Coursework 2: STAT 570

Philip Pham

October 12, 2018

1. Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ i = 1, \ldots, n,$$

where the error terms ϵ_i are such that $\mathbb{E}[\epsilon_i] = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$, and $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$.

In the following you will consider $x_i \sim_{\text{iid}} \mathcal{N}(20, 3^2)$, with $\beta_0 = 2$ and $\beta_1 = -2.5$ and $n = 15, 30$.

Consider the model in Equation 1 with the error terms ϵ_i, independent and identically distributed, from the distributions:

- The normal distribution with mean 0 and variance 2^2.
- The uniform distribution on the range $(-r, r)$ for $r = 2$.
- A skew normal distribution with $\alpha = 5$, $\omega = 1$, and ξ chosen to given mean 0.

(a) What is the theoretical bias for $\hat{\beta}$ if the errors are of the form specified?

Solution: The theoretical bias for $\hat{\beta}$ is 0. Let

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \tag{1}$$

If we use the least squares estimate, we have

$$\begin{aligned} \hat{\beta} &= (X^\intercal X)^{-1} X^\intercal y \\ &= (X^\intercal X)^{-1} X^\intercal (X\beta + \epsilon) \\ &= \beta + (X^\intercal X)^{-1} X^\intercal \epsilon, \end{aligned} \tag{2}$$

Thus, using Equation 2 and linearity of expectations, we have

$$\boxed{\text{bias}\left(\hat{\beta}\right) = \mathbb{E}\left[\hat{\beta}\right] - \beta = \beta + (X^\intercal X)^{-1} X^\intercal \mathbb{E}[\epsilon] - \beta = 0.} \tag{3}$$

(b) Compare the variance of the estimator as reported by least squares, with that which follows from the sampling distribution of the estimator.

Solution: The variances reported by least squares are in the column titled **Least-squares Variance** in Tables 1a and 1b. This number is the variance reported by least squares averaged across the simulations.
To obtain the sample variance, 2,048 simulations were done. The results are reported in the column titled **Sample Variance** in Tables 1a and 1b.
The two variance estimates are nearly identical except when the errors have a t distribution.
Code for calculations can be found in `estimator_variance.ipynb`.

(c) Examine the distribution of the resultant estimators (across simulations) of β_0 and β_1, in particular with respect to normality. For each parameter find the coverage probability of a 95% confidence interval, that is the proportion of times that the confidence intervals contain the true value.

Solution: Several tests of normality were done. First, I checked how frequently the 95% confidence interval contains the true value of β_j. In the second-to-last column of Tables 1a and 1b, this was always close to 95% as one would expect.
I also performed the Shapiro-Wilk test, whose null hypothesis is that the estimates were drawn from a normal distribution. From the last column of Tables 1a and 1b, we can clearly reject the null hypothesis when the errors come from a t-distribution. For the other error distributions, the evidence is not as conclusive.
Finally, I did a qualitative evaluation by plotting the histogram against a fitted normal distribution. In Figures 1 and 2, we see the normal distribution accurately describes the data except when the errors are t-distributed.
Code for calculations and plots can be found in `estimator_variance.ipynb`.

(d) **Bonus:** Can you "break" least squares? i.e., find a distribution of the errors (with mean zero) that produces poor confidence interval coverage?

Solution: Yes, it is possible to "break" least squares. The Gauss-Markov theorem gives us the conditions under which the least squares estimate is the best linear unbiased estimator: (1) the errors have mean 0, (2) they are homoscedastic, and (3) they are uncorrelated.
The t-distribution with 3/2 degrees has infinite variance. If our errors are distributed in such a manner, from Tables 1a and 1b and Figures 1 and 2, our estimates no longer have a normal distribution.
In particular, we see that least squares poorly estimates the variance, and the estimates have the greatest error from the true values of β_j. Despite the non-normality and badly estimated variances, confidence interval coverage is actually quite good.

2. Consider the exponential regression problem with independent responses

$$p\left(y_i \mid \lambda_i\right) = \lambda_i \exp\left(-\lambda_i y_i\right), y_i > 0, \tag{4}$$

and $\log \lambda_i = \beta_0 + \beta_1 x_i$ for given covariates x_i, $i = 1, \ldots, n$. We wish to estimate the 2×1 regression parameter $\beta = \begin{pmatrix} \beta_0 & \beta_1 \end{pmatrix}^\intercal$ using maximum likelihood estimation (MLE).

