definition, see Appendix B.7

Just like in the expansion of the log likelihood, the first term is not dependent on $\boldsymbol{\beta}$ can therefore be pulled out of the maximising function. Substituting this value for the log prior in the BLR maximising function yields the following output:

$$\begin{aligned}
\underset{\boldsymbol{\beta}}{\operatorname{argmax}}\left[\ln P(\mathbf{Y}|\boldsymbol{\beta}) + \ln P(\boldsymbol{\beta})\right] &= \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2 - \frac{1}{b}\sum_{j=1}^{p}|\beta_j|\right] \\
&= \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}\left(Y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2 + \frac{2\sigma^2}{b}\sum_{j=1}^{p}|\beta_j|\right)\right] \\
&= \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left[\sum_{i=1}^{n}\left(Y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2 + \frac{2\sigma^2}{b}\sum_{j=1}^{p}|\beta_j|\right]
\end{aligned}$$

Making the fraction $\frac{2\sigma^2}{b}$ its own parameter λ turns the optimising function into

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left[\sum_{i=1}^{n}\left(Y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right]$$

which is the exact same optimising function as the one in Lasso regression, showing that Lasso regression is a special case of BLR when the prior distribution is chosen to be a Laplacian distribution with mean 0 and diversity b.[4]

Similary, making the assumption that the researchers chose the prior distribution to be a normal distribution with mean 0 and standard deviation τ. $\therefore \beta_j \sim N(0, \tau^2)$. The log prior can now be expanded as follows:

$$\begin{aligned}
\ln P(\boldsymbol{\beta}) = \ln\left[\prod_{j=1}^{p}\frac{1}{\tau\sqrt{2\pi}}e^{-\frac{\beta_j^2}{2\tau^2}}\right] &= \sum_{j=1}^{p}\ln\left[\frac{1}{\tau\sqrt{2\pi}}e^{-\frac{\beta_j^2}{2\tau^2}}\right] \\
&= \sum_{j=1}^{p}\left[\ln\left(\frac{1}{\tau\sqrt{2\pi}}\right) - \frac{1}{2\tau^2}\beta_j^2\right] \\
&= \sum_{j=1}^{p}\ln\left(\frac{1}{\tau\sqrt{2\pi}}\right) - \frac{1}{2\tau^2}\sum_{j=1}^{p}\beta_j^2
\end{aligned}$$

Again, the first term is not dependent on $\boldsymbol{\beta}$ and can therefore be pulled out of the maximising function. Substituting this value for the log prior into the maximising function yields the following:

$$\begin{aligned}
\underset{\boldsymbol{\beta}}{\operatorname{argmax}}\left[\ln P(\mathbf{Y}|\boldsymbol{\beta}) + \ln P(\boldsymbol{\beta})\right] &= \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2 - \frac{1}{2\tau^2}\sum_{j=1}^{p}\beta_j^2\right] \\
&= \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}\left(Y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2 + \frac{\sigma^2}{\tau^2}\sum_{j=1}^{p}\beta_j^2\right)\right] \\
&= \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left[\sum_{i=1}^{n}\left(Y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2 + \frac{\sigma^2}{\tau^2}\sum_{j=1}^{p}\beta_j^2\right]
\end{aligned}$$

Making the ratio of the two variances σ^2 and τ^2 their own parameter λ turns the optimising function into

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left[\sum_{i=1}^{n}\left(Y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right]$$

which is the exact same optimising function as the one in Ridge regression, showing that Ridge regression is a special case of BLR when the prior distribution is chosen to be normally distributed with a mean at 0 and a constant variance.[5]

The fact that Lasso and Ridge regression are essentially special cases of BLR is interesting as it shows the motivation and the origin behind the two regularisation techniques. The only difference in their implementation between BLR and regular linear regression is that the final predictions for BLR are probability distributions while in linear regression they're point estimates.

3 Data

To draw comparisions between the two regression models I'll be using the a divorce data set pulled from the following url: `https://raw.githubusercontent.com/bzaharatos/-Statistical-Modeling-for-Data-Science-Applications/master/Modern%20Regression%20Analysis%20/Datasets/divusa.txt`.

