Coursework 6: STAT 570

Philip Pham

November 8, 2018

1. In this question, you will implement the algorithm you described in Question 4 of Exercises 5. The algorithm derived in 4(b) will now be implemented for the prostate cancer data. These data are available in the R package `lasso2` and are named `Prostate`. Take Y as log prostate specific antigen and x as log cancer volume. Implement the blocked Gibbs sampling algorithm using the prior given in the first equation of the aforementioned question, with $m_0 = m_1 = 0$, $v_{00} = v_{11} = 2$, $v_{01} = 0$, and $a = b = 0$. Run two chains, one with starting values corresponding to the unbiased estimates of the parameters and one starting from a point randomly generated from the prior $\pi(\beta_0, \beta_1)$. Report:

   (a) Histogram representations of the univariate marginal distributions $p(\beta_0 \mid y)$, $p(\beta_1 \mid y)$ and $p(\sigma \mid y)$, and scatterplots of the bivariate marginal distributions $p(\beta_0, \beta_1 \mid y)$, $p(\beta_0, \sigma \mid y)$, and $p(\beta_1, \sigma \mid y)$.

   Solution: The empirical univariate distributions from Gibbs sampling can be found in Figure 1. The joint distributions are found in Figures 2, 3, and 4.
   From the univariate distributions in Figure 1, we get samples close to the MLE estimates (see Table 5.3 of Wakefield's *Bayesian and Frequentist Regression Methods*). The dataset is rather large, so this is not surprising. The distributions for β_j are symmetrical. The distribution for σ seems to skew slightly to the right.
   From the bivariate distributions in Figures 3 and 4, we see that there is not much correlation between the β_j and σ, which is expected since σ is an ancillary statistic in the frequentist setting.
   From Figure 2, we see a negative correlation between β_0 and β_1, which is also expected since if we discount the effect of x on y, the estimate for β_0 must compensate.
   Code for the Gibbs sampler and plots can be found at `prostate.ipynb`.

   (b) The posterior means, standard deviations and 10%, 50%, 90% quantiles of β_0, β_1, and σ.

	Posterior mean	Standard deviation	10% quantile	50% quantile	90% quantile
σ	0.796154	0.056701	0.726527	0.793442	0.871269
β_0	1.495039	0.120843	1.341545	1.494805	1.652094
β_1	0.723290	0.068159	0.638009	0.724150	0.808873

Table 1: Summary statistics calculated from samples drawn with Gibbs sampling.

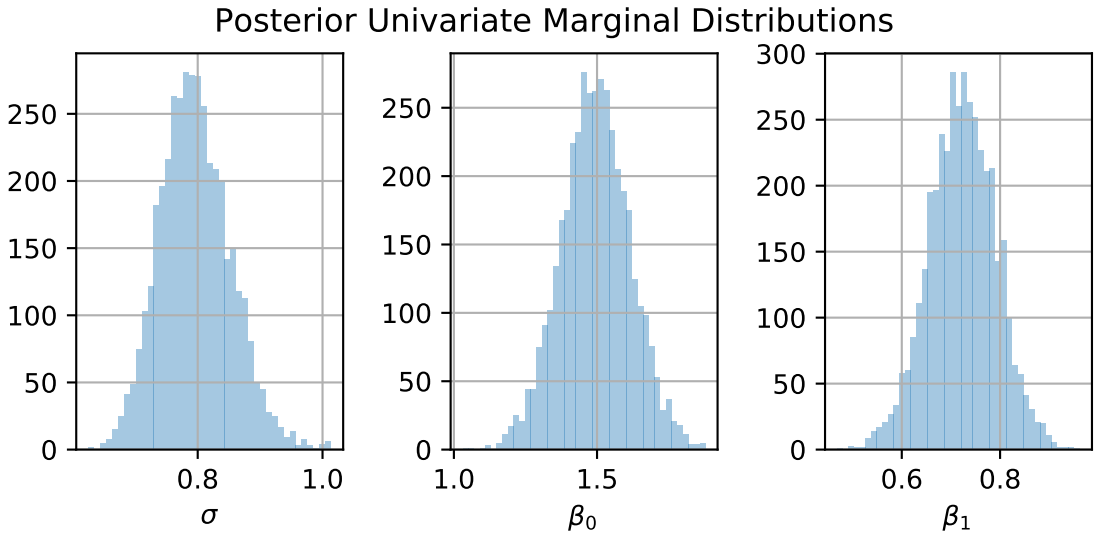


Figure 1: Emprical univariate distributions from Gibbs sampling.

Figure 2: Emprical joint distribution for $(\beta_0, \beta_1) \mid y$ from Gibbs sampling.

