مجلة دجلة • المجلد ٦ ،العدد ١ (٢٠٢٣)، ص: ١٠ ـ ٢٦
ISSN: 2222-6583

Comparison between Multiple Regression and Bayesian Regression with Application

Assistant professor :Dr.Samira Muhamad Salh
University of sulaimani
Department of statistic & information
Samira.muhamad@univsul.edu.iq

Abstract:

This research performs linear regression modeling with Bayesian approach. There are many kinds of regression analysis. The most repeatedly manipulated by researchers is the Frequentist methods that are often utilized in linear regression are Ordinary Least Square (OLS) and Maximum Likelihood Estimation (MLE), it has dissimilarity among two types of regression Frequentist and Bayesian in expression of looking at a parameter, along with the Bayesian development, assorted studies have manifested better modeling outcomes than the Frequentist method. After Appling two regression methods, we conclude this result: Three variables (affected the numbers of eggs in infertility, or parameters of the three independent variables are significant, the value of explain these variables is 0.698 .The MSE are equal in two methods. We conclude that the Bayesian regression method is best model for this data. As well as Bayesian approach can be used as an alternative method for the model. The comparison between Bayesian and Frequentist (OLS) modeling results have shown by using several criteria's such as RMSE, MAPE , AIC,BIC, ,MAD, and R^2. The results showed that the linear regression method by using Bayesian approach is better than Frequentist method using OLS.

Keywords: linear regression, Bayesian regression, Ordinary least square (OLS), prior distribution, posterior distribution.

الملخص :

قمنا من خلال هذا البحث بتطبيق نمذجة الانحدارالخطي باستخدام نهج بايزي.هناك عدة انواع من تحليل الانحدار،أكثرالطرق التي يتم استخدامها بشكل متكررمن قبل الباحثين هي الطرق المتكررة او (الطريقة الكلاسيكية) وهي المربعات الصغرى العادية (OLS) وتقدير الامكان الاعظم (MLE) ، والتي له تباين بين نوعي الإنحدار التكراري والبايزي في التعبير عن النظر إلى المعلمة إلى جانب تطوير طريقة Bayesian أظهرت دراسات متنوعة بأن نتائج البيزي أفضل من الطريقة الكلاسيكية و المتكررة. بعد تطبيق طريقتي الإنحدار توصلنا إلى ما يلي اثرت أربعة متغيرات على أعداد البويضات في العقم ، و أن قيمة تفسيرية لمعامل الاختلاف لمتغيرات الاربعة مستقلة معنوية هي ٠.٠٦٩٨ و ظهرت في النتائج أن طريقة الانحدار البايزي هي أفضل نموذج لهذه البيانات و كما يمكن استخدام نموذج بايزي كطريقة بديلة.وأيضا أظهرت المقارنة بين نتائج النموذج البايزي والمتكررة

مجلة دجلة • المجلد ٦ ،العدد ١ (٢٠٢٣)، ص: ١٠ ـ ٢٦
ISSN: 2222-6583

(OLS) باستخدام عدة معايير مثل (RMSE، MAPE، AIC، BIC، MAD، R^2) أن طريقة الانحدار الخطي باستخدام طريقة بايزي أفضل من طريقة التكرار باستخدام OLS.

الكلمات الافتتاحية:الانحدار الخطي ،الانحدار البيزي،المربعات الصفرى الاعتيادية ،التوزيع السابق ،التوزيع اللاحق.

Aim of research: In this work we can compare between two methods of regression which is (OLS Ordinary least square and Bayesian regression) or the foremost objectives of this study includes:

(1) A linear regression model is built by using Ordinary Least Square (OLS) method.
(2) A linear regression model is built by using Bayesian approach.
(3) To compare between two method and discover the best linear model.

Introduction:

In statistic one of the most habitually method used is linear regression analysis to analyzing data. There are two types of attributes in this method, first known as dependent variable which is variables that effected or the values that depend on other variables, and second variable is independent or explanatory variables which impact or affected to dependent variable. In regression analysis some response variables can only be one variable and the data scale is interval / ratio so that data can only be in the form of numerical data [22]. While the independent variable can be more than one variable and the data scale can be (nominal / ordinal) meaning that categorical data or (interval / ratio) of numerical data. If only one independent variable is used then the model is called simple linear regression analysis, whereas if the number of variables are more than one then it is called multiple linear regression analysis [30]. There are three goals in linear regression analysis to know the relationship among the response variable and the independent variables, to test whether the independent variables have effect over the dependent variable, and to predict the value of variable dependent based on independent variables that have been determined. Formation of the regression model is done by estimating the parameters of the regression model. So that will yield regression coefficient for each independent variable [22]. The most frequent method used by researchers is the Frequentist or Classic method by using OLS (Ordinary Least Square) or MLE (Maximum Likelihood Estimation) [3]. First method is also called the least squares method, done by minimizing the number of errors from the regression equation. The parameters are obtained from minimizing the error function of the equation [21]. While MLE is done by maximizing the probability density function (probability density function) of a data. So by using both of the methods, there are classical assumptions that must be met from the results of regression modeling. Some assumptions that must be met are error independent, identical and normal distribution [23] .In addition to these two methods, there are other methods that can be used to estimate the parameters of regression model which is Bayesian approach. The difference between Frequentist and Bayesian methods is the point of view of

the parameters. The Bayesian approach view's the parameters as a random variable so that the value is not a single value as in the Frequentist point of view [1] .