The full data set contains 77 entries with 7 variables for each of those entries. These variables are the year the entry was recorded — starting from 1920 up until 1996, the number of divorces per 1000 women aged 15 and over, the unemployment rate, the female percentage of the labour force who are aged 16 and above, the number of marriages per 1000 women aged 16 and above, the number of births per 1000 women between the ages of 15 and 44, and the percentage of the population that is in the military.

3.1 Model selection

The goal is to predict the divorce rate per 1000 women aged 16 and above using the variables as predictors with both classic linear regression and BLR. However, not all the features might be relevant or statistically significant when it comes to predicting the divorce rate. To find the optimal model, a model selection criterion must be used for each of our models.

The summary output for the coefficients of a linear regression model with divorce rate as the response and all other variables as predictors yields the following:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 380.14761   99.20371   3.832 0.000274 ***
year         -0.20312    0.05333  -3.809 0.000297 ***
unemployed   -0.04933    0.05378  -0.917 0.362171
femlab        0.80793    0.11487   7.033 1.09e-09 ***
marriage      0.14977    0.02382   6.287 2.42e-08 ***
birth        -0.11695    0.01470  -7.957 2.19e-11 ***
military     -0.04276    0.01372  -3.117 0.002652 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using a significance level of 5% and examining p-values, all of the predictors appear to be statistically significant, except for the unemployment rate. However, the computed p-values might prove to be erroneous due to compounded errors in the calculation of the t-statistic.[8] To combat this, a more rigorous approach such as the Akaike Information Criterion[9] (AIC) can be used.

The best way to use the AIC would be to compare AIC values across models of different sizes. However, multiple models using 2 predictors for example are possible. The best way to determine the best model with k predictors is to model all possible models with k predictors and then choose the model with the lowest

[8] For definition, see Appendix B.8

[9] For definition, see Appendix B.9

RSS. The results from conducting this process are shown in the table below:

k	(Intercept)	year	unemployed	femlab	marriage	birth	military
1	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
2	TRUE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE
3	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE
4	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE
5	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE
6	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE

Table 1: The best models with k predictors in terms of RSS for the divorce data set

From the table, the best linear regression model with one predictor is using the percentage of female labour, the best two predictor model includes female labour and the birth rate, and so on. The AIC can now be used to compare these models to determine the optimal model for predicting the divorce rate per 1000 women aged 15 and above. The model with the lowest AIC will be chosen as the optimal model.

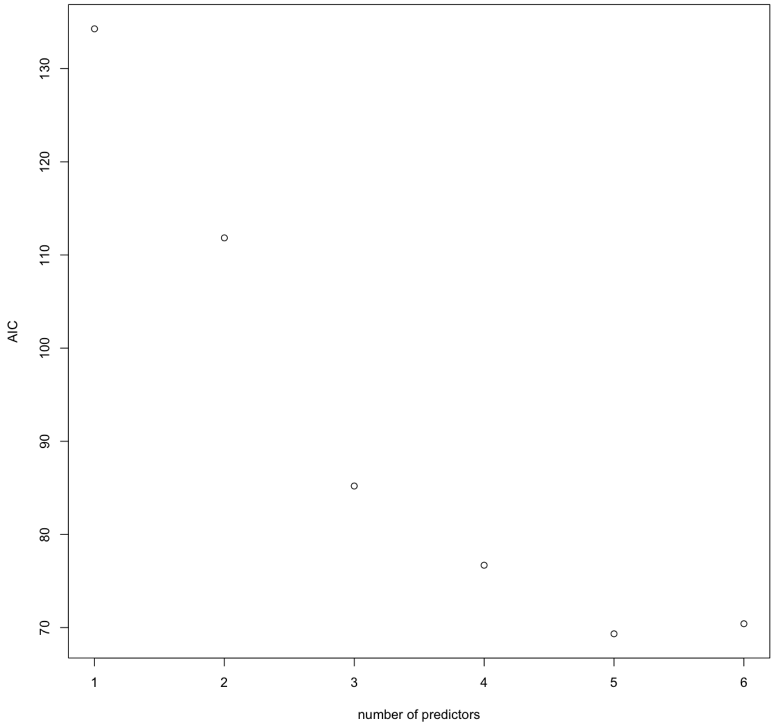

Figure 1: AIC values for the best k predictor models

According to the plot above, the model with the lowest AIC is the one with 5 predictors which, according to Table 1, is the model without the unemployment predictor. Removing the unemployment predictor from the model yields the optimal model for predicting the divorce rate using classic linear regression.