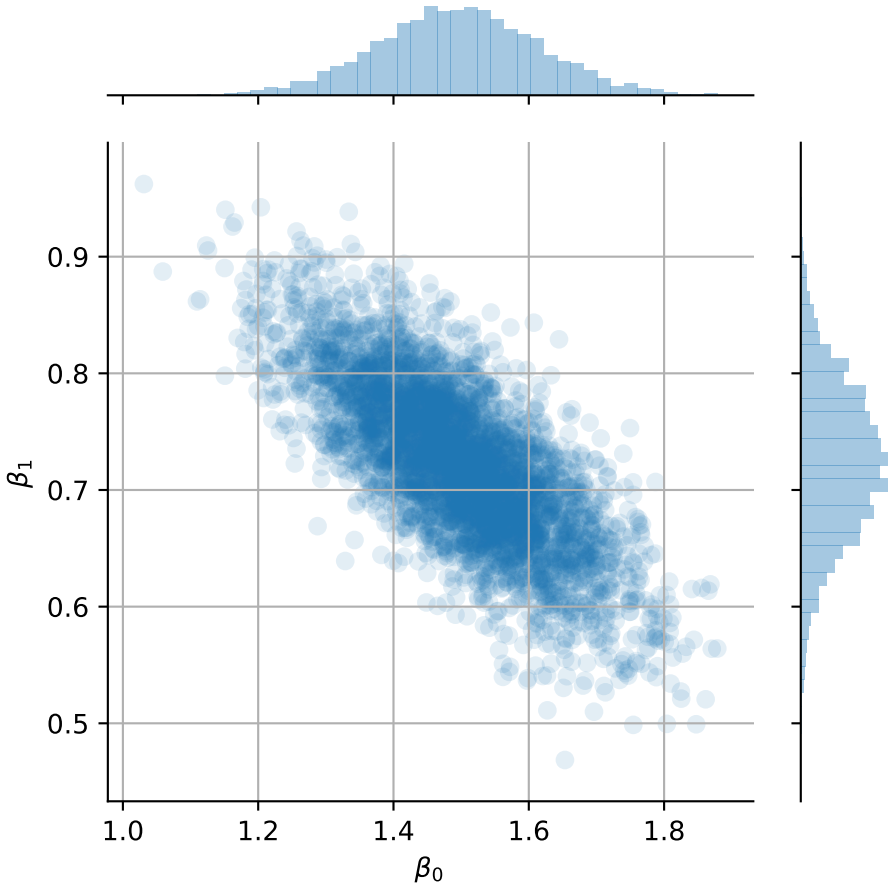


Figure 3: Emprical joint distribution for $(\beta_0, \sigma) \mid y$ from Gibbs sampling.

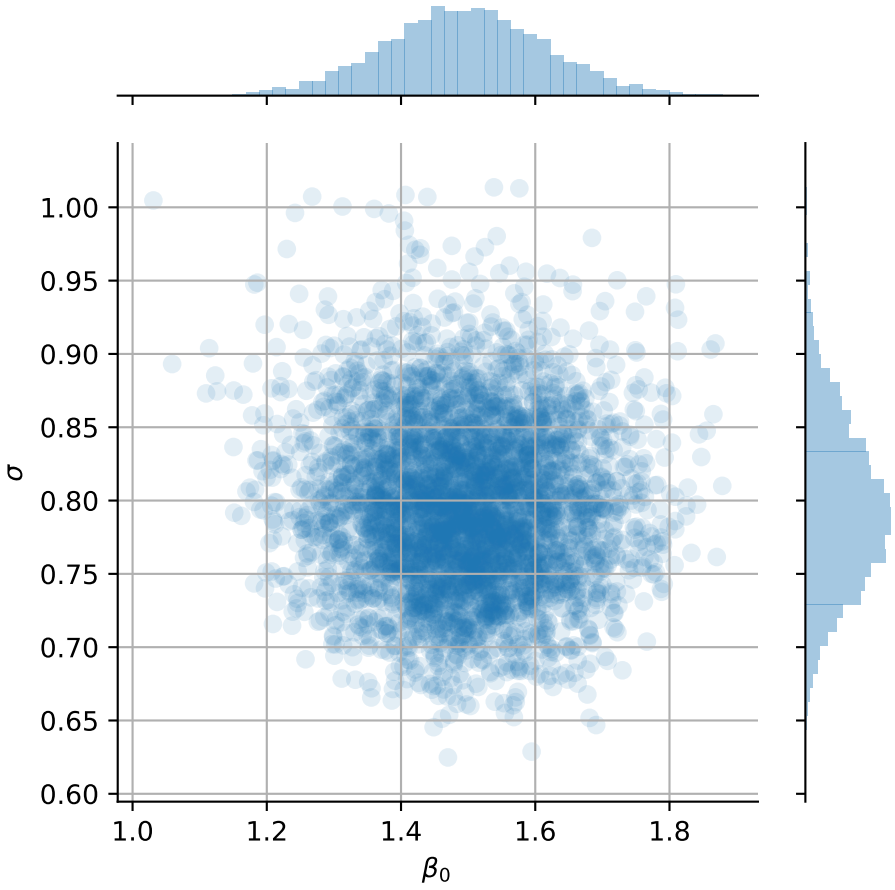
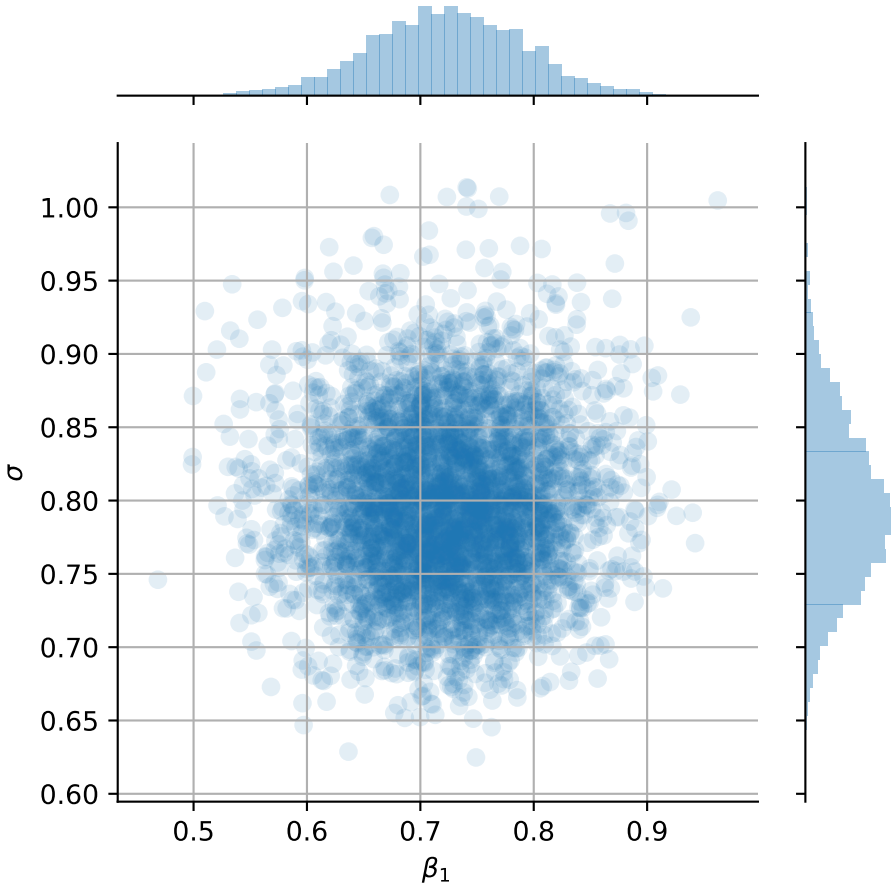


Figure 4: Emprical joint distribution for $(\beta_1, \sigma) \mid y$ from Gibbs sampling.



Solution: The summary statistics for the empirical posterior distributions can be found in Table 1. Samples from both chains were used for a total 4,096 samples.