Relate work or literature Reviews:

There are many study have been carried out in Bayesian method such as:

1. Edward R. ,et al (2010): The main target of this study is to improve a Bayesian "sum-of-trees" model where each tree is constrained by a regularization prior to be a weak learner, and fitting and inferenceare accomplished via an iterative Bayesian backfitting MCMC algorithmthat generates samples from a posterior .
2. Brodersen, et al (2015): The main aim of this research is to proposes to infer causal impact on the funamental point of a diffusion-regression state-space model that predicts the counter factual market response in a synthetic control that would have occurred had no intervention taken place .
3. In (2018) ,Qian, H.: published the research the aim of this research is to establish the normal-inverse-gamma summation operator, which combines Bayesian regression results from different data sources and conduct to a simple split-and-merge algorithm for big data regressions.
4. Hubin, A.,ET AL in (2018):worked in Bayesian model the main objective of this study is to will adapt an advanced developmental algorithm called GMJMCMC (Genetically modified Mode Jumping Markov Chain Monte Carlo) to execute Bayesian model selection in the space of logic regression models.
5. Neelon, B. (2018): The main intention of this study is to propose an efficient Bayesian approach to fitting ZINB models. Moreover, the proposed data-augmented Gibbs sampler makes use of easily sampled Polya-Gamma random variables; conditional on these latent variables, inference proceeds via straightforward Bayesian inference for linear models.

1. Linear Regression Model:

One of the customary statistical technique that used to elucidate the relationship among (independent) explanatory variables and the (response) dependent variable is Regression analysis. Therefore, the necessity of this analysis is selecting a pertinent model for the available data is. It has been used in many areas of applications, including dental health, medical research, and economics [23]. Regression analysis includes several variations, such as linear, multiple linear, and nonlinear. There are two descriptions of linearity one of them is linearity for the parameter and the other is linearity for the variable. Linear model is called modeling if the model is linear for the parameter. The model is called linear model as long as the method is linear to the parameter even if the model is not linear to the attribute or linear to the attribute [31],[13]. Multiple linear regression models can be utilized to assess the relationship between the dependent variable with two or more independent variables. Before doing linear regression modeling, first it must be ensured that there is a relationship or correlation between each explanatory variable to the response variable and the relationship between the outcome variable with all independent variables are linear. Simple linear and multiple linear are the most familiar models. Nonlinear regression analysis is regularly used for more convoluted data sets in which the response and explanatory variables show a nonlinear relationship [24]. We start with multiple linear regression, the "simplest" of the

regression analyses. It is defined to be the least squares estimator of a regression model with two or more than two explanatory (independent) variable in which a straight line is fit through (n) points so that the sum of squared residuals (SSR), The goal is predicting the unknown parameters by minimize ($\sum \varepsilon i^2$). The intention is to select the parameters that the residuals are "small." While there are various methods with which to minimize the residuals [11]. The model can be thought of as a similar to the slope-intercept form of a line with an error (residual) term. The simple linear regression model is given by:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon_i \quad \ldots (1)$$

Where: i= 1, 2, ..., n, yi is the response (dependent) variable, Xi is the explanatory /predictor (independent) variables, and $(\beta_0, \beta_{i)}$ are the unknown parameters that are to be estimated, and (ε_i) is the residual term[30] . the, term (ε_i) is independent and identically distributed (i.i.d) with a normal distribution with (0) mean and unknown (σ^2) variance[12],[29] the multiple linear regression reduces to simple linear regression and it is a generalization of the simple linear regression. by using matrix notation the equation can be writing as :

$$Y = X'\beta + \varepsilon \quad \ldots (2)$$

The estimation of the parameters are defined $\beta_0, \beta_1, .., +\beta_n$ as ($\beta_0, \beta_1, .., +\beta_n$) .
Where minimize $\varepsilon = Y - X'\beta$), Or $\varepsilon_i = y_i - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_{n)}$, the least square estimates is : $\beta = (X'X)^{-1} X'Y$

Where minimize the ($\sum_{i=1}^{n} \varepsilon_i^2$) is to find a line that "best" fits the data. To do this we use an estimated regression line [1] .

2. Bayesian Linear Regression:

Before we introduce Bayesian Regression Analysis, it is predominant to first step Bayes' Theorem, the idea behind Bayesian Analysis. Bayes' Theorem is also known as Bayes' rule. Bayes' rule was introduced by Thomas Bayes as a means for calculating Conditional probabilities. For two events, A and B,

$$P(AlB) = \frac{P(B/A)P(A)}{P(B)} \quad \ldots..(3)$$

Where $P(B) \neq 0$. Even though the inference methods we present in this work are likelihood-based (also called the frequentist's view)[15],[28], Bayesian methods are widely used and accepted in practice. The concept of Bayes model is to compute posteriors based on specified priors and the likelihood function of data. It essential researchers to appropriately identify priors given inappropriate priors may escort to biased estimates or make computation of posteriors arduous. In this section, we will briefly go through several most common priors, how posteriors are calculated based on priors, and how the selection of priors influence posterior computation. In a Bayesian analysis, we start with our model, which is the distribution of the observed data conditioned on its parameters. This is also termed the

likelihood and, for most practical purposes, can simply be viewed as equivalent to the traditional likelihood function. In order to update our model, we define a distribution for the parameter(s) in this model, which is based on previous beliefs. This distribution is called a prior distribution and the parameters in this distribution are called hyper parameters. We multiply the likelihood function of our model by the prior distribution also then divide by the marginal density function, which is the joint density function with the parameter integrated out. The result is called the posterior distribution. Fortunately, the marginal density function is just a normalizing constant and does not usually have to be calculated in practice. We refer to the additional information obtained as the likelihood and marginal likelihood denoted by P(B I A) and P(B), respectively[15]. Thus, Bayes' rule can be written as:

$$posterior = \frac{likelihood \times prior}{m\arg inal \;\; likelihood} \qquad \textbf{...(4)}$$