Unfortunately, there is no quick and easy way to test the significance of the predictors for BLR such as a t-test for linear regression. However, pseudo measures to test significance do exist. The Probability of Distribution metric is one such measure. The metric outputs a value between 50% and 100% which is the probability of whether a predictor is strictly positive or negative and is highly correlated to the p-value.

The probability of distribution for the full BLR model - divorce rate as the response variable and all other variables as predictors along with the assumption that the priors are normally distributed - produces the following:

```
Parameter   |     pd
--------------------
(Intercept) |   100%
year        |   100%
unemployed  | 82.50%
femlab      |   100%
marriage    |   100%
birth       |   100%
military    | 99.90%
```

Any parameter with a probability of distribution (pd) value higher than 98% can be considered to be statistically significant. From the output above, all the predictors seems to be statistically significant other than the unemployment rate. The conclusion from the probability of distribution test is the same as the conclusion from the t-tests conducted for the linear regression model.

4 Results

4.1 Linear Regression Model

The summary output of the reduced linear regression model without unemployment as a predictor is as follows:

```
Call:
lm(formula = divorce ~ . - unemployed, data = divorce)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7586 -1.0494 -0.0424  0.7201  3.3075

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 405.61670   95.13189   4.264 6.09e-05 ***
year         -0.21790    0.05078  -4.291 5.52e-05 ***
femlab        0.85480    0.10276   8.318 4.29e-12 ***
marriage      0.15934    0.02140   7.447 1.76e-10 ***
birth        -0.11012    0.01266  -8.700 8.43e-13 ***
military     -0.04120    0.01360  -3.030  0.00341 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.511 on 71 degrees of freedom
Multiple R-squared:  0.9336,Adjusted R-squared:  0.929
F-statistic: 199.7 on 5 and 71 DF,  p-value: < 2.2e-16
```

The R^2 value of the model is 0.9336, which means that 93.36% of the variation in the response variable is explained by the predictors. All predictors are statistically significant as their p-values are lower than the chosen significance level of 0.05. All this suggests that the linear regression model does a good job of estimating the relationship between the divorce rate and the year, proportion of female labour in the labour market, the marriage rate, the birthrate, and proportion of military population.

The model produces point estimates for the five coefficient values and the intercept term. According to the model and the data used to train it, the average number of divorces per 1000 women aged 15 and above is 405.61670 when all the other predictors are set to 0. A one unit increase in the year causes a 0.21790 decrease in the divorce rate on average when all the other predictors are kept constant. A one unit increase in the proportion of female labour results in average increase of 0.85480 in the divorce rate given all other predictors are help constant. A one unit increase in the marriage rate causes the divorce rate to increase by 0.15934 on average given that all other predictors are held constant. A one unit increase in birthrate causes the average divorce rate to go down by 0.11012 given all other predictors are held constant and a one unit increase in the military population reduces the average divorce rate by 0.04120 given, of course, that all other variables are held constant.

These estimated coefficient values can now be used to predict the divorce rate per 1000 women as long as values for the other predictors have been measured. The output will predict the average divorce rate for those input values as a point estimate. The predicted values can be compared to the actual observed values to gauge the predictive performance of the model.

