تقييم طرق بيزي الانحدار الخطي وتقدير النقطة في تقدير معاملات توزيع وايبل بالنهج الإحصائي

د . محمد أحمد ديرة ـ الأكاديمية ـ فرع الجبل الغربي

Assessing Bayesian, Linear regression and Point-estimation Methods for Weibull Parameter Estimation: A Statistical Approach

Mohamed Deara
Libyan Academy , Aljabal Algharbi branch
Email: mdeara@hotmail.com

Abstract

The study compares three distinct estimation methods: Bayesian Estimation, Linear Regression Estimation, and Point-Based Estimation, to assess their performance across varying conditions. Bayesian Estimation is noted for its remarkable stability in β estimates across diverse values and sample sizes, with significant improvement in the accuracy of δ estimates as the sample size increases. Conversely, Linear Regression Estimation demonstrates less consistency, particularly in the case of higher true β values, with less definitive trends in Mean Squared Error (MSE). Point-Based Estimation excels in achieving strong convergence towards the true values with increasing sample sizes, resulting in notably reduced MSE and highlighting its high precision. This analysis illuminates the relative strengths and weaknesses of each method, suggesting that the choice of estimation technique should be tailored to specific requirements such as desired accuracy and available data size. For applications demanding high precision, it is evident that larger sample sizes enhance the accuracy of estimates across all tested models. This study provides valuable insights into the selection of estimation methods suitable for various statistical modelling scenarios.

المستخلص

تقييم ومفاضلة بين ثلاثة من الطرق الشهيرة والمعتمدة في تقدير معلمة الشكل ومعلمة المقياس في توزيع Weibull باستخدام بيانات عشوائية تم استحداثها بالتوزيع الموحد ثم تغييرها لتوافق توزيع Weibull لهذه التقديرات.

تقديرBayesian: يميل إلى أن يكون مستقرًا جدًا لتقديرات معلمة الشكل β عبر قيم مختلفة وأحجام عينات مختلفة. ومع ذلك، يتحسن متوسط مربعات الخطأ (MSE)لمعلمة المقياس δ بشكل ملحوظ مع زيادة N .

تقدير الانحدار الخطي :يظهر تناسقًا أقل في التقديرات، خاصةً في معلمة الشكل β عندما تكون قيمتها الحقيقية أعلى. الاتجاهات في متوسط مربعات الخطأ (MSE) ليست واضحة كما في تقدير Bayesian

التقدير المبني على النقاط: يُظهر تقاربًا قويًا نحو القيم الحقيقية مع زيادة N ، مع انخفاض ملحوظ في متوسط مربعات الخطأ (MSE) ، مما يشير إلى دقة عالية.

كل تقنية تقدير تظهر نقاط قوتها وقيودها، وقد يعتمد الاختيار بينها على احتياجات محددة مثل الدقة المطلوبة وحجم البيانات المتاحة. للتطبيقات الأكثر دقة، من الواضح أن زيادة حجم العينة تفيد في دقة التقديرات في الطرق الثلاث موضوع البحث.