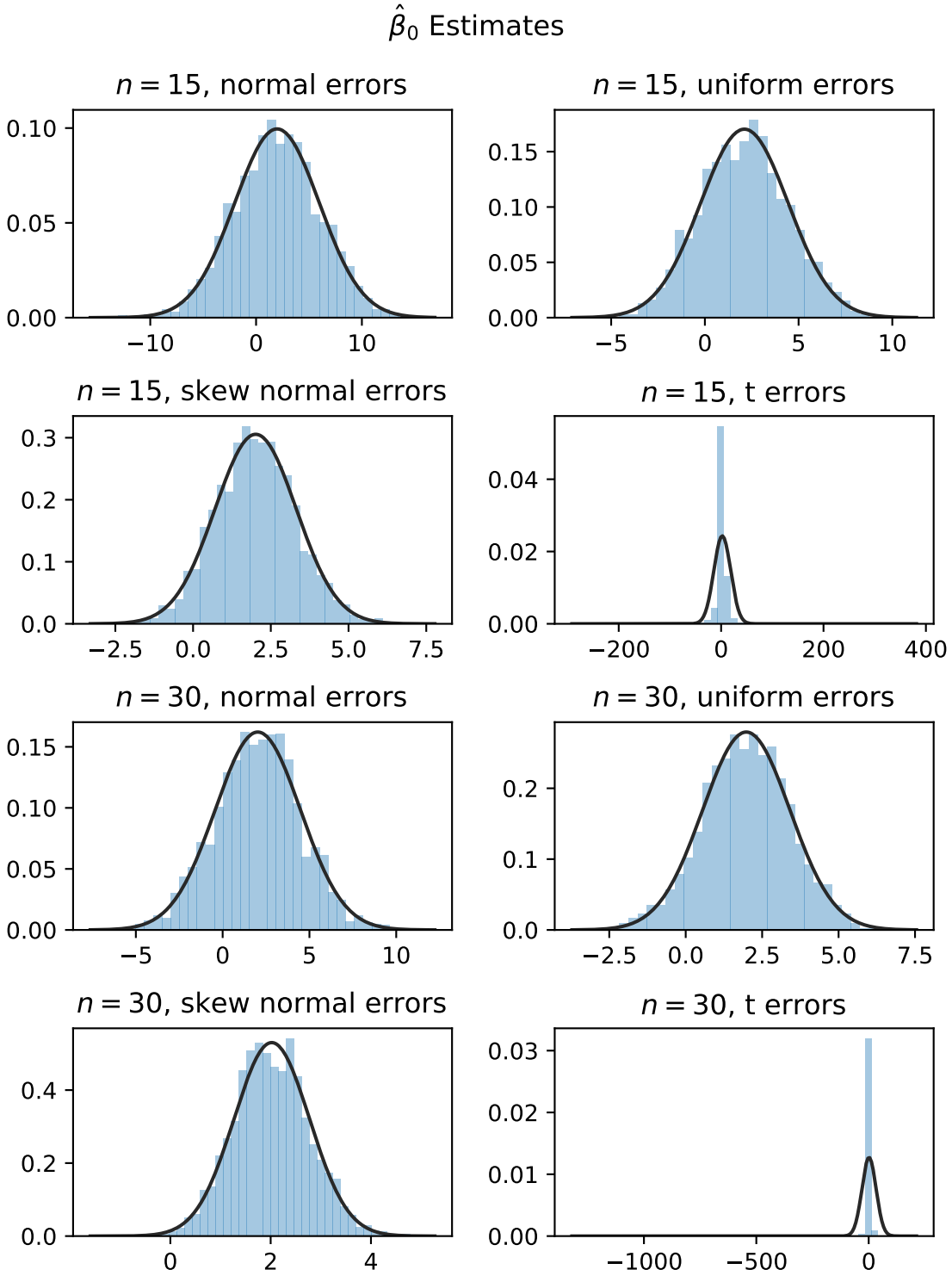


Figure 1: The distributions of $\hat{\beta}_0$ compared to a fitted normal.