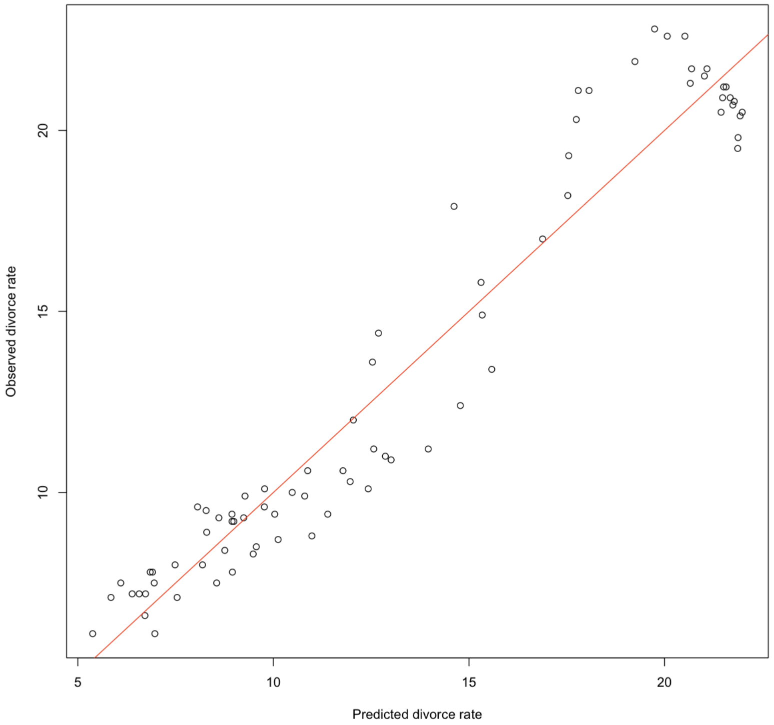

Figure 2: Observed divorce rate vs predicted divorce rate

If the model was perfect, the points would all fall on the straight line going through the origin with a slope of 1. However, the model seems to be systematically under-predicting and over-predicting the divorce rates for certain intervals. There is no way to gauge just how likely these predictions are. Prediction intervals can be calculated, but they represent a window within which a prediction can be observed with a certain certainty. Perhaps the Bayesian Linear Model can help fix this.

4.2 Bayesian Regression Model

Using the same model within a Bayesian Regression framework gives different results. Instead of point estimates, the coefficients are given as normal probability distributions, owing to the fact that the prior distribution was assumed to be normal. The estimates for the distributions of the coefficients are shown below:

```
Estimates:
```

```
              mean   sd    10%   50%   90%
(Intercept) 396.7   94.1 273.6 397.3 517.5
year         -0.2    0.1  -0.3  -0.2  -0.1
femlab        0.8    0.1   0.7   0.8   1.0
marriage      0.2    0.0   0.1   0.2   0.2
birth        -0.1    0.0  -0.1  -0.1  -0.1
military      0.0    0.0  -0.1   0.0   0.0
```

The means of the probability estimates are comparable to the point estimate values generated by classic linear regression. However, this time they are accompanied by standard deviation values which reflect a degree of uncertainty to the estimates. Plots of the coefficient estimate distributions are given below:

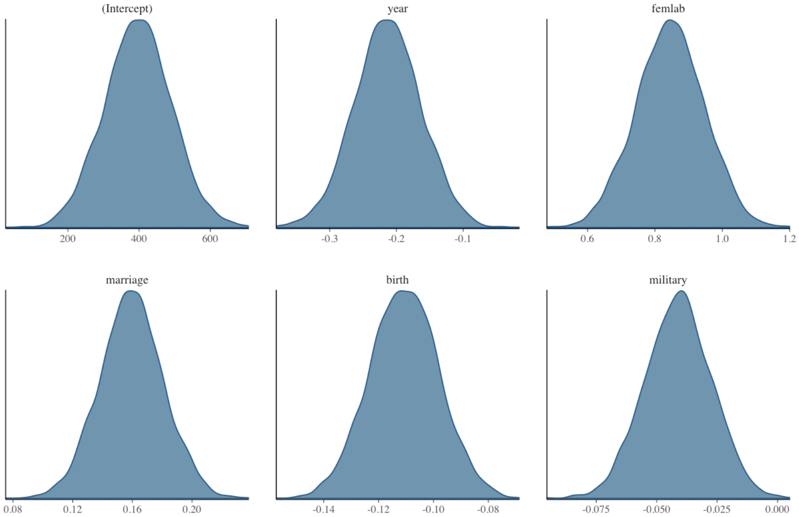

Figure 3: Probability distributions of the estimated coefficients

The probability distributions of the estimated coefficients manage to be much more descriptive of the effects of the predictor variables. A change in a variable is now no longer guaranteed to predict a certain amount of average change in the response variable. The changes are now probabilistic in nature.