In the Bayesian approach there is a prior, likelihood distribution and posterior distribution. Parameter estimation by Bayesian approach is done by processing posterior distribution which multiplies the prior distribution with likelihood. In linear regression model using OLS estimation method, there is normal distributed error assumption that is $\varepsilon \sim N(0,\sigma^2)$. Since the error is normally distributed, the variables $(Y/X,\beta,\sigma^2)$ are also normally distributed. Thus the variables $(Y/X,\beta,\sigma^2) \sim N(X\beta,\sigma^2)$ and probability density function (pdf) of these variables are as follows [18],[3].

$$P(Y/X,\beta,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} EXP\left[-\frac{1}{2}(Y-X\beta)'(Y-X\beta)\right] \qquad \ldots(5)$$

Based on the probability density function (pdf) above can be defined likelihood function of these variables are as follows [26],[4]:

$$P(Y/X,\beta,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} EXP\left[-\frac{1}{2}(Y-X\beta)'(Y-X\beta)\right]$$

$$P(Y/X,\beta,\sigma^2) = (\sigma^2)^{\frac{-n}{2}} EXP\left[-\frac{1}{2}(Y-X\beta)'(Y-X\beta)\right]$$

$$P(Y/X,\beta,\sigma^2) \propto (\sigma^2)^{\frac{-\nu}{2}} EXP\left[-\frac{\nu S^2}{2\sigma^2}\right] \times (\sigma^2)^{\frac{-n}{2}} EXP\left[-\frac{1}{2}(Y-X\beta)'(Y-X\beta)\right] \quad \ldots.(6)$$

There are several prior distributions that can be operate in the Bayesian approach of linear regression model, one of which is the distribution of prior conjugate [8],[23]. Estimation of regression model parameters with Bayesian approach can be done by iterating at marginal

posterior. Posterior distribution is calculated by multiplying the prior distribution and likelihood function [3],[10],[20].

$$Posterior \propto Likelihood \times Prior$$

$$p(\beta,\sigma^2/Y,X) \propto P(Y/X,\beta,\sigma^2)P(\sigma^2)P(\beta/\sigma^2)$$

$$p(\beta,\sigma^2/Y,X) \propto (\sigma^2)^{\frac{-n}{2}} \times EXP\left[-\frac{1}{2\sigma^2}(Y-X\beta)'(Y-X\beta)\right]$$

$$\times(\sigma^2)^{-(\frac{\nu}{2}+1)} EXP\left[-\frac{\nu S^2}{2\sigma^2}\right]\times(\sigma^2)^{\frac{-k}{2}} \times EXP\left[-\frac{1}{2\sigma^2}(\beta-\mu)'\Lambda(\beta-\mu)\right]$$

…(7)

The process of obtaining the estimation of regression model parameters with Bayesian approach can be done by using MCMC (Markov Chain Monte Carlo) algorithm [9],[21] [23],.

3. Priors:

Statistical models by utilizing Bayesian statistics vital the formulation of a set of prior distributions for any unknown parameters. The probability distributions express one's uncertainty about an unknown quantity, p, before the "data" is taken into account. It is indicated to the attribute's uncertainty rather than randomness to the uncertain quantity. The unknown quantity may be a parameter or latent attribute. It is often the purely subjective assessment of an experienced specialist. The most controversial element of Bayesian Analysis is utilize of a prior distribution. Criticized for introducing subjective information, the use of a prior is purely an educated guess and can vary from one scientist to another. There are two types of priors. The first type, called the Conjugate Prior, occurs when the posterior distribution has the same form as the prior distribution. The second type, called the Non informative Prior, is used when we have very little knowledge or information about the prior distribution. The non-informative prior is used to "conform to the Bayesian model in a correct parametric form." [3],[30].

Choose conjugate prior for (β,σ^2) to ensure an analytic expression for the posterior distribution. With both (β) and (σ^2) unknown, the conjugate prior is specified as $p(\beta,\sigma^2) = p(\beta/\sigma^2)p(\sigma^2)$ where:

$$\beta/\sigma^2 \sim N(\bar{\beta},\sigma^2 A^{-1})$$

$$\sigma^2 \sim \frac{\nu_0 S_0^2}{\chi^2_{\nu 0}}$$

[Hence β/σ^2 has a normal prior and σ^2 has selected inverse chi-square prior.

$$p(\beta,\sigma^2/y,X)=p(y/X,\beta/\sigma^2)p(\beta/\sigma^2)P(\sigma^2)$$
$$\propto(\sigma^2)^{\frac{-n}{2}}\exp\left\{\frac{-1}{2}(y-X\beta)'(y-X\beta)\right\}\times(\sigma^2)^{\frac{k}{2}}\exp\left\{\frac{-1}{2\sigma^2}(\beta-\bar{\beta})'(\beta-\bar{\beta})\right\}$$
$$\times(\sigma^2)^{-(\frac{v_0}{2}+1)}\exp\left\{-\frac{v_0S_0^2}{2\sigma^2}\right\} \qquad (8)$$

And the joint posterior then takes the form: If the posterior distributions p(θ|x) are in the same family as the prior probability distribution p(θ), conjugate distributions are called for both of the prior and posterior ,also the prior is called a conjugate prior for the likelihood[16],[26].

Below for some likelihood functions, if you choose a certain prior, the posterior ends up being in the same distribution as the prior. Such a prior then is called a Conjugate Prior. The conjugate prior simplifies the mathematics involved in Bayesian data analysis [22],[26].Some commonly used conjugate priors are listed in table (1).