Both the posterior mean and standard deviation agree closely with the MLE estimates in Table 5.3 of Wakefield's *Bayesian and Frequentist Regression Methods*. From the quantiles, we get an empirical estimate of the 80% credible interval, which appear to be symmetrical with respect to the median.

(c) $\mathbb{P}(\beta_1 > 0.5 \mid y)$

Solution: One way to interpret the significance of x's effect on y is to look at the distribution of β_1. The emprical estimate for $\mathbb{P}(\beta_1 > 0.5 \mid y)$ is $\boxed{0.9990}$, so the effect is likely significant, both statistically and in size.

This was calculated simply by taking the proportion of samples of β_1 that exceeded 0.5. See `prostate.ipynb` for the calculation.

(d) Justify your choice of *burn-in* period. For example, you may present the trace plots $\beta_0^{(t)}$, $\beta_1^{(t)}$, $\left(\log \sigma^2\right)^{(t)}$ versus t.

Solution: The trace plots for $\log \sigma^2$, β_0, and β_1 are shown in Figure 5. The first 128 results are plotted. The MLE chain immediately is stationary. The prior chain quickly becomes stationary in about 10 steps.

I specified a burn-in period of 128 steps for good measure. Then, I took 2,048 samples from each chain with no thinning.

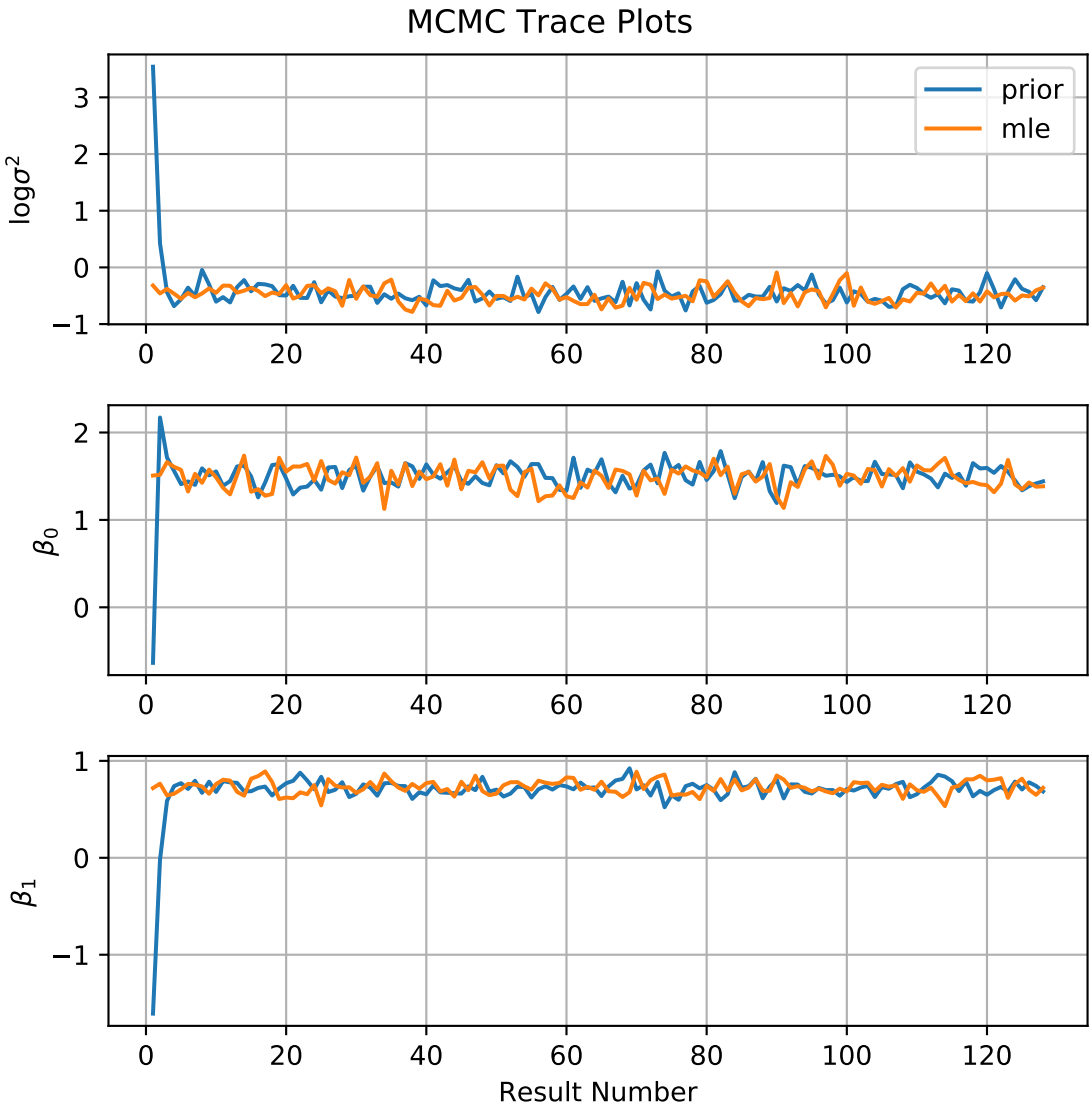


Figure 5: The trace plots show the sampled posterior parameters at each step in the MCMC chain.

Length (mm)	0	1	2	3	4	5	6	7	8	9	10	11	12
1	2.247	2.640	2.842	2.908	3.099	3.126	3.245	3.328	3.355	3.383	3.572	3.581	3.681
10	1.901	2.132	2.203	2.228	2.257	2.350	2.361	2.396	2.397	2.445	2.454	2.454	2.474
20	1.312	1.314	1.479	1.552	1.700	1.803	1.861	1.865	1.944	1.958	1.966	1.997	2.006
50	1.339	1.434	1.549	1.574	1.589	1.613	1.746	1.753	1.764	1.807	1.812	1.840	1.852

Table 2: Failure stress data for four groups of fibers.

2. Consider the data in Table 2 contain data on a typical reliability experiment and give the failure stresses (in GPa) of four samples of carbon fibers of lengths 1, 10, 20 and 50mm.