The probability distributions can now be used to formulate probability distributions for the number of divorces per 1000 women. These predicted distributions can be compared to the actual observed values via a posterior predictive check.

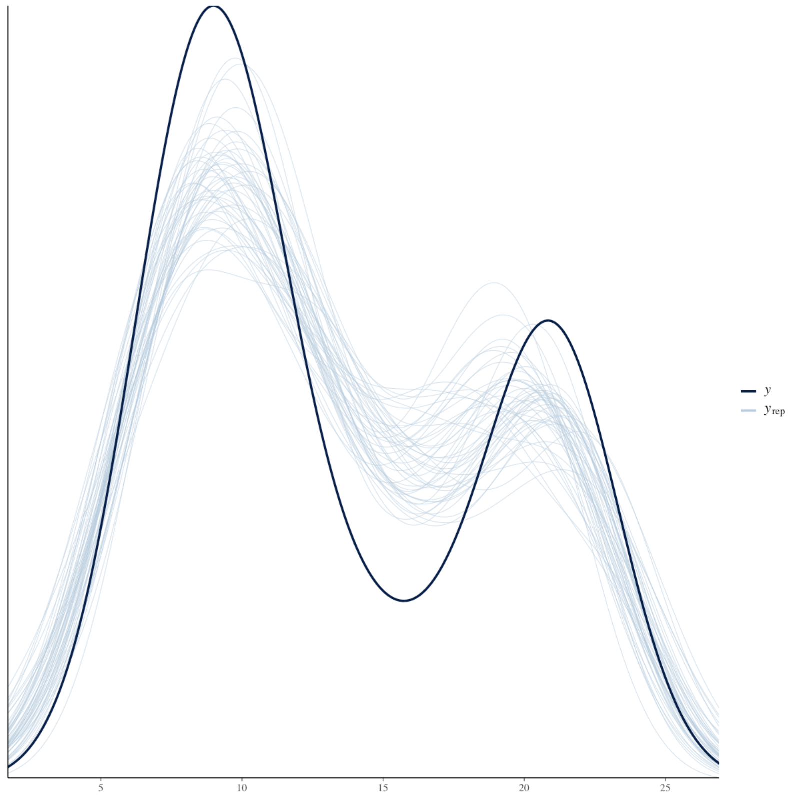

Figure 4: Observed values and multiple replicated values

The plot above showcases the distribution of the actual divorce rates along with multiple replications of the divorce rates that have been produced using the posterior predictive distribution of the coefficients, meaning the values of the coefficients used to produce the replicated values have been sampled in accordance to their predictive distribution. The distribution of the replicated values, more or less, follows the distribution of the actual observed data. This suggests that the BLR model performs reasonably well when it comes to predicting the divorce rate using the provided predictor values.

4.3 Comparisons

The fundamental difference between classic linear regression and Bayesian linear regression is the nature of the outputs of the two models. Linear regression under LSE is a frequentist approach and relies on the frequency of data points. A lot of observations are required for a linear regression model to be able to predict the response variable. The prediction is an average point estimate that provides no further information than a belief in what the average value of the response variable should be, given certain inputs.

BLR, on the other hand, outputs a probability distribution as a prediction and it doesn't need a lot of observed values either. A posterior predictive distribution can potentially be produced using only a single observation. A distribution can also be much more descriptive of an outcome rather than a point estimate for an average value as a degree on uncertainty is included in the prediction.

However, the means of the distributions can be treated as point estimates and then be compared to the point estimate predictions from LSE to facilitate direct comparisons. This data for the first 5 predictions has been gathered into a table below.