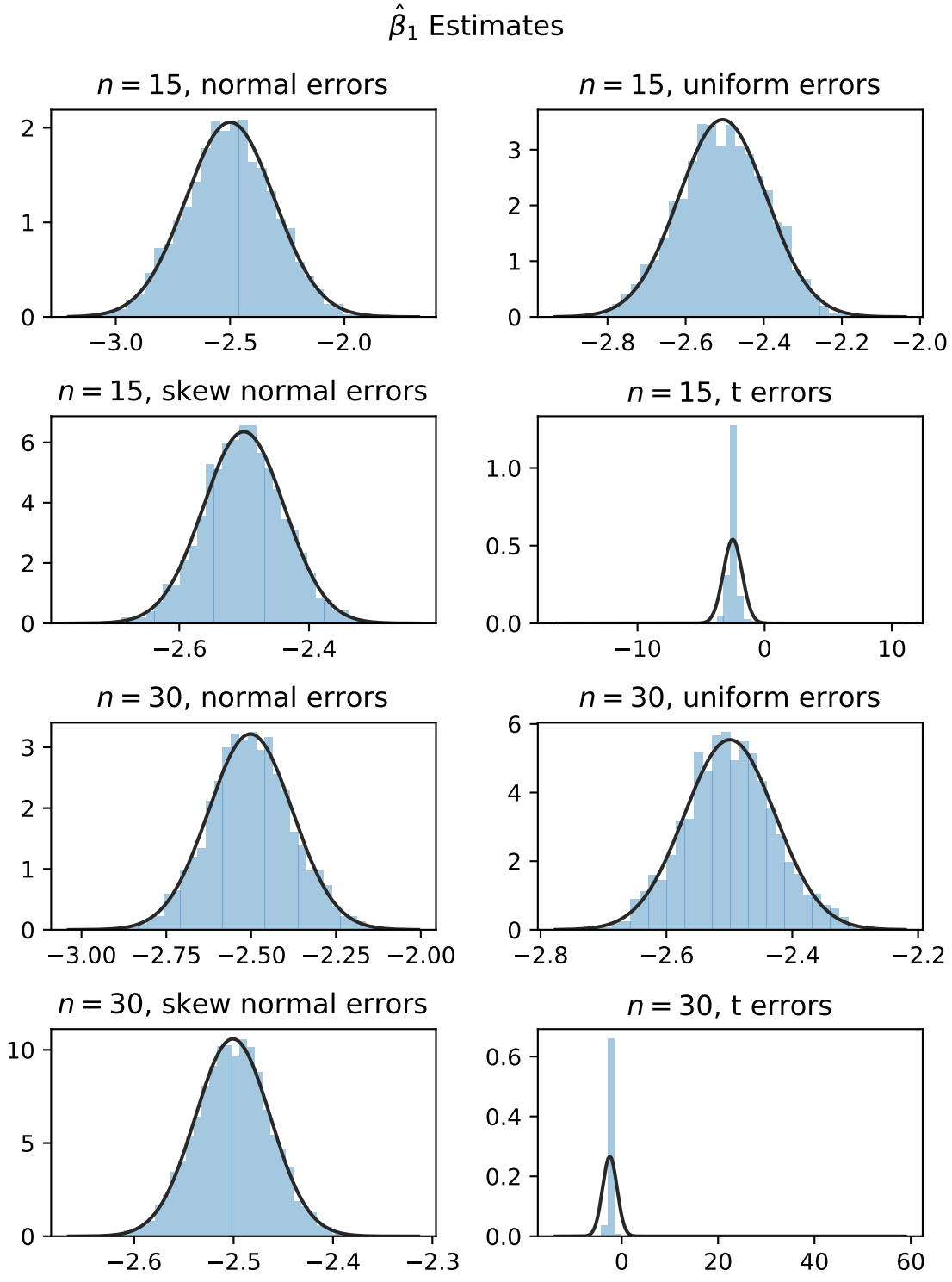


Figure 2: The distributions of $\hat{\beta}_1$ compared to a fitted normal.

n	Error Distribution	$\hat{\beta}_0$ Estimate	Sample Variance	Least-squares Variance	95% CI Coverage	Shapiro-Wilk p-value
15	normal	1.981029	16.066709	16.462905	0.954102	0.440320
15	uniform	2.108807	5.488093	5.457236	0.948730	0.052002
15	skew normal	2.013222	1.706046	1.624730	0.943359	0.558366
15	t	2.223518	268.857783	447.847807	0.950195	0.000000
30	normal	2.023642	6.058565	6.067812	0.949707	0.635323
30	uniform	1.980790	2.039209	2.001988	0.941895	0.307742
30	skew normal	2.021499	0.566532	0.591055	0.954590	0.099939
30	t	1.468913	981.048353	559.866048	0.950684	0.000000

(a) Simulations of for $\hat{\beta}_0$. Recall that $\beta_0 = 2$.