Table (1) conjugate priors

Prior Distribution	Likelihood	Posterior distribution
Beta	Binomial	Beta
Gamma	Poisson	Gamma
Gamma	Exponential	Gamma
Normal	Normal	Normal

4. Posterior probability distribution:

The posterior probability of a random event or an uncertain proposition is the conditional probability that is assigned after the relevant evidence is taken into account [5],[7]. The posterior probability distribution of one random variable given the value of another can be calculated with Bayes' theorem by multiplying the prior probability distribution by the likelihood function, and then dividing by the normalizing constant, as follows[25],[27]:

$$f\,{}^{X}\!/_{Y}=y(x)=\frac{f_X(x)L\,X/Y=y(x)}{\int_{-\infty}^{\infty}f_X(x)L\,X/Y=y(x)dx}$$

5. Model selection criteria:

In this work has many criteria are used to comparison between models such as (R^2, AIC, BIC, RMSE, MAPE, MAD and MSE).and it is calculated according to the following formula [6],[20]:

criteria		formula	Descriptions
1	R^2	$R^2=\frac{MSE}{MST}$	MSE :Mean square error MST :Mean square total
2	AIC	$AIC=-2\log L+2(k+p)$	Akaike Information Criterion
3	BIC	$AIC=-2\log L+2\log n(k+p)$	Bayesian Information Criterion

مجلة دجلة • المجلد ٦ ،العدد ١ (٢٠٢٣)، ص: ١٠ - ٢٦
ISSN: 2222-6583

4	MSE	$MSE = \frac{SSE}{d.f}$	MSE :Mean square error
5	RMSE	$RMSE = \sqrt{MSE}$	Root square of MSE :Mean square error
6	MAD	$MAD = \frac{\sum_{i=1}^{n} \lvert x_i - \bar{x} \rvert}{n}$	Mean absolute division
7	MAPE	$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left\lvert \frac{x_i - \hat{x}}{x_i} \right\rvert$	Mean Absolute Percentage Error (MAPE)

2. Application part:

In this section we discusse practical part and all results will be presented to the applied of the study by using statically package (JASP and SYSTAT12) program.

2.1- Data Description and analysis:

2.1.1: Data Description:

Infertility is a medical condition that leads to psychological, physical, mental, spiritual and medical detriments to the patient .this medical condition affecting both partners as a couple, their prospective is to achieve pregnancy in one menstrual cycle .therefore, counseling the patient on referrals and providing basic education and understanding of this medical condition is one of the basic criteria that will also assist the healthcare team [19],[26]. In this study a sample was taken from (182) women suffering from infertility, with six independent variables which are {Age, (TSH) Hormone thyroid stimulation, (LH) luteinizing hormone, (FSH) follicle stimulating hormone,(PL) Prolactin , and (P4)Progesterone , with the dependent (y=response) variable is the numbers of eggs (oocyte), The goal of this study is to determine the effect of these variables on the oocyte (y).The explanatory variables are taken in this study are the follow

X_1= Age measured by year.

X_2= TSH Hormone thyroid stimulation measured in (mlu/ml) milli-internatinal unite per milliliter.

X_3= luteinizing hormone LH expressed in units of (mlu/ml) milli-internatinal unite per milliliter.

X_4= follicle stimulating hormone FSH, measured in (mlu/ml) milli-internatinal unite per milliliter.

X_5= Prolactin PL expressed in units of (mlu/ml) milli-internatinal unite per milliliter.

X_6= Progesterone P4 expressed in units of (ng/ml).

And y= oocyte, (response) variable is the numbers of eggs.

2.1.2: Analysis and Result:

At the first by using (JASP and SYSTA13 program package) we show that the result of descriptive statistics for all variable's entered in study.

مجلة دجلة • المجلد ٦ ،العدد ١ (٢٠٢٣)، ص: ١٠ ـ ٢٦
ISSN: 2222-6583

Table (1) descriptive statistics for all variable's

Variables	Y	X_1	X_2	X_3	X4	X5	X6
mean	11.53	39.28	3.16	6.43	8.15	18.09	0.67
SD	6.47	7.10	2.32	4.17	3.56	7.91	0.43

And then to determine whether there is all independent variables are effected the depended variable or not multiple regression analysis according to the (OLS) method or (frequentist method).

Table (2) Fitting Multiple Regression Model (OLS)

variables	Estimate	Std.Error	t-value	Sig.	95% confidence interval Lower	upper	Collinearity Tolerance	VIF
Intercept	42.142	1.92	21.84	0.000	38.33	45.94		
X1	-0.573	0.038	-14.90	0.000	-0.649	-0.497	0.939	1.06
X2	0.272	0.131	2.07	0.003	0.013	0.532	0.753	1.32
X3	-0.094	0.078	-1.20	0.228	-0.248	0,060	0.663	1.50
X4	-0.838	0.089	-9.45	0.000	-1.013	-0,663	0.701	1.42
X5	-0.102	0.034	-3.03	0.003	-0.196	-0.036	0.980	1.02
X6	0.478	0.67	0.70	0.470	-0.853	1.810	0.802	1.24

we found that the coefficient of (independent variables X_1=age, X_2=TSH, X_3=LH, X_4=FSH, X_5=PL and X_6=P4) affected the dependent variable (y= oocyte) .when we compare the p-value of these variables with level of significant (α=0.01).show that the p-value of these variables (Age, FSH,PL ,and TSH) are less than (0.01) conclude that these variables affected the numbers of eggs in table (2) .