Observed	LSE prediction	BLR prediction (mean)
8.0	8.188121	8.147563
7.2	6.390035	6.322208
6.6	6.715821	6.692991
7.1	7.538527	7.540150
7.2	6.569368	6.532420

Table 2: Observed divorce rates along with LSE predicted average divorce rates and the BLR means of predicted divorce rates

The means of the BLR predicted distributions are pretty comparable to the LSE point estimate predictions. The means represent the most probable outcome from the predicted BLR distributions and appear to be slightly lower than the LSE point estimates on average.

A plot of the probability distribution of the first prediction under the BLR model is given below.

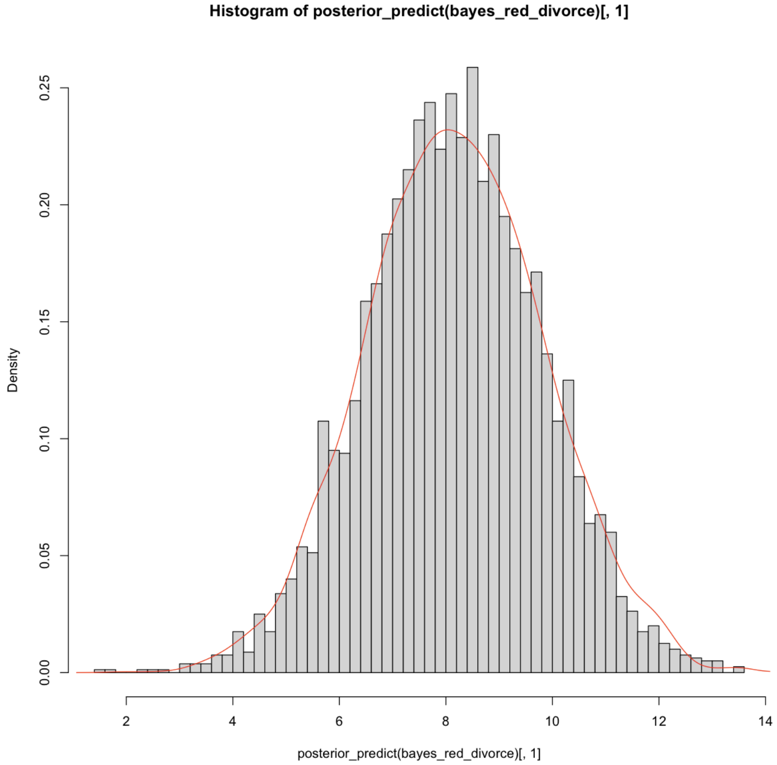

Figure 5: Normal distribution of the 1st predicted response

The distribution peaks at the mean of about 8.14 which has an approximate probability of 0.25 of occurring. The distribution output enables researchers to make more informed predictions than using point estimates. When predicting divorce rates for predictor values that are yet unobserved, the posterior predictive distribution provides a very good idea of the values that are likely to be observed. The best a point estimate can do is provide a prediction interval that is, more often than not, incredibly wide ranging and not very useful.

5 Conclusion

In conclusion, this paper has explored the use of Bayesian Linear Regression as well as Classical Liner Regression by using both models to predict divorce rates among women aged 15 and above using a range of predictors such as the year, marriage rate, birthrate, etc. Classical linear regression represents a frequentist approach to prediction and is therefore highly dependent on the observed values being used to train the model. Bayesian Linear Regression, on the other hand gives more power to the researchers in the form of choosing the prior distribution of the coefficients. This allows researchers to make decisions that are more informed, drawing information from the domain being predicted, and to also make predictions that are more informative.

In certain fields, such as Machine Learning, it is much more fruitful to state that a certain value for a response variable has a certain probability of being observed, given a set of inputs, than saying that the response variable will amount to a certain value on average. As a statistician, it is important to know of the existence of multiple approaches to prediction and the best use cases for those approaches. Classical Linear Regression and BLR are widely different approaches and should be used according to what a statistician wants to achieve and their specific case.