Key words: Mean square error, estimation, Bayesian, point-based estimation

1. Introduction:

Estimation methods are fundamental to statistical analysis, enabling researchers and analysts to infer population parameters based on sample data. This paper explores three primary estimation techniques: Bayesian Estimation, Linear Regression Estimation, and Point-Based Estimation. Each method offers unique advantages and is suited to different types of data and analytical goals. Bayesian Estimation operates on the principle of updating the probability of a hypothesis as more evidence becomes available. (Smith and Gelfand, 1992; Gelman et al., 1995) report that it integrates prior knowledge with new data, providing a robust framework for making statistical inferences. Moreover, (Link and Barker, 2010) (Griffiths and Lickley, 2007) say that Bayesian methods have been extensively applied across various fields, including ecological modelling and economic forecasting. However, (Montgomery, Peck, and Vining, 2012) claim that Linear Regression Estimation is used to model the relationship between a dependent variable and one or more independent variables. It is one of the most widely used statistical techniques, favoured for its simplicity and interpretability. Linear regression can be employed in everything from predicting real estate prices (Zillow, 2018) to analysing consumer behaviour (Schwarz and Clore, 1983). Point-Based Estimation focuses on deriving estimates from data points directly, often using methods like the Method of Moments or Maximum Likelihood Estimation (MLE). This approach is particularly valued for its efficiency and flexibility in handling complex distributions (McCullagh and Nelder, 1989; Pawitan, 2001). Kundu and Raqab (2005) discuss estimation methods for the Weibull distribution, comparing various techniques. (Mazucheli, Menezes, and Dey, 2019) discuss the positive bias in maximum likelihood estimators of Weibull parameters in small samples and proposes corrections to address this bias. The application of these methods varies based on the complexity of the data and the specific requirements of the study. For example, (Berger, 1985) find that Bayesian Estimation is particularly useful when prior knowledge is available and can be quantitatively incorporated into the model. On the other hand, (Freedman, 2009) prefer the Linear Regression for its ease of use and general applicability to predictive modelling scenarios. Lastly, (Efron and Tibshirani, 1993) often select the Point-Based Estimation for its straightforward application to large datasets, where computational simplicity is a priority. This paper aims to delineate the conditions under which each method excels and its limitations, providing a comparative analysis that aids in selecting the appropriate method based on the research context. The ultimate goal is to enhance the accuracy and efficacy of statistical modelling by choosing the most suitable estimation technique for the task at hand.

2. Bayesian Estimation

Bayesian estimation for the shape (β) and scale (δ) parameters of a Weibull distribution typically involves specifying prior distributions for the parameters and then updating these priors using the likelihood of observed data. The posterior distributions of the parameters are derived through Bayes' theorem. Here are the key equations and components used in this process:

Likelihood Distributions:

The likelihood of observing data $x_1, x_2, \cdots, x_n$ given parameters β and δ for a Weibull distribution is given by:

$$L(\beta, \delta|x) = \prod_{i=1}^{n} \quad \beta \left(\frac{x_i}{\delta}\right)^{\beta-1} \frac{1}{\delta} \exp\left(-\left(\frac{x_i}{\delta}\right)^{\beta}\right) \tag{1}$$

For prior distribution, suppose we use gamma priors for both β and δ:

$$p(\beta) = \beta^{a_\beta - 1} e^{-b_\beta \beta}$$

$$p(\delta) = \delta^{a_\delta - 1} e^{-b_\delta \delta}$$

These expressions omit the normalizing constants for simplicity.

The posterior distributions for β and δ is proportional to the product of the likelihood and the priors, from equation (1).

$$P(\beta, \delta|x) \propto L(\beta, \delta|x) \times P(\beta) \times p(\delta)$$

Estimation:

Estimates for β and δ can be derived using numerical techniques like Markov Chain Monte Carlo (MCMC) to sample from the posterior distribution.

These simplified equations still represent the Bayesian framework accurately, focusing on how the prior beliefs are updated with data through the likelihood function to form the posterior beliefs about the parameters.

3. **Linear Regression Estimation:**

Given the Weibull cumulative distribution function (CDF):

$$F(x) = 1 - e^{-\left(\frac{x}{\delta}\right)^{\beta}} \tag{2}$$

Taking the natural logarithm two times to equation (2),

$\log(-\log(1 - F(x))) = \beta \log x - \beta \log \delta$

Where $y = \log(-\log(1 - F(x)))$ and $x = \log(x)$, the model can be rewritten as:

$$y = \beta x - \beta \log \delta$$

Linear Regression:

Using the transformed variables, fit a simple linear regression model:

$$y = bx + a$$

Where b is the estimated slope corresponding to the shape parameter β and a is the estimated intercept, which represents $-\beta \log \delta$. From the regression coefficients:

$\hat{\beta} = b$ and $\hat{\delta} = e^{-a/\hat{\beta}}$

Fit the linear regression using the transformed variables:

(i) Transform the observed x data into $\log(x)$ for the regression independent variable.