مجلة دجلة • المجلد ٦ ،العدد ١ (٢٠٢٣)، ص: ١٠ ـ ٢٦
ISSN: 2222-6583

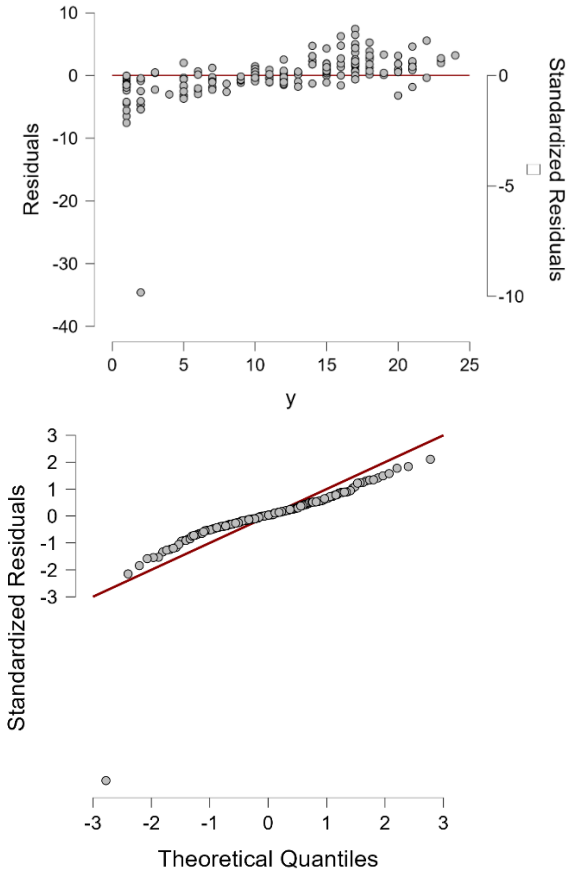
Table (3) Model Summery

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
	0.841	0.708	0.698	3.560	0.708	70.57	6	175	0.000	2.076

Table (4) Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	5365.75	6	894.29	70.57	0.000
Residual	2217.55	175	12.67		
Total	7583.30				

One of the basic assumption before analysis is the realization of the normality for residual we can improve that throw Figure (1)



Residual vs. dependent variable **Q-Q Plot standardized residuals**

2.1.3: Result and discussion for OLS method:

After applying, multiple regression with least square method (OLS) on the sample data size (n=182) for the cases under the infertility. By taking the hypothesis that the (Oocyte y) depends on the expletory variables {X_1=age, X_2=TSH, X_3=LH, X_4=FSH, X_5=PL and X_6=P4} and comparing their results, the following important points are concluded below:

1. Table (1) shows all measures of descriptive statistics. Descriptive statistics give an overview we are working with as can be seen.
2. From table (2) the four variables (Age, FSH,PL and TSH) have highly significant effects on the oocyte. In addition, the intercept has a highly significant effect. However, the two variables (LH, P4) appeared to have no significant effects. And the last column of the table represent the value of VIF which is (1.06,1.32,1.50,1.42,1.02,1.24) and this indicates that there is no correlation among (explanatory variables.
3. The output from Table (3) shows the results of fitting a multiple linear regression model to describe the relationship between Oocyte and six independent variables. Since the P-value in the ANOVA table is less than 0.01, there is a statistically significant relationship between the variables at the 95.0% confidence level. The R-Squared statistic indicates that the model as fitted explains 70% of the variability in Oocytes. The adjusted R-squared statistic, which is more suitable for comparing models with different numbers of independent variables, is 0.698%. Last Column shows that there is no independence problem because the Durbin- Watson test is equal to 2.076 which is a possible range between zero and four.
4. Table (4) shows the (ANOVA table) analysis of variance. The result represent that sum squares for regression, residual, total is equal to (5365.75, 2217.55, 7583.30) respectively, in addition, mean square for regression and residual is equal to (894.29, 12.67) and the F is (70.57).

5. Figure (1) shows the normality of residuals while all plots are on the line.

2.2: Analysis for Bayesian regression:

2.1.1: In this section Appling Bayesian regression is shown in table (5, 6) the variable contains dependent and independent variables.

Table (5) Model (Comparison – y) for Bayesian regression

Models	P(M)	P(M|data)	BF_M	BF_{10}	R^2
x1 + x4 + x5	0.016	0.428	47.066	1.000	0.698
x1 + x2 + x4 + x5	0.016	0.271	23.437	0.634	0.705
x1 + x4 + x5 + x6	0.016	0.071	4.785	0.165	0.700
x1 + x2 + x3 + x4 + x5	0.016	0.052	3.483	0.123	0.707
x1 + x4	0.016	0.048	3.178	0.112	0.681
x1 + x3 + x4 + x5	0.016	0.040	2.656	0.095	0.698
x1 + x2 + x4 + x5 + x6	0.016	0.033	2.130	0.076	0.705

مجلة دجلة • المجلد ٦ ،العدد ١ (٢٠٢٣)، ص: ١٠ ـ ٢٦
ISSN: 2222-6583

Table (5) Model (Comparison – y) for Bayesian regression

Models	P(M)	P(M|data)	BF_M	BF_{10}	R^2
x1 + x2 + x4	0.016	0.020	1.293	0.047	0.687
x1 + x3 + x4 + x5 + x6	0.016	0.008	0.522	0.019	0.700
x1 + x2 + x3 + x4 + x5 + x6	0.016	0.008	0.477	0.018	0.708

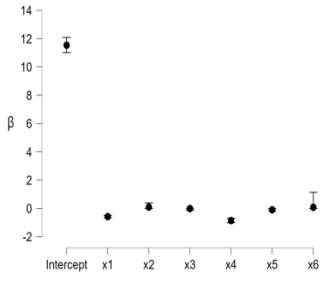
Table (6)Posterior Summaries of Coefficients