This exploration has been rather interesting. Learning about Bayesian linear regression made me realise just how limited my knowledge was about the variety of statistical methods out there in the world and the variety of use cases that might arise in the real world. The idea that there are vast pools of statistical knowledge that I have yet to discover is an exciting thought to have.

Finally, this exploration could have been further extended by exploring the divorce data set further during the model selection phase. The predictors could have been analysed for correlations between them, raising the question if individual predictors could truly be manipulated while keeping others the same. The significance of the predictors for Bayesian Linear Regression could also have been analysed more rigorously using methods such as Zellner's G-Prior for variable selection.[6]

References

[1] Ranstam, J, and J A Cook. "Lasso Regression." British Journal of Surgery, vol. 105, no. 10, 2018, pp. 1348–1348., https://doi.org/10.1002/bjs.10895.

[2] Saunders, C., et al. [PDF] Ridge Regression Learning Algorithm in Dual Variables: Semantic Scholar. 1 Jan. 1998, https://www.semanticscholar.org/paper/Ridge-Regression-Learning-Algorithm-in-Dual-Saunders-Gammerman/922b81f11a71aa64cda78914e6356cce89cd4f86.

[3] Baldwin, Scott A., and Michael J. Larson. "An Introduction to Using Bayesian Linear Regression with Clinical Data." Behaviour Research and Therapy, vol. 98, 2017, pp. 58–75., https://doi.org/10.1016/j.brat.2016.12.016.

[4] Park, Trevor, and George Casella. "The Bayesian Lasso." Journal of the American Statistical Association, vol. 103, no. 482, 2008, pp. 681–686., https://doi.org/10.1198/016214508000000337.

[5] van Wieringen, Wessel N. "Lecture notes on ridge regression." arXiv preprint arXiv:1509.09169 (2015).

[6] Feng Liang, Rui Paulo, German Molina, Merlise A Clyde & Jim O Berger (2008) Mixtures of g Priors for Bayesian Variable Selection, Journal of the American Statistical Association, 103:481, 410-423, DOI: 10.1198/016214507000001337

Appendix

A Code

All code used to produce plots and statistical outputs can be found here: Github.

B Definitions and Derivations

1. The Least Square Estimator, $\hat{\boldsymbol{\beta}}$, is derived by minimising the RSS. The RSS, in matrix form, is given as

$$\begin{aligned}\text{RSS} &= (\mathbf{Y}-\mathbf{x}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{Y}-\mathbf{x}\boldsymbol{\beta}) = \left(\mathbf{Y}^{\mathrm{T}}-\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}^{\mathrm{T}}\right)(\mathbf{Y}-\mathbf{x}\boldsymbol{\beta}) \\ &= \mathbf{Y}^{\mathrm{T}}\mathbf{Y}-\mathbf{Y}^{\mathrm{T}}\mathbf{x}\boldsymbol{\beta}-\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}^{\mathrm{T}}\mathbf{Y}+\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}^{\mathrm{T}}\mathbf{x}\boldsymbol{\beta} \\ &= \mathbf{Y}^{\mathrm{T}}\mathbf{Y}-2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}^{\mathrm{T}}\mathbf{Y}+\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}^{\mathrm{T}}\mathbf{x}\boldsymbol{\beta}.\end{aligned}$$

   Differentiating the RSS with respect to $\boldsymbol{\beta}$ yields the following:

$$\frac{\partial \text{RSS}}{\partial \boldsymbol{\beta}} = -2\mathbf{x}^{\mathrm{T}}\mathbf{Y}+2\mathbf{x}^{\mathrm{T}}\mathbf{x}\boldsymbol{\beta}.$$

   The Least Square Estimator can be derived by setting this derivative equal to 0 and solving for $\boldsymbol{\beta}$.

$$\begin{aligned}-2\mathbf{x}^{\mathrm{T}}\mathbf{Y}+2\mathbf{x}^{\mathrm{T}}\mathbf{x}\boldsymbol{\beta} &= 0 \\ \mathbf{x}^{\mathrm{T}}\mathbf{x}\boldsymbol{\beta} &= \mathbf{x}^{\mathrm{T}}\mathbf{Y} \\ \hat{\boldsymbol{\beta}} &= \left(\mathbf{x}^{\mathrm{T}}\mathbf{x}\right)^{-1}\mathbf{x}^{\mathrm{T}}\mathbf{Y}\end{aligned}$$