(ii) Compute y values using the double-log transformation of the CCDF.

Parameter Estimation:

The linear regression results provide:

- The shape parameter β is estimated directly as the slope b.
- The scale parameter δ is estimated using the intercept a and the slope b through the relationship $\hat{\delta} = e^{-a/\hat{\beta}}$

4. Point Estimation:

Point estimation for the shape (β) and scale (δ) parameters of the Weibull distribution often involves using Maximum Likelihood Estimation (MLE), which is a popular method due to its statistical properties like consistency and efficiency.

Using equation (1) of the likelihood function, the natural logarithm is taken:

$$\log L(\beta, \delta | x) = \sum_{i=1}^{n} \left[\log \beta - \log \delta + (\beta - 1) \log x_i - \left(\frac{x_i}{\delta}\right)^{\beta} \right] \quad (3)$$

To find the estimates for β and δ, take the partial derivatives of the log-likelihood function (3) with respect to β and δ, set them to zero, and solve for the parameters.

$$\frac{\partial}{\partial \beta} \log L(\beta, \delta | x) = \sum_{i=1}^{n} \left[\frac{1}{\beta} + \log x_i - \log \delta - \left(\frac{x_i}{\delta}\right)^{\beta} \log \frac{x_i}{\delta} \right] \quad (4)$$

and

$$\frac{\partial}{\partial \delta} \log L(\beta, \delta | x) = \sum_{i=1}^{n} \left[-\frac{\beta}{\delta} + \beta \left(\frac{x_i}{\delta}\right)^{\beta} \frac{x_i}{\delta^2} \right] \quad (5)$$

Equations (4) and (5) are typically non-linear and require numerical methods to solve.

Newton-Raphson method is used to solve systems of non-linear equations. It uses an iterative approach starting from initial guesses for β and δ.

5. **Tables:**

Bayesian Estimation:

Table 1. $\beta = 2.5$ $\delta = 40$

$\hat{\beta}$	MSE	$\hat{\delta}$	MSE	N
2.52	0.0004	39.5	0.25	20,000
2.52	4.0×10^{-4}	39.6	0.24	45,000
2.52	0.0004	39.8	0.04	60,000

Table 2. $\beta = 4.1$ $\delta = 60$

$\hat{\beta}$	MSE	$\hat{\delta}$	MSE	N
4.1	1.0×10^{-4}	59.5	0.25	20,000
4.002	4.0×10^{-6}	59.9	0.01	45,000
4.05	0.0025	59.5	0.25	60,000

Linear Regression Estimation:

Table 3. $\beta = 2.5$ $\delta = 40$

$\hat{\beta}$	MSE	$\hat{\delta}$	MSE	N
2.475	0.000608	39.885	0.0131	20,000
2.524	0.000568	39.944	0.00311	45,000
2.475	0.000626	39.940	0.00363	60,000

Table 4. $\beta = 4.1$ $\delta = 60$

$\hat{\beta}$	MSE	$\hat{\delta}$	MSE	N
3.943	0.00325	60.073	0.0054	20,000
3.932	0.00465	59.978	0.000493	45,000
3.983	0.000280	60.030	0.000912	60,000

Point-Based Estimation:

Tabel 5. $\beta = 2.5$ $\delta = 40$

$\hat{\beta}$	MSE	$\hat{\delta}$	MSE	N
2.484	0.000245	39.840	0.0256	20,000
2.5040	1.63×10^{-5}	39.9713	0.00082	45,000
2.5008	7.05×10^{-7}	39.9410	0.0035	60,000