   (a) Consider a Bayesian analysis with a Weibull likelihood and independent lognormal priors, $\eta \sim \text{LogNormal}(\mu_\eta, \sigma_\eta)$, $\alpha \sim \text{LogNormal}(\mu_\alpha, \sigma_\alpha)$. Choose μ_η, σ_η so that the prior probability that η lies between 0.5 and 30 is 0.9, and μ_α, σ_α so that the prior probability that Îś lies between 1 and 4 is 0.9.

   Solution: $\log \eta \sim \mathcal{N}(\mu_\eta, \sigma_\eta)$ by definition of the lognormal distribution. Since log is a monotonic transformation,

$$\begin{aligned} 0.9 = \mathbb{P}(1/2 \leq \eta \leq 30) &= \mathbb{P}\left(\log \frac{1}{2} \leq \log \eta \leq \log 30\right) \\ &= \mathbb{P}\left(\Phi^{-1}(0.05) \leq \frac{\log \eta - \mu_\eta}{\sigma_\eta} \leq \Phi^{-1}(0.95)\right), \quad (1) \end{aligned}$$

   where Φ is the cumulative distribution function of a standard normal. Equation 1 implies that

$$\begin{aligned} \frac{\log(1/2) - \mu_\eta}{\sigma_\eta} &= \Phi^{-1}(0.05) \\ \frac{\log 30 - \mu_\eta}{\sigma_\eta} &= \Phi^{-1}(0.95). \end{aligned}$$

   Solving, we have that

$$\begin{aligned} \sigma_\eta &= \frac{\log 30 - \log \frac{1}{2}}{\Phi^{-1}(0.95) - \Phi^{-1}(0.05)} \approx 1.2446 \\ \mu_\eta &= \log 30 - \sigma_\eta \Phi^{-1}(0.95) = \log \frac{1}{2} - \sigma_\eta \Phi^{-1}(0.05) \approx 1.3540 \end{aligned}$$

   Repeating the calculating for α, we have $\mu_\alpha \approx 0.6931$ and $\sigma_\alpha \approx 0.4214$. Calculations can be found in `failure_stresses.ipynb`.

   (b) Run MCMC for summarizing the posterior $p(\eta, \alpha \mid y)$, and implement this algorithm for each of the groups in Table 2. Report the posterior medians and 90% credible intervals for η and α and give histograms representations of the posterior margins for η and α, and a scatterplot representation of $p(\eta, \alpha \mid y)$.

   Solution: For each length, 32,768 MCMC samples were generated with Hamiltonian Monte Carlo. Some thinning was used: 2 steps were skipped between each sample.

   The trace plots can be seen in Figure 6. A significant number of burn-in steps (65,536) were needed before the distribution became stationary.

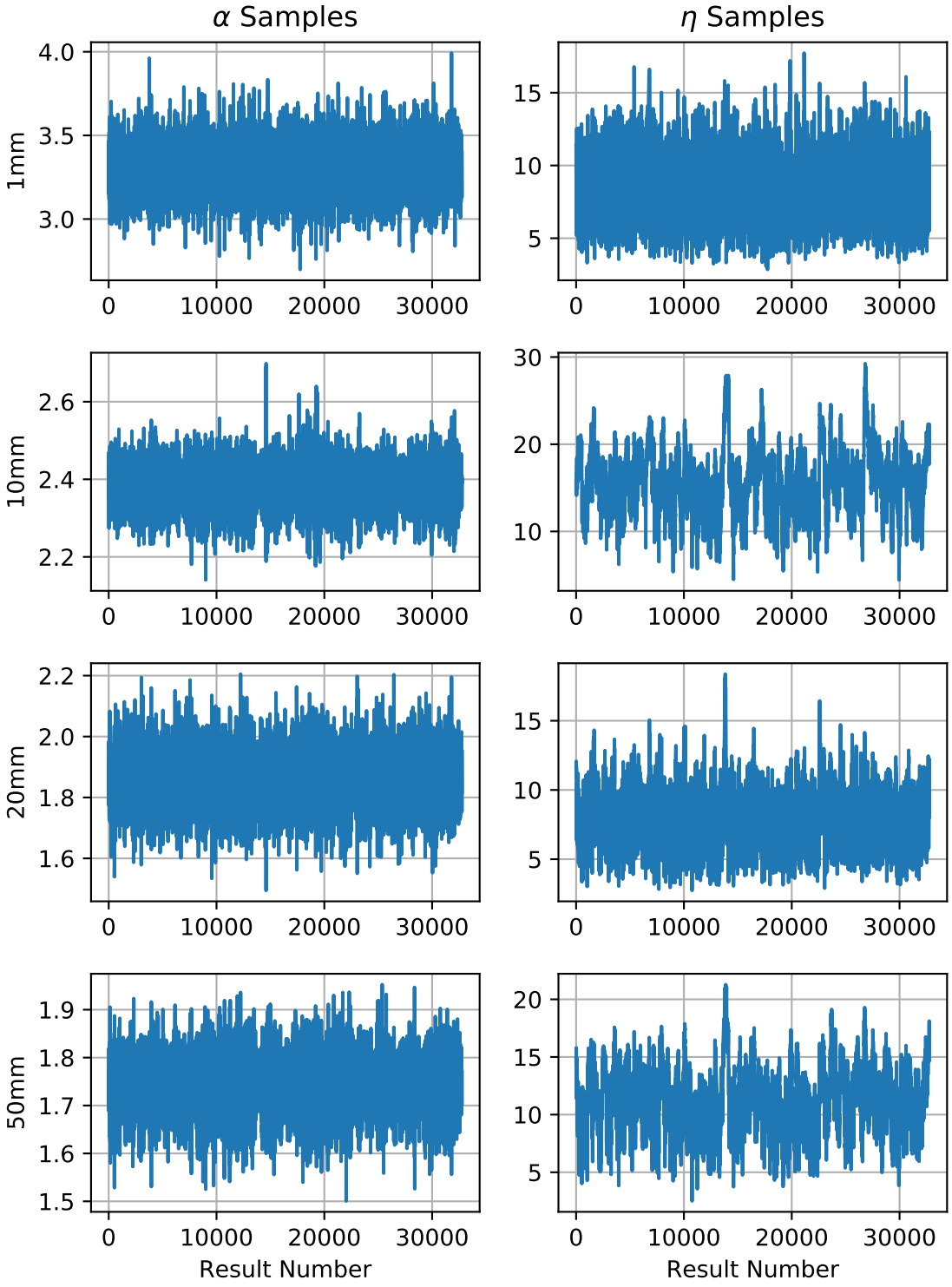


Figure 6: Trace plots for the MCMC samples of the Weibull parameters following burn-in.