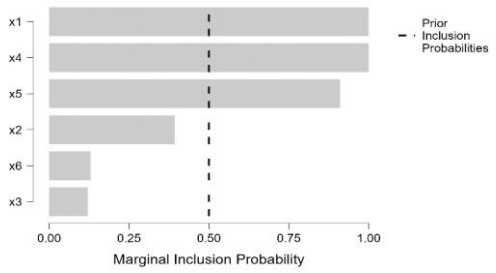
Coefficient	P(incl)	P(excl)	P(incl|data)	P(excl|data)	$BF_{inclusion}$	Mean	SD	95% Credible Interval	
								Lower	Upper
Intercept	1.000	0.000	1.000	0.000	1.000	11.533	0.266	11.008	12.086
x1	0.500	0.500	1.000	5.781e-31	1.730e+30	-0.576	0.038	-0.655	-0.502
x2	0.500	0.500	0.393	0.607	0.647	0.094	0.140	0.000	0.391
x3	0.500	0.500	0.121	0.879	0.138	-0.007	0.035	-0.118	0.013
x4	0.500	0.500	1.000	3.498e-16	2.859e+15	-0.857	0.080	-1.011	-0.691
x5	0.500	0.500	0.911	0.089	10.190	-0.096	0.044	-0.160	0.000
x6	0.500	0.500	0.129	0.871	0.149	0.079	0.317	-0.006	1.141

We can write the regression models as follow:

$Y=11.533-0.576(X_1\text{-mean } X_1)+0.094(X_2\text{-mean}X_2)-\ldots\ldots+0.079(X_6\text{-mean } X_6)$

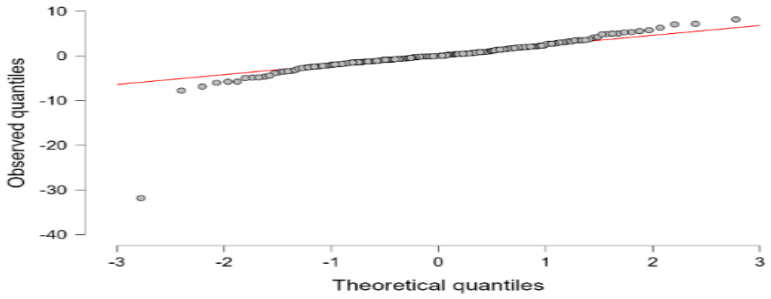
Posterior Coefficients with 95% Credible Interval **Inclusion Probabilities**



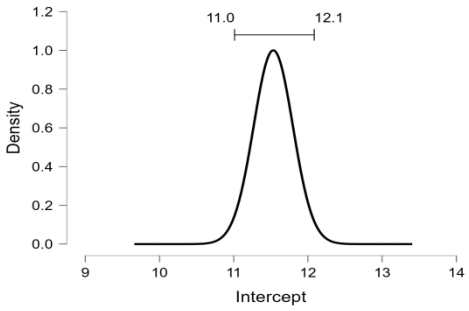


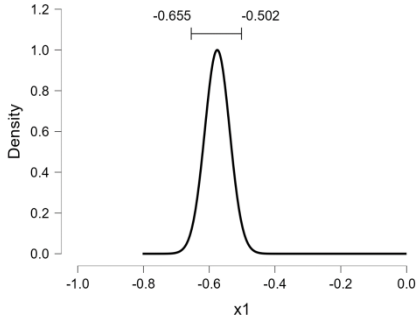
مجلة دجلة • المجلد ٦ ،العدد ١ (٢٠٢٣)، ص: ١٠ ـ ٢٦
ISSN: 2222-6583

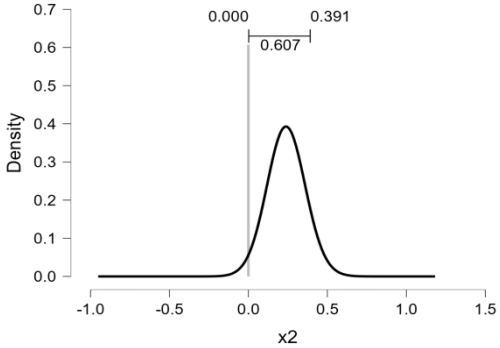
Q-Q Plot

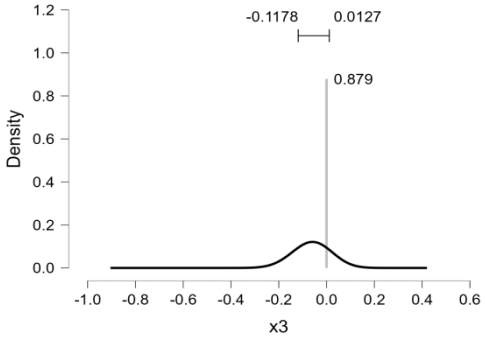


Marginal Posterior Distributions for coefficient of regression:

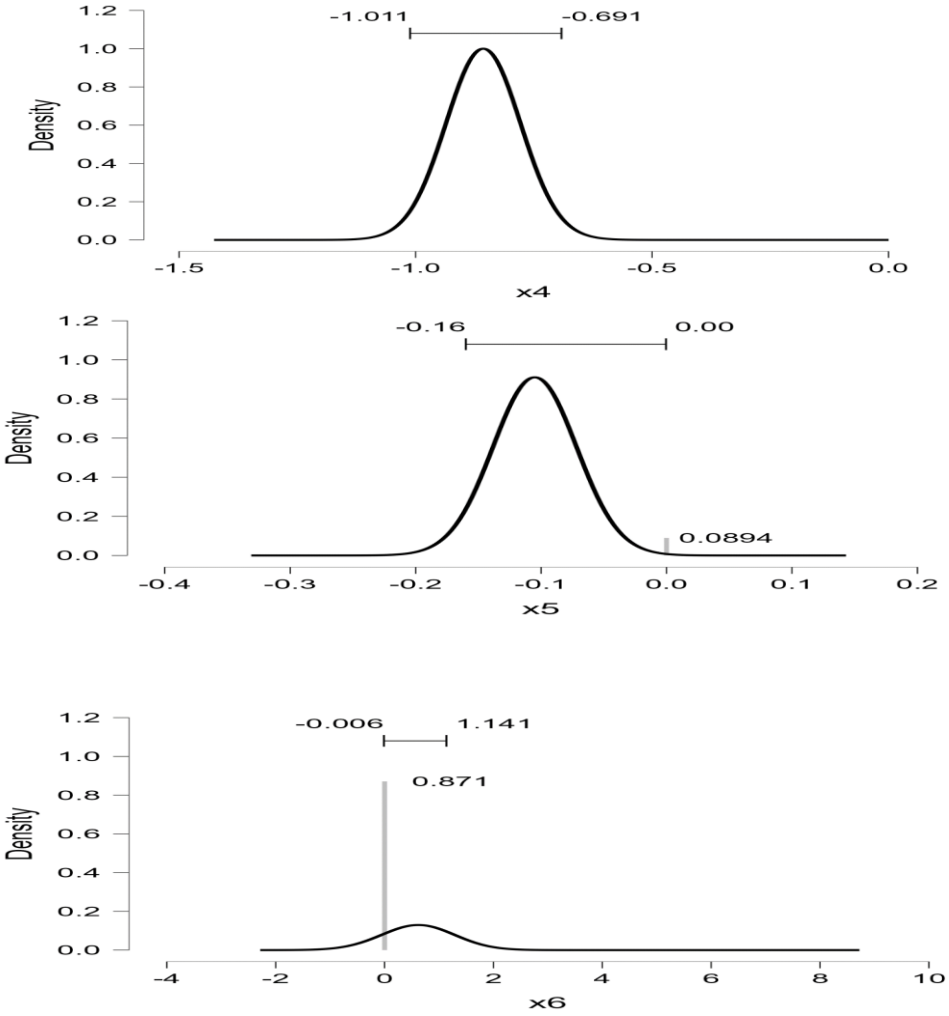








مجلة دجلة • المجلد ٦ ،العدد ١ (٢٠٢٣)، ص: ١٠ ـ ٢٦
ISSN: 2222-6583



2.1.2: Result and discussion for Bayesian regression:

After applying, multiple regression with Bayesian approach **c** on same sample data size (n=182) for the cases under the infertility. By taking the hypothesis that the (Oocyte y) depends on the expletory variables {X_1=age, X_2=TSH, X_3=LH, X_4=FSH, X_5=PL and X_6=P4} and comparing their results, the following important points are concluded below:

1) Table (5) shows that each of these variables appears in 10 models. Each of these models has its own importance. We can explain it in several points .Each of the variables (X_1, X_4 and X_5) participated in all models, so it is considered the three most important variables of the study and the main component of all models against all other variables. Then Variable (X_2= TSH) appeared in 6 models, it is clear that the variable TSH also one of the important variables affects the numbers of Oocyte.
2) In the second column P(M)=prior model probability we see that the Prior for each of these models is 0.016 .
3) In column P(M/data) posterior model probability we see that the buster of this model is (0.428) but where is the others (0.497) .

مجلة دجلة • المجلد ٦ ،العدد ١ (٢٠٢٣)، ص: ١٠ ـ ٢٦
ISSN: 2222-6583

4) In column (BF_M) Bayes factor represent the change from prior model odds to posterior model odds a best model selected is 47.066 compared with other models .
5) Also Bayes factor (BF_{10}) on the top is 1 against for each models.
6) last column it is represent the R^2 square, the value for the three variables is (0.698) it is highest for three variables while the R^2 square for the all variables is (0.708) that is, the difference between the value of all variables against three variables (x_1, x_4, x_5) only 0.01 .that is mean affect these three variables in the study is highest.
7) In Table (6) Posterior Summaries of Coefficients we see that the two parts of the result ,the first part including the Posterior and prior with ($BF_{inclusion}$) helps us to determine each possible predictor .The second part consist the coefficient and confidence interval .in this table according to ($BF_{inclusion}$) we show that the variable X5 is most important than other variables the inclusion Bayes factor is (10.190).this means that the data have increased our prior odds for including x5 as a predictor factor by (10.190) in the model .
8) The posterior mean of the coefficient is (-0.576, 0.094,-0.007,-0.857,-0.096, 0.079).
9) Last column consist of the 95% Credible Interval for each variables.
10) From the marginal Posterior Distributions for coefficient of regression show that the variables (X_3 and X_6), although they do not have a significant effect on the model, but through the probability of the model, we can say that if these variables is included in the model, the probability of this including is (0.879,0.871) there is a large spike at lower confident interval to upper bound of two variables which is (-0.118 ,0.013) and (-0.006,1.141) respectively .

3: Comparison between two methods throw the Model selection criteria.