n	Error Distribution	$\hat{\beta}_1$ Estimate	Sample Variance	Least-squares Variance	95% CI Coverage	Shapiro-Wilk p-value
15	normal	-2.499933	0.037590	0.038154	0.953613	0.447169
15	uniform	-2.505455	0.012717	0.012647	0.948242	0.045104
15	skew normal	-2.500507	0.003944	0.003765	0.944336	0.907067
15	t	-2.512479	0.543700	1.037909	0.948730	0.000000
30	normal	-2.502034	0.015370	0.015198	0.949219	0.744967
30	uniform	-2.499262	0.005189	0.005014	0.936523	0.101658
30	skew normal	-2.500797	0.001421	0.001480	0.956055	0.850093
30	t	-2.475815	2.229763	1.402287	0.950684	0.000000

(b) Simulations of for $\hat{\beta}_1$. Recall that $\beta_1 = -2.5$.

Table 1: Results for different sample sizes and error distributions.

(a) Find expressions for the likelihood function $L(\beta)$, log-likelihood function $l(\beta)$, score function $S(\beta)$, and Fisher's information matrix $I(\beta)$.

Solution: We can rewrite Equation 4 in terms of β, which gives us

$$\begin{aligned} p(y_i \mid \beta_0, \beta_1) &= \exp(\beta_0 + \beta_1 x_i) \exp(-y_i \exp(\beta_0 + \beta_1 x_i)) \\ &= \exp(\beta_0 + \beta_1 x_i - y_i \exp(\beta_0 + \beta_1 x_i)). \end{aligned} \tag{5}$$

Using Equation 5, we can write the likelihood function

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{n} p(y_i \mid \beta_0, \beta_1) \\ &= \exp\left(n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i \exp(\beta_0 + \beta_1 x_i)\right). \end{aligned} \tag{6}$$

Taking the log of Equation 6, we have the log-likelihood function as

$$\begin{aligned} l(\beta) &= \log L(\beta) \\ &= n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i \exp(\beta_0 + \beta_1 x_i). \end{aligned} \tag{7}$$

Taking the gradient of Equation 7, we have the score function

$$\begin{aligned} S(\beta) &= \nabla l(\beta) \\ &= \begin{pmatrix} n - \sum_{i=1}^{n} y_i \exp(\beta_0 + \beta_1 x_i) \\ \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_i \exp(\beta_0 + \beta_1 x_i) \end{pmatrix}. \end{aligned} \tag{8}$$

One definition of the Fisher information is the expected value of the observed information which is the negative of the second derivative of the log-likelihood

function. For a single observation,

$$\begin{aligned}\mathcal{I}_1(\beta) &= \mathbb{E}\left[\begin{pmatrix} Y\exp(\beta_0+\beta_1 x_i) & x_i Y\exp(\beta_0+\beta_1 x_i) \\ x_i Y\exp(\beta_0+\beta_1 x_i) & x_i^2 Y\exp(\beta_0+\beta_1 x_i)\end{pmatrix} \mid X = x_i\right] \\ &= \frac{1}{\exp(\beta_0+\beta_1 x_i)}\begin{pmatrix} \exp(\beta_0+\beta_1 x_i) & x_i\exp(\beta_0+\beta_1 x_i) \\ x_i\exp(\beta_0+\beta_1 x_i) & x_i^2\exp(\beta_0+\beta_1 x_i)\end{pmatrix} \\ &= \begin{pmatrix} 1 & x_i \\ x_i & x_i^2\end{pmatrix} \end{aligned} \tag{9}$$

by properties of the exponential distribution. Thus, Fisher information is

$$\mathcal{I}_n(\beta) = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2\end{pmatrix}. \tag{10}$$

(b) Find expressions for the maximum likelihood estimate $\hat{\beta}$. If no closed form solution exists, then instead provide a functional form that could be simply implemented for solution.

Solution: We can solve for $\hat{\beta}_0$ in terms of $\hat{\beta}_1$. We know that $S\left(\hat{\beta}\right) = \mathbf{0}$.