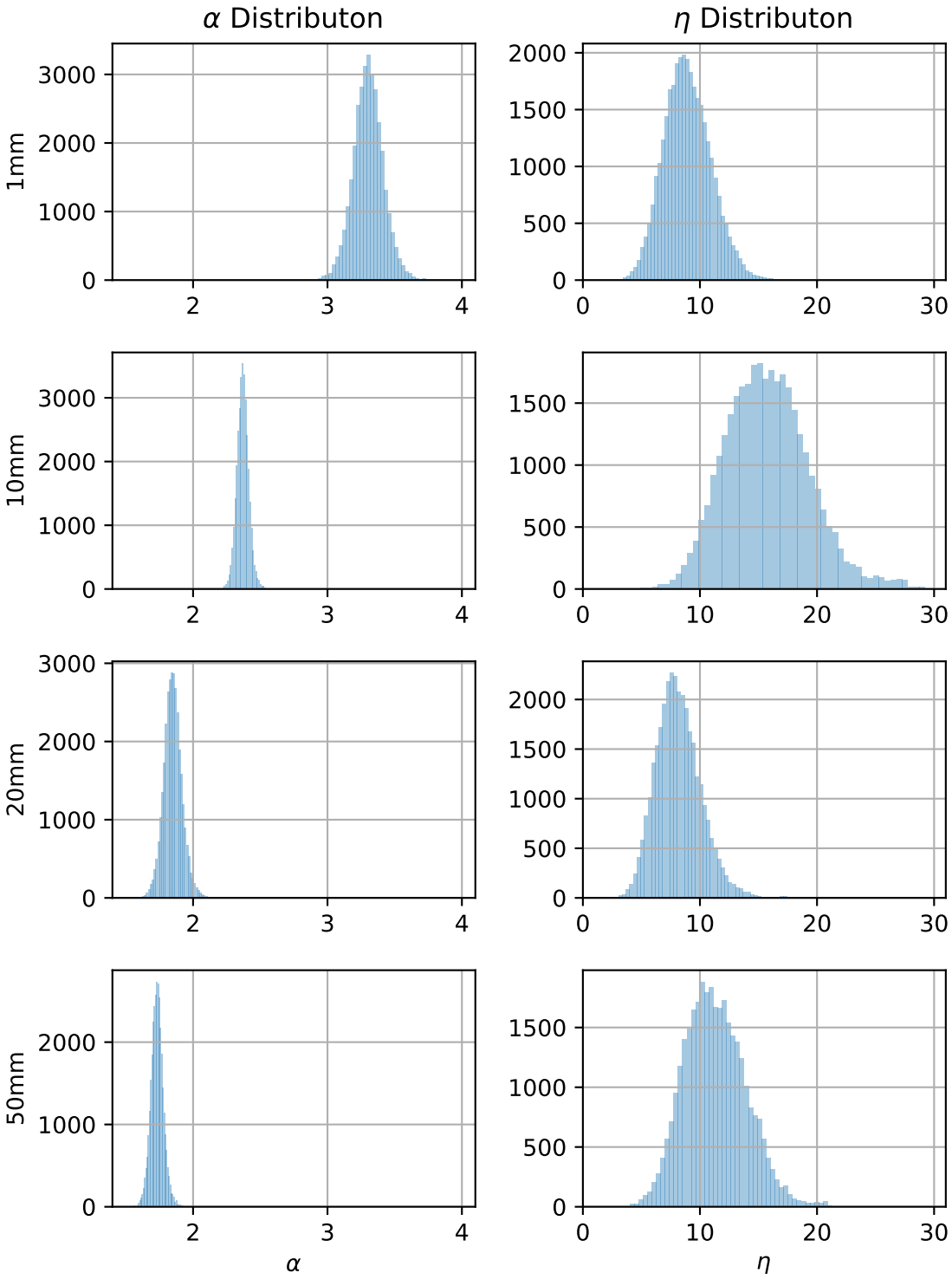


Figure 7: Samples from MCMC were used to approximate the posterior distribution of the parameters.