Table (7) Model selection criteria

criteria	OLS method	Bayesian method
R^2	0.708	0.698
AIC	987.52	474.009
BIC	1013.15	486.312
MSE	12.67	12.67
RMSE	3.560	3.560
MAPE	11.67%	11.07%
MAD	0.8532	0.8037

From table (7) above we showed that the throw the three criteria (AIC,BIC,MAD,and MAPE) the Bayesian regression approach is better than OLS ,while throw the two criteria (MSE,RMSE) for two methods are equal ,and (R^2) for ordinary least square largest than Bayesian method ,knowing that the value (R^2=0.698) in Bayesian for three variable only . then the Bayesian regression method is best model for this data .

4: Conclusion & Recommendation:

4-1: Conclusion:

Form the Appling two regression method, we conclude this result:

1. Four variable's (affected the numbers of eggs in infertility, or parameters of these independent variables are significant, the value of explain these variables is 0.698.
2. The MSE are equal in two methods.
3. According to the criteria we conclude that the Bayesian regression method is best model for this data.
4. It is not necessary for all criteria suitable for every models.

4-2: Recommendation:

Finally, although Bayesian approach is difficult we recommended to:

1) Using Bayesian method and comparing all Bayesian methods with non- Bayesian methods.
2) Appling Bayesian method in medical filed.
3) We recommended researcher's to use 0dBayesian method

Reference:

1.Ali .Ahmed, Inglis. Alan n., Prado. Estevão, Bruna Wundervald,(2013).” Bayesian Linear Regression “1Statistics Solutions. (2013).

2.A.Hubin,, G. Storvik,, & F.Frommlet, (2018). A novel algorithmic approach to Bayesian Logic Regression. Bayesian Analysis.

3. Ausin .M. Concepcion ,(2009),”Bayesian Inference “Universidad Carlos III de Madrid , Master in Business Administration and Quantitative Methods

4. Ahmad .Sheikh P. (2011).” Bayesian Regression Analysis with Examples in S- PLUS and R “Journal of Modern Applied Statistical Methods,Vol.10,Issue1, Article 24 .

5. Bagnell.D, Hua.Rushane, and Kambam, Dheeraj R.” Statistical Techniques in Robotics “ (16-831, F14), Lecture#16, (Thursday November 6).

6. Burnham,K.P.etal.,(2004),)”multimodel inference understanding AIC and BIC in selection”Sociological methods & Research ,33(2):p(261-304).

7. Chipman.Hugh A.,Edward I. and McCulloch George .E.,(2010),”Additive Regression Trees “.the annals of applied statistics ,Vol 4 ,No.1 ,Page (266-298) .

8.Clyde.Merlise.M.,etal.,(2021), “. An Introduction to Bayesian Thinking a Companion to the Statistics with R Course “

9.Congdon .Peter, (2003) ,” Applied Bayesian Modelling”, John Wiley & Sons, Ltd.

10. Congdon Peter ,(2004), “ Bayesian model choice based on Monte Carlo estimates of posterior model probabilities “Department of Geography, Queen Mary University of London, Mile End Road, London E1 4NS, UK

11. Esian .Nelsy , (2018).” . An Introduction to Bayesian Linear Regression”

12. Evans .Sara, (2012),”Bayesian regression analysis” University of Louisville, M.A., University of Louisville.

13. Fox. John,(2018),” Bayesian Estimation of Regression Models “An Appendix to Fox & Weisberg an R Companion to Applied Regression, third edition.

14. H. Qian,, (2018). Big Data Bayesian Linear Regression and Variable Selection by Normal-Inverse-Gamma Summation. Bayesian Analysis, 13(4), 1007-1031.
15. Johnson, Nels,(2011).”Bayesian Methods for Regression in R”, Laboratory for Interdisciplinary Statistical Analysis.
16. K.H.,Brodersen ,F. Gallusser,, Koehler, J., Remy, N., & Scott, S. L. (2015). Inferring causal impact using Bayesian structural time-series models. The Annals of Applied Statistics, 9(1), 247-274.
17. Neelon, B. (2018). Bayesian Zero-Inflated Negative Binomial Regression Based on Pólya-Gamma Mixtures. Bayesian Analysis.
18. Miroshikov .Alexey , Savelev .Evgeny. and , Conlon .Erin M.(2015), “An R package for Bayesian Linear Models for Big Data and Data Science “
19. M.H.Walker, and Tobler,K.J.,2020.Female infertility .Stat Pears [Internet].
20. Minka.Thomas P.,(2010),”Bayesian linear regression “
21. Permaia.S.D., Heruna Tanty, (2018) ,” Linear regression model using Bayesian approach for energy performance of residential building “ 3rd International Conference on Computer Science and Computational Intelligence . Science Direct. Procedia Computer Science 135 (671–677).
22. Rice.Ken ,2017) ,”Bayesian Statistics for Genetics “Lecture 4: Linear Regression
23. Smith. Adam N, (2021).”Notes on Bayesian Linear Regression “,July 31.
24. Shekaramiz, Mohammad,. Moon. Todd. K,(2019),” Note on Bayesian Linear Regression”Electrical and Computer Engineering Faculty Publications
25. Wakefield, Jon,(2013) ,”Bayesian and Frequentist Regression Methods . Springer New York .
26. Yun Lin ,Meng,(2013)," Bayesian statistics “technical Report No.2 ,May 6 .
27. Look at website: Wallace H, Matthew and Tobler J .Kyle,(2021),concept of the infertility is available from (https://www.ncbi.nlm.nil.gov/books/NBK556033/).
28. Bayesian linear regression ST440/540: Applied Bayesian Statistics Spring, 2018
29. Look at www.inf.ed.ac.uk/teaching/couses /mlpr MLPR: w3b Iain Murray and Arno Onken, (2019) .
30. Look at APPM 5720: Bayesian Computation.
31. Look at , http://www.inf.ed.ac.uk/teaching/courses/mlpr/2019/.