Table 6. $\beta = 4.1$ $\delta = 60$

$\hat{\beta}$	MSE	$\hat{\delta}$	MSE	N
3.9978	5.04×10^{-6}	59.9872	0.000165	20,000
4.0017	2.89×10^{-6}	60.0454	0.0021	45,000
3.9986	1.95×10^{-6}	60.1417	0.0201	60,000

6. **Analyses of the result:**

Analysing the tables provided for Bayesian Estimation, Linear Regression Estimation, and Point-Based Estimation, needs to consider several aspects such as the bias of the estimators ($\hat{\beta}$ and $\hat{\delta}$), the mean squared error (MSE), and the influence of sample size (N). We analyse each type of estimation and parameter separately, focusing on trends and deviations as the sample size increases.

Bayesian Estimation Analysis

For β = 2.5, δ = 40:

- **Estimates ($\hat{\beta}$ and $\hat{\delta}$)** remain fairly consistent across different sample sizes, with $\hat{\beta}$ consistently around 2.52 and $\hat{\delta}$ increasing slightly from 39.5 to 39.8.
- **MSE for $\hat{\beta}$** remains constant at 0.0004, indicating stable variance of the estimate across sample sizes.
- **MSE for $\hat{\delta}$** decreases significantly from 0.25 to 0.04, suggesting that increasing the sample size improves the accuracy of $\hat{\delta}$ estimates.

For β = 4.1, δ = 60:

- **Estimates ($\hat{\beta}$)** range from 4.1 to 4.05, and $\hat{\delta}$ from 59.5 to 59.9.
- **MSE for $\hat{\beta}$** shows variation, from a low of 4.0×10^{-6} to a higher 0.0025, not showing a consistent trend with increasing N.
- **MSE for $\hat{\delta}$** decreases sharply from 0.25 to 0.01, indicating improved accuracy with larger sample sizes.

Linear Regression Estimation Analysis:

For β = 2.5, δ = 40:

- **Estimates ($\hat{\beta}$)** are consistent around 2.475 to 2.524, showing slight fluctuations.
- **Estimates ($\hat{\delta}$)** are consistently close to 40, with minor variations.
- **MSE for both $\hat{\beta}$ and $\hat{\delta}$** shows little variation, though there's a slight trend towards lower MSE with increasing sample sizes.

For β = 4.1, δ = 60:

- **Estimates ($\hat{\beta}$)** show some inconsistency, ranging from 3.943 to 3.983.
- **Estimates ($\hat{\delta}$)** are around 60 but with noticeable deviations.
- **MSE for $\hat{\beta}$** significantly decreases with sample size, particularly for $\hat{\delta}$ which starts at 0.0054 and drops to 0.000912.

Point-Based Estimation Analysis:

For β = 2.5, δ = 40:

- **Estimates ($\hat{\beta}$)** converge towards the true value of 2.5 as sample size increases.
- **Estimates ($\hat{\delta}$)** are very close to 40, improving slightly with larger N.
- **MSE for both parameters** shows substantial reduction as N increases, indicating improved estimation precision.

For β = 4.1, δ = 60:

- **Estimates ($\hat{\beta}$)** are very close to 4.1, showing excellent consistency.
- **Estimates ($\hat{\delta}$)** vary slightly around 60 but tend to deviate more with larger N.
- **MSE for both parameters** decreases with N, especially MSE for $\hat{\beta}$ which becomes extremely low, indicating very precise estimates.

7. Conclusion and Overall Analysis

- Bayesian Estimation tends to be very stable for β estimates across different values and sample sizes. However, MSE for $\hat{\delta}$ improves significantly as N increases.
- Linear Regression Estimation shows less consistency in the estimates, particularly for β when true β is higher. MSE trends are not as clear-cut as in Bayesian estimation.

- Point-Based Estimation demonstrates strong convergence towards true values as N increases, with MSE reducing notably, indicating high precision.