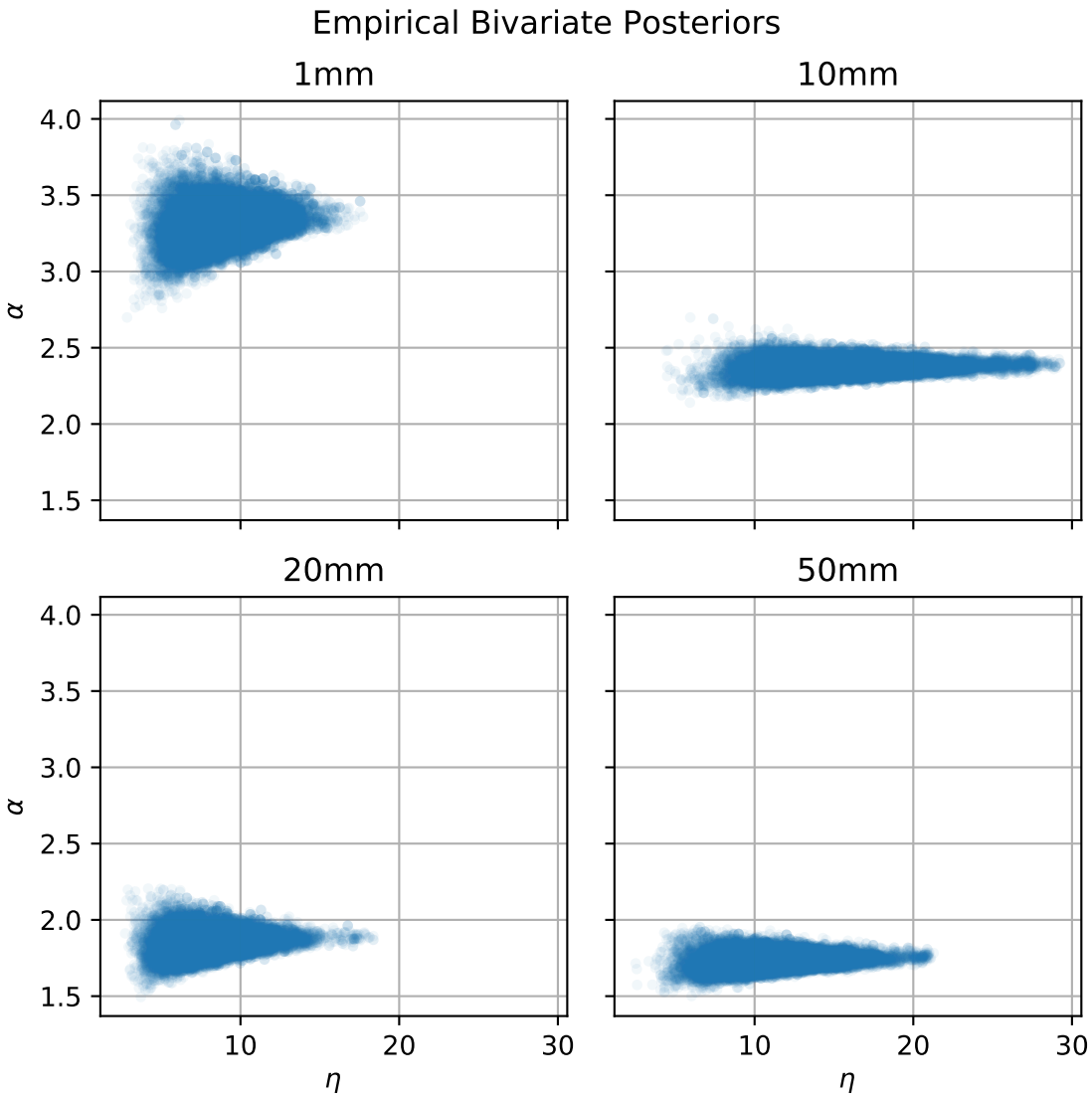


Figure 8: Correlation between the parameters can be estimated with the empirical joint distributions.

Length (mm)	Parameter	Mean	Standard deviation	Median	5% quantile	95% quantile
1	α	3.304722	0.113267	3.303521	3.122608	3.490909
	η	8.945404	1.989280	8.837034	5.846150	12.363148
10	α	2.369973	0.045122	2.369507	2.298008	2.443862
	η	15.779253	3.575655	15.611898	10.309790	21.820534
20	α	1.849104	0.069687	1.847223	1.738584	1.965827
	η	8.176279	1.934859	8.019274	5.281827	11.535180
50	α	1.732334	0.047075	1.731810	1.656464	1.810707
	η	11.302895	2.636021	11.133819	7.279310	15.768627

Table 3: Posterior estimates of summary statistics calculated from the MCMC samples.

The empirical univariate distributions can be found in Figure 7. The lognormal assumption seems appropriate.
The empirical joint distributions can be found in Figure 8. There is a triangle shape: as the shape parameter η increases there is less variance in the estimate of α.
Posterior summary statistics can be found in Table 3. The 90% credible interval can be taken from the last two columns. These numbers are similar to the MLEs calculated in Homework 3, but pulled towards the prior means and and standard deviation. Since the number of samples is small the prior has a significant effect.
Code to run the MCMC chain and recreate the plots can be found in `failure_stresses.ipynb`.

i	time	drug concentration
1	2	1.63
2	4	1.01
3	6	0.73
4	8	0.55
5	10	0.41
6	24	0.01
7	28	0.06
8	32	0.02

Table 4: Concentrations of the drug Cadralazine (in mg/liter, y_i) as a function of time (in hours, x_i), for $i = 1, \ldots, 8$.

3. The data in Table 4, taken from Wakefield et al. (1994), were collected following the administration of a single 30mg dose of the drug Cadralazine to a cardiac failure patient. The response y_i represents the drug concentration at time x_i, $i = 1, \ldots, 8$. The most straightforward model for these data is to assume

$$\log y_i = \mu(\beta) + \epsilon_i = \log\left[\frac{D}{V}\exp(-k_e x_i)\right] + \epsilon_i \tag{2}$$

where $\epsilon_i \mid \sigma^2 \sim_{\text{iid}} \mathcal{N}(0, \sigma^2)$, $\beta = \begin{pmatrix} V & k_e \end{pmatrix}^\intercal$ and the dose is $D = 30$. The parameters are the volume of distribution $V > 0$ and the elimination rate k_e.

(a) For this model obtain expressions for:

i. The log-likelihood function $l(\beta, \sigma^2)$.

ii. The score function $S(\beta, \sigma^2)$.

iii. The expected information matrix $I(\beta, \sigma^2)$.

Solution: We have that

$$\frac{\log y_i - \log D + \log V + k_e x_i}{\sigma} = \epsilon_i \sim \mathcal{N}(0, 1),$$

so $Y_i \sim \text{LogNormal}\left(\log D - \log V - k_e x_i, \sqrt{\sigma^2}\right)$ so the log-likelihood function is

$$l\left(\beta, \sigma^2\right) = \log L\left(\beta, \sigma^2\right) = \sum_{i=1}^{n} \log p\left(y_i \mid x_i, \beta, \sigma^2\right) \tag{3}$$

$$= -\frac{n}{2}\log(2\pi) - \sum_{i=1}^{n}\log y_i - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(\log y_i - \log D + \log V + k_e x_i\right)^2,$$

where $n = 8$.

Taking the gradient of Equation 3, we find the score function

$$S(\beta) = \nabla^\intercal l\left(\beta, \sigma^2\right)$$

$$= \begin{pmatrix} -\frac{1}{V\sigma^2}\sum_{i=1}^{n}\left(y_i - \log D + \log V + k_e x_i\right) \\ -\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i\left(y_i - \log D + \log V + k_e x_i\right) \\ -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}\left(y_i - \log D + \log V + k_e x_i\right)^2 \end{pmatrix}. \tag{4}$$

	Estimate	Standard error	95% CI lower bound	95% CI upper bound
$\hat{V}$	16.663309	6.165432	4.579285	28.747333
$\hat{k}_e$	0.152106	0.020508	0.111911	0.192302
$\hat{\sigma}^2$	0.411963	0.205981	0.008247	0.815679

Table 5: Parameter estimates are computed using the MLE, and confidence intervals use Fisher information.