From Equation 8, we can solve for $\hat{\beta}_0$,

$$\hat{\beta}_0 = \log n - \log\sum_{i=1}^n y_i \exp\left(\hat{\beta}_1 x_i\right). \tag{11}$$

Substituing Equation 11 into the second entry of Equation 8, we have

$$\begin{aligned} 0 &= \sum_{i=1}^n x_i - \exp\left(\hat{\beta}_0\right)\sum_{i=1}^n x_i y_i \exp\left(\hat{\beta}_1 x_i\right) \\ &= \sum_{i=1}^n x_i - \frac{n}{\sum_{i=1}^n y_i \exp\left(\hat{\beta}_1 x_i\right)}\sum_{i=1}^n x_i y_i \exp\left(\hat{\beta}_1 x_i\right), \end{aligned} \tag{12}$$

which we can solve numerically with a root-finding algorithm.

(c) For the data in Table 2, numerically maximize the likelihood function to obtain estimates of β. These data consist of the survival times (y) of rats as function of concentration of a contaminant (x). Find the asymptotic covariance matrix for your estimate using the information $\mathcal{I}(\beta)$. Provide a 95% confidence interval for each element of β_0 and β_1.

Solution: Numerically solving Equations 11 and 12, we have that

$$\hat{\beta} = \begin{pmatrix}\hat{\beta}_0 \\ \hat{\beta}_1\end{pmatrix} = \begin{pmatrix}-2.821150253077923 \\ 0.30133576292327585\end{pmatrix}. \tag{13}$$

The Fisher information gives a lower bound on the variance according to the Cramér-Rao bound. Asymptotic normality of the MLE tells us that

$$\hat{\beta}_n - \beta \to \mathcal{N}\left(0, \mathcal{I}_n^{-1}(\beta)\right)$$

in distribution.

i	x_i	y_i
1	6.1	0.8
2	4.2	3.5
3	0.5	12.4
4	8.8	1.1
5	1.5	8.9
6	9.2	2.4
7	8.5	0.1
8	8.7	0.4
9	6.7	3.5
10	6.5	8.3
11	6.3	2.6
12	6.7	1.5
13	0.2	16.6
14	8.7	0.1
15	7.5	1.3

Table 2: Each observation is a rat. x_i are the concentrations of the contaminant, and y_i are the survival times.

Thus, we have the covariance matrix

$$\operatorname{Var}\left(\hat{\beta}\right) \approx \begin{pmatrix} 15 & 90.1 \\ 90.1 & 671.07 \end{pmatrix}^{-1} = \begin{pmatrix} 0.34448471 & -0.04625162 \\ -0.04625162 & 0.00770005 \end{pmatrix}. \quad (14)$$

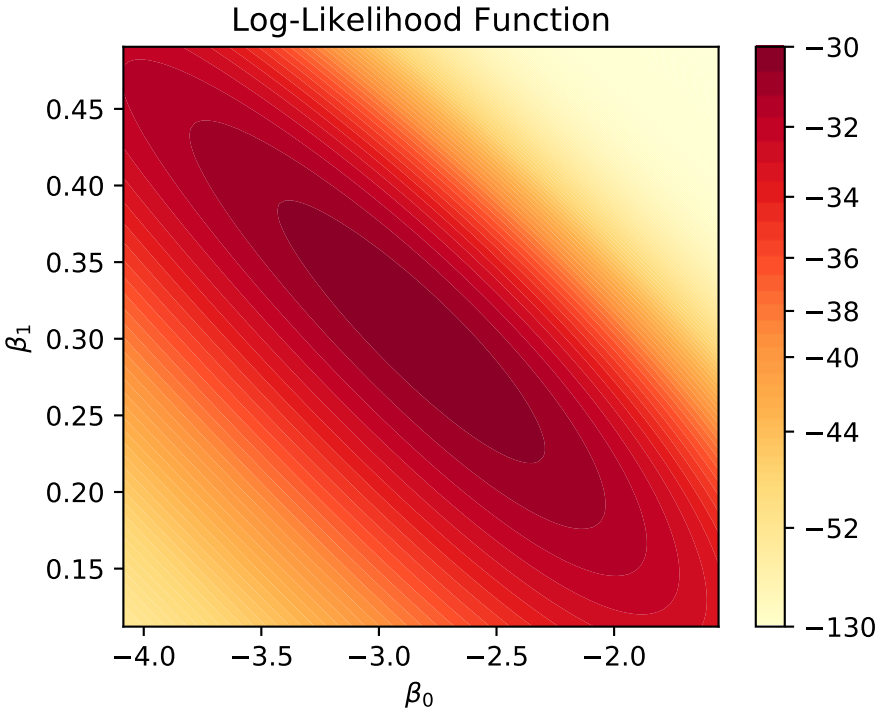
Using this we can approximate 95% confidence intervals as $\hat{\beta}_j \pm z_{0.975}\sqrt{\operatorname{Var}\left(\hat{\beta}_j\right)}$, where $z_p = \Phi^{-1}(p)$ and Φ is the cumulative distribution function of the normal distribution.
We have the confidence intervals

$$\begin{aligned} \left(\hat{\beta}_0 - 1.150358, \hat{\beta}_0 + 1.150358\right) &= (-3.97150839, -1.67079212) \\ \left(\hat{\beta}_1 - 0.171986669, \hat{\beta}_1 + 0.171986669\right) &= (0.12934909, 0.47332243) \end{aligned} \quad (15)$$