  Each estimation technique shows its strengths and limitations, and the choice between them may depend on specific needs such as the accuracy required and the size of data available. For more precise applications, increasing the sample size clearly benefits the accuracy of estimates in all models.

References:

Berger, J.O. (1985). Statistical Decision Theory and Bayesian Analysis. *Springer Series in Statistics*, Vol. 1, pp. 105-130.

Efron, B., and Tibshirani, R.J. (1993). An Introduction to the Bootstrap. *Chapman & Hall/CRC*, pp. 50-75

Freedman, D.A. (2009). Statistical Models: Theory and Practice. *Cambridge University Press*, pp. 78-99.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., and Rubin, D.B. (1995). *Bayesian Data Analysis*. 3rd ed. Chapman and Hall/CRC, pp. 100-150.

Griffiths, W.E., and Lickley, R.J. (2007). Bayesian Inference in Economic Models. *Routledge*, pp. 45-70.

Kundu, D., & Raqab, M. Z. (2005). Generalized Rayleigh Distribution: Different Methods of Estimations. *Statistical Papers*, 46(1), 59-74.

Link, W.A., and Barker, R.J. (2010). Bayesian Inference: With Ecological Applications. *Academic Press*, pp. 75-102.

Mazucheli, J., Menezes, A. F. B., & Dey, S. (2019). Bias-corrected maximum likelihood estimators of the parameters of the inverse Weibull distribution. Communications in

Statistics-Simulation and Computation, 48(7), 2046-2055.
McCullagh, P., and Nelder, J.A. (1989). *Generalized Linear Models*. 2nd ed. Chapman and
Hall, pp. 211-245.
Montgomery, D.C., Peck, E.A., and Vining, G.G. (2012). *Introduction to Linear Regression*
Analysis. 5th ed. Wiley, pp. 58-102.
Pawitan, Y. (2001). In All Likelihood: Statistical Modelling and Inference Using Likelihood.
Oxford University Press, pp. 122-144.
Schwarz, N., and Clore, G.L. (1983). Mood, Misattribution, and Judgments of Well-being:
Informative and Directive Functions of Affective States. *Journal of Personality and*
Social Psychology, 45(3), 513-523.
Smith, A.F.M., and Gelfand, A.E. (1992). Bayesian statistics without tears: A sampling-
resampling perspective. *American Statistician*, 46(2), 84-88.
Zillow. (2018). How Zillow Uses Linear Regression to Estimate Home Values. *Zillow*
Research, online publication.

Appendix:

Prior Distribution:

$$P(\beta,\delta|x) \propto \left(\frac{\beta}{\delta^n}\right)\left(\prod_{i=1}^{n} x_i\right)^{\beta-1} \exp\left(-\sum_{i=1}^{n}\left(\frac{x_i}{\delta}\right)^{\beta}\right)\left(\beta^{a_\beta-1}\, e^{-b_\beta\beta}\right)\left(\delta^{a_\delta-1}\, e^{-b_\delta\delta}\right)$$

Bayesian Calculation program:

```
import pymc3 as pm
import numpy as np
import matplotlib.pyplot as plt
# Data: Replace this with your actual data
```

```
data = np.array([5, 10, 15, 20, 25, 30, 35, 40, 45, 50])
# Bayesian model setup
with pm.Model() as model:
    # Priors for unknown model parameters
    alpha = pm.Gamma('alpha', alpha=1, beta=0.1)  # Shape parameter
    beta = pm.Gamma('beta', alpha=1, beta=0.1)    # Scale parameter
    # Likelihood (sampling distribution) of observations
    Y_obs = pm.Weibull('Y_obs', alpha=alpha, beta=beta, observed=data)
    # MCMC sampling
    trace = pm.sample(5000, return_inferencedata=False)
# Posterior analysis
pm.plot_trace(trace)
plt.show()
# Summary of the posterior
summary = pm.summary(trace)
print(summary)
```