Let $\delta_i(\beta) = y_i - \log D + \log V + k_e x_i$ and $RSS(\beta) = \sum_{i=1}^n [\delta_i(\beta)]^2$. The expected information matrix can also be calculated from Equation 3

$$
\begin{aligned}
I_n\left(\beta, \sigma^2\right) &= -\mathbb{E}\left[\nabla\nabla^\intercal l\left(\beta, \sigma^2\right)\right] \qquad (5)\\
&= \mathbb{E}\left[\begin{pmatrix} \frac{n}{V^2\sigma^2} - \frac{1}{V^2\sigma^2}\sum_{i=1}^n \delta_i(\beta) & \frac{1}{V\sigma^2}\sum_{i=1}^n x_i & -\frac{1}{V(\sigma^2)^2}\sum_{i=1}^n \delta_i(\beta) \\ \frac{1}{V\sigma^2}\sum_{i=1}^n x_i & \frac{1}{\sigma^2}\sum_{i=1}^n x_i^2 & -\frac{1}{(\sigma^2)^2}\sum_{i=1}^n x_i\delta_i(\beta) \\ -\frac{1}{V(\sigma^2)^2}\sum_{i=1}^n \delta_i(\beta) & -\frac{1}{(\sigma^2)^2}\sum_{i=1}^n x_i\delta_i(\beta) & -\frac{n}{2(\sigma^2)^2} + \frac{1}{(\sigma^2)^3}RSS(\beta) \end{pmatrix}\right].\\
&= \sum_{i=1}^n \begin{pmatrix} \frac{1}{V^2\sigma^2} & \frac{1}{V\sigma^2}x_i & 0 \\ \frac{1}{V\sigma^2}x_i & \frac{1}{\sigma^2}x_i^2 & 0 \\ 0 & 0 & \frac{1}{2(\sigma^2)^2} \end{pmatrix}.
\end{aligned}
$$

since $\mathbb{E}[RSS(\beta)] = n\sigma^2$ and $\mathbb{E}[S(\beta, \sigma)] = \mathbf{0} \implies \mathbb{E}[\delta_i] = 0$.

(b) Obtain the MLE, and give an asymptotic 95% confidence interval for each element of β.

Solution: Let $\bar{x}$ and $\bar{y}$ be the empirical means of the x_i and y_i respectively. The MLE can be obtained from Equation 4: we solve $S\left(\hat{\beta}, \hat{\sigma}^2\right) = \mathbf{0}$ to obtain

$$
\begin{aligned}
\hat{V} &= \exp\left(-\left(\bar{y} - \log D - \hat{k}_e\bar{x}\right)\right)\\
\hat{k}_e &= -\frac{n^{-1}\sum_{i=1}^n x_i y_i - \bar{x}\bar{y}}{n^{-1}\sum_{i=1}^n x_i - \bar{x}^2}\\
\hat{\sigma}^2 &= \frac{1}{n}RSS\left(\hat{\beta}\right). \qquad (6)
\end{aligned}
$$

The approximate covariance matrix can be obtained by inverting Equation 5:

$$
\begin{aligned}
\mathrm{Var}\left(\hat{\beta}, \hat{\sigma}^2\right) &\approx \left[I_n\left(\hat{\beta}, \hat{\sigma}^2\right)\right]^{-1} \qquad (7)\\
&= \begin{pmatrix} \frac{\hat{\sigma}^2\hat{V}^2}{n^2\hat{\mathrm{Var}}(X)}\sum_{i=1}^n x_i^2 & -\frac{\hat{\sigma}^2\hat{V}}{n\hat{\mathrm{Var}}(X)}\bar{x} & 0 \\ -\frac{\hat{\sigma}^2\hat{V}}{n\hat{\mathrm{Var}}(X)}\bar{x} & \frac{\hat{\sigma}^2}{n\hat{\mathrm{Var}}(X)} & 0 \\ 0 & 0 & 2\frac{\left(\hat{\sigma}^2\right)^2}{n} \end{pmatrix},
\end{aligned}
$$

where $\hat{\mathrm{Var}}(X) = n^{-1}\sum_{i=1}^n x_i^2 - \bar{x}^2$ is the empirical variance estimate of the $\{x_i\}$.

Applying the formulas in Equations 6 and 7, we can compute parameter estimates and 95% confidence intervals, which can be found in Table 5.

(c) Plot the data, along with the fitted curve.

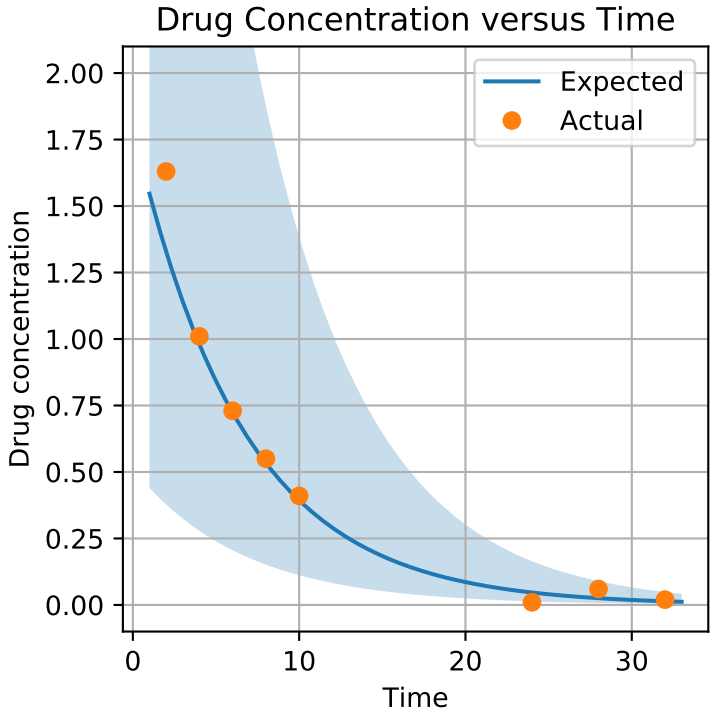


Figure 9: Parameters of Equation 2 were estimated with the MLE. The model is plotted with the data. The shaded region is the 95% confidence interval.