   $\therefore$ The LSE can be calculated as $\hat{\boldsymbol{\beta}} = \left(\mathbf{x}^{\mathrm{T}}\mathbf{x}\right)^{-1}\mathbf{x}^{\mathrm{T}}\mathbf{Y}$ when $\left(\mathbf{x}^{\mathrm{T}}\mathbf{x}\right)^{-1}$ exists.

2. The LSE, $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$ is $E\left[\hat{\boldsymbol{\beta}}\right] = \boldsymbol{\beta}$.

$$\begin{aligned}E\left[\hat{\boldsymbol{\beta}}\right] &= E\left[(\mathbf{x}^{\mathrm{T}}\mathbf{x})^{-1}\mathbf{x}^{\mathrm{T}}\mathbf{Y}\right] = (\mathbf{x}^{\mathrm{T}}\mathbf{x})^{-1}\mathbf{x}^{\mathrm{T}}E\left[\mathbf{Y}\right] && (E\left[kX\right] = kE\left[X\right] \text{ where } k \text{ is a constant}) \\ &= (\mathbf{x}^{\mathrm{T}}\mathbf{x})^{-1}\mathbf{x}^{\mathrm{T}}\mathbf{x}\boldsymbol{\beta} && (\text{From the normal distribution of } \mathbf{Y}) \\ &= \boldsymbol{\beta}\end{aligned}$$

   $\therefore E\left[\hat{\boldsymbol{\beta}}\right] = \boldsymbol{\beta}$.

3. The L$_1$ norm of a vector, $\mathbf{v}$, with n elements is defined as the sum of its magnitudes in space. Mathematically, this is represented as

$$||\mathbf{v}||_1 = \sum_{i=1}^{n} |v_i|.$$

4. The L$_2$ norm of a vector, $\mathbf{v}$, with n elements is defined as the distance between the vector coordinates and the origin of the vector space. Mathematically, this is represented as

$$||\mathbf{v}||_2 = \sqrt{\sum_{i=1}^{n} v_i^2}.$$

5. Bayes' Theorem, named after Thomas Bayes, gives the probability of an event occurring given another event has already occurred. Mathematically, Bayes' Theorem is presented as

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

   where $P(A|B)$ is the probability of event A occurring given that event B has occurred, $(PB|A)$ is the probability of event B occurring given that event A has occurred, $P(A)$ is the probability of event A occurring, and $P(B)$ is the probability of event B occurring.

6. The normal distribution is a probability distribution with parameters mean (μ) and standard deviation (σ). The probability density function for the normal distribution is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

7. The Laplace distribution is a probability distribution with parameters mean (μ) and diversity (b). The probability density function for the Laplace distribution is

$$f(x) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}.$$

8. Assume that $\hat{\beta}_j$ is an estimator for β_j in a statistical model and the following hypotheses need to be tested

$$H_0 : \hat{\beta}_j = c \quad \text{vs} \quad H_a : \hat{\beta}_j \neq c$$

where $c \in \mathbb{R}$. Then the t statistic for the above hypothesis test is given as

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j - c}{\text{se}(\hat{\beta}_j)}$$

where $\text{se}(\hat{\beta}_j)$ is the standard error of $\hat{\beta}_j$.

9. The Akaike Information Criterion is a prediction error metric for statistical models. It is defined as

$$\text{AIC} = 2(p+1) - 2\ln\left[L(\hat{\boldsymbol{\beta}})\right]$$

where p is the number of parameters in the model, $L(\hat{\boldsymbol{\beta}})$ is the likelihood function with the parameter estimators, $\hat{\boldsymbol{\beta}}$ as its input. For classic linear regression the AIC formula reduces to

$$\text{AIC} = 2(p+1) + n\ln\left(\frac{\text{RSS}}{n}\right)$$

where n is the number of observations.