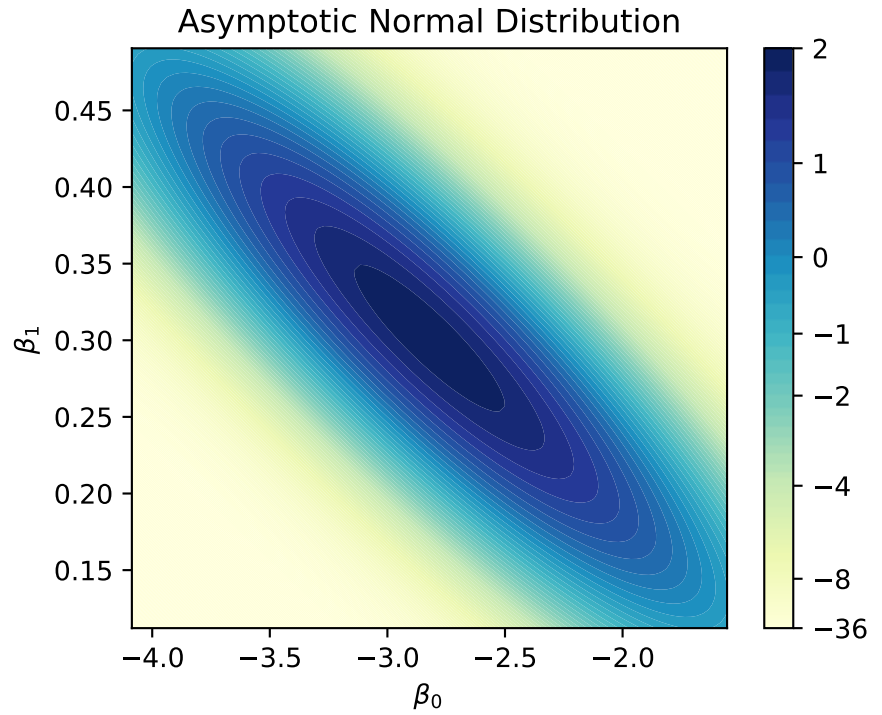
for $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively.
All calcuations can be found in `exponential_regression.ipynb`.

(d) Plot the log-likelihood function $l(\beta_0, \beta_1)$ and compare with the log of the asymptotic normal approximation to the sampling distribution of the MLE.

Solution: The two plots can be found in Figure 3.
The log-likelihood function can be found in Figure 3a, and the asymptotic normal approximation is plotted in Figure 3b.
The distributions are similar, and the asymptotic normal approximation accurately captures the covariance structure of $\hat{\beta}$. The asymptotic normal approximation, however, does underestimate the variance. This is unsurprising since the inverse of the Fisher information matrix is a lower bound on variance according to the Cramér-Rao bound.
Code for the plots can be found in `exponential_regression.ipynb`.



(a) The log-likelihood function is centered at the MLE estimate.



(b) The asymptotic normal approximation mirrors the log-likelihood with slightly less variance.

Figure 3: Plots of the distributions of $\hat{\beta}$ for Part 2d.

(e) Summarize the results of the estimation presented above in a manner that would address the question of whether increasing concentrations of the contaminant had an effect on a rat's life expectancy.

Solution: λ_i can be viewed as the rate of death, so the mean survival time is λ_i^{-1}. $\log \lambda_i$ has a linear relationship with concentrations of the contaminant described by β_1. $\beta_1 > 0$ indicates that increasing concentrations of the contaminant increase death rates, and therefore, decrease life expectancy. Indeed from Equation 13, our estimate of β_1, $\hat{\beta}_1 = 0.30133576292327585 > 0$. From Equation 15, we see that the 95% confidence interval lies entirely on the positive half-line, so the result is statistically significant at level 0.05. Thus, it is quite likely that increasing concentrations of the contaminant decrease life expectancy.