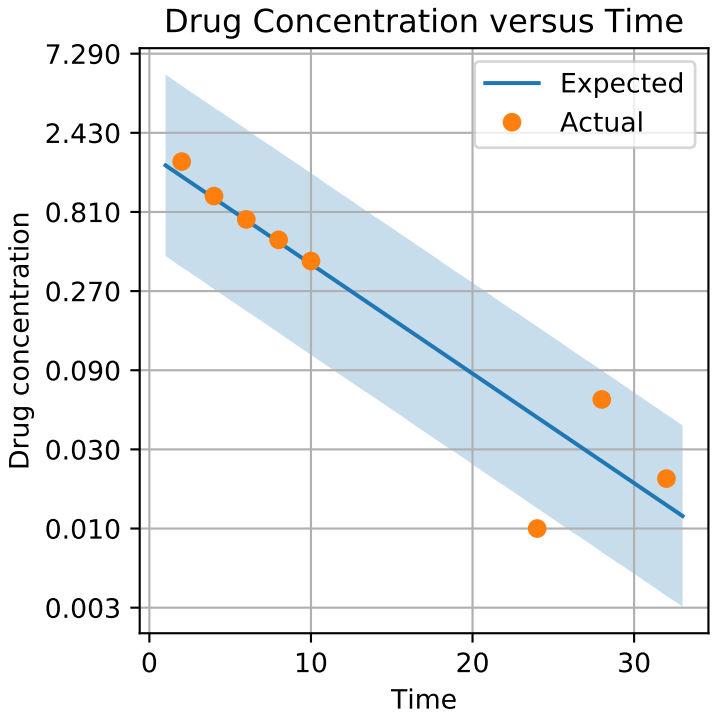


Figure 10: The same plot as Figure 9, but the with the y-axis scaled logarithmically.

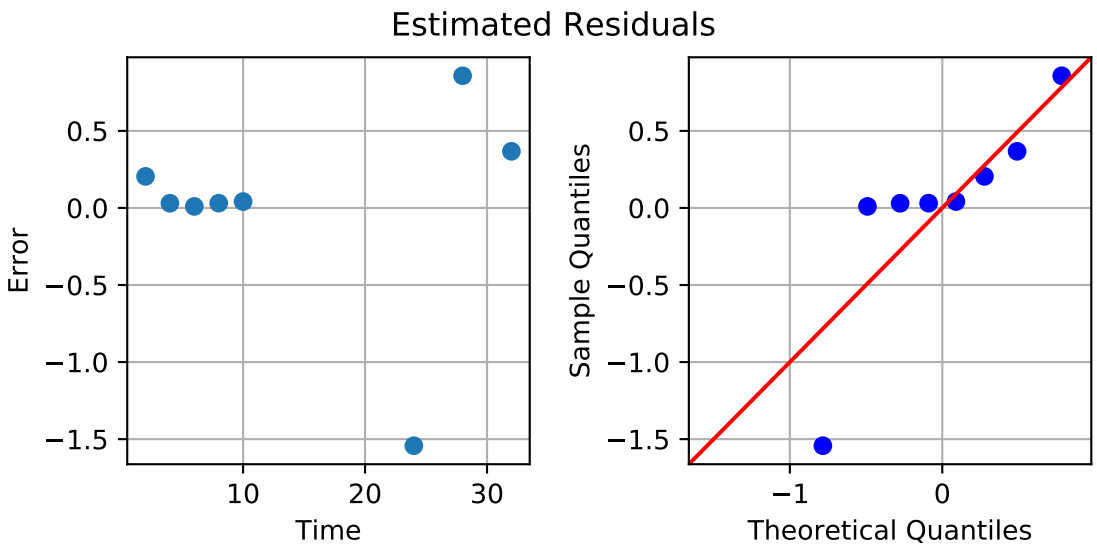


Figure 11: On the left, residuals are plotted against time. On the right, residuals are plotted in a Q-Q plot.

	Estimate	Standard error	95% CI lower bound	95% CI upper bound
Clearance (Cl)	2.534595	0.700000	1.162621	3.90657
Elimination half-life ($x_{1/2}$)	4.556990	0.614409	3.352771	5.76121

Table 6: Estimates and confidence intervals for parameters of interest.

Solution: Plots of the data along with the fitted curve and 95% confidence interval regions can be seen in Figures 9 and 10. In the first plot, the untransformed data is plotted. In the second, a log-scaling is applied to the drug concentration. Errors appear to increase with time.

(d) Using residuals, examine the appropriateness of the assumptions of the above model. Does the model seem reasonable for these data?

Solution: Residuals $\hat{\epsilon}_i, \ldots, \hat{\epsilon}_n$ are estimated with the MLE and plotted in Figure 11. The normal model for the residauls does not appear to be appropriate. In the Q-Q plot the lower-tailed residuals are far from the $y = x$ line, so the errors are not normal. From the plot against time, the residuals appear dependent on the time which violates the independence assumption.

(e) The clearance $Cl = V \times k_e$ and elimination half-life $x_{1/2} = \frac{\log 2}{k_e}$ are parameters of interest in this experiment. Find the MLEs of these parameters along with asymptotic 95% confidence intervals.

Solution: The MLE is invariant to reparameterization, so computing the MLEs is easy:

$$\begin{aligned} \hat{Cl} &= \hat{V} \times \hat{k}_e \\ \hat{x}_{1/2} &= \frac{\log 2}{\hat{k}_e}. \end{aligned} \tag{8}$$

Approximate variance can be computed with the delta method and previous estimated variances:

$$\hat{\mathrm{Var}}\left(\hat{Cl}\right) = \frac{\hat{\sigma}^2}{n\hat{\mathrm{Var}}\left(X\right)} \begin{pmatrix} \hat{k}_e & \hat{V} \end{pmatrix} \begin{pmatrix} \frac{\hat{V}^2}{n}\sum_{i=1}^{n} x_i^2 & -\hat{V}\bar{x} \\ -\hat{V}\bar{x} & 1 \end{pmatrix} \begin{pmatrix} \hat{k}_e \\ \hat{V} \end{pmatrix}.$$

$$\hat{\mathrm{Var}}\left(\hat{x}_{1/2}\right) = \frac{\hat{\sigma}^2}{n\hat{\mathrm{Var}}\left(X\right)} \left(\frac{\log 2}{\hat{k}_e^2}\right)^2. \tag{9}$$

The resulting MLE estimates and confidence intervals by applying Equations 8 and 9 can be found in Table 6.
Code to reproduce this analysis can be found in `cadralazine.ipynb`.