

Statistics for Data Science - Exercise Set 12 - Solution

1. Attached is a data file `mrfit.csv` and an R code file `logistic.r` for running a logistic regression analysis on the data.

a. Compute a 95% confidence interval for the coefficient associated with the variable `diab`.

b. Compute a 95% confidence interval for $p(x) = P(Y = 1|X = x)$ for the following configuration:

`age = 50`

`diab = 0`

`chol = 200`

`map = 100`

`smoke = 1`

Solution

a. The formula for the asymptotic confidence interval for a logistic regression coefficient is

$$\hat{\beta}_j \pm z_{1-\alpha/2} \text{SE}(\hat{\beta}_j)$$

where

$$\text{SE}(\hat{\beta}_j) = \text{Var}(\hat{\beta}_j)^{1/2} = \hat{V}_{jj}^{1/2}, \quad \hat{V} = (X^T \hat{W} X)^{-1}$$

From the output we have $\hat{\beta}_j = 0.6539$ and $\text{SE}(\hat{\beta}_j) = 0.1182$. The critical value is $z_{0.975} = 1.9600$. From here, the rest is simple arithmetic, and the confidence interval obtained is $[0.4222, 0.8856]$.

b. We have $p(x) = g(a^T \beta)$, where $a = [1, 50, 0, 200, 100, 1]^T$ and $g(u) = e^u / (1 + e^u)$. The point estimate for $p(x)$ is $\hat{p}(x) = g(a^T \hat{\beta}) = 0.2316$. For the confidence interval, we start, as explained in class, by forming a confidence interval for $\psi = a^T \beta$. This is given by

$$a^T \hat{\beta} \pm z_{1-\alpha/2} (a^T \hat{V} a)^{1/2}$$

where \hat{V} is as defined above. The matrix \hat{V} is extracted using the command `vcov`. We get the confidence interval $[\psi_L, \psi_U]$. The confidence interval for $p(x)$ is then given by $[p_L, p_U]$ where $p_L = g(\psi_L)$ and $p_U = g(\psi_U)$. Below is R code for carrying out the calculation:

```
zcrit = qnorm(0.975)
expit = function(u) exp(u)/(1+exp(u))
a = c(1, 50, 0, 200, 100, 1)
lr1 = glm(died10 ~ age + diab + chol + map + smoke, family=binomial(link=logit))
print(summary(lr1))
lammat = vcov(lr1)
print(lammat)
beta = coef(lr1)
psihat = sum(a*beta)
print(psihat)
```

```

phat = expit(psihat)
se_psi = sqrt(t(a) %*% lammat %*% a)
cihw_psi = zcrit*se_psi
psi_lo = psihat - cihw_psi
psi_hi = psihat + cihw_psi
p_lo = expit(psi_lo)
p_hi = expit(psi_hi)
print(cbind(p_lo,p_hi))

```

The resulting confidence interval is [0.2117, 0.2527].

2. Let Y_1, \dots, Y_n iid $N(\mu, \sigma^2)$ with σ^2 known, and consider Bayesian inference for μ with the prior $N(\mu_0, \tau^2)$. We stated in class that the posterior distribution of μ is $N(\bar{\mu}, v)$, where

$$\begin{aligned}\bar{\mu} &= w\bar{Y} + (1-w)\mu_0 \\ w &= \left(\frac{n}{\sigma^2}\right) / \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right) \\ v &= \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}\end{aligned}$$

Present the details of the derivation of this result.

Solution

In the following, we will use the notation C to denote a constant that does not depend on μ , where that constant may change from line to line.

We have

$$\begin{aligned}\pi(\mu) &= (2\pi\tau^2)^{-1/2} \exp\left(-\frac{1}{2\tau^2}(\mu - \mu_0)^2\right) \\ f_{Y|\mu}(Y|\mu) &= \prod_{i=1}^n \left[(2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}(Y_i - \mu)^2\right) \right] \\ &= C \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \mu)^2\right) \\ \pi(\mu)f_{Y|\mu}(Y|\mu) &= C \exp\left(-\frac{1}{2\tau^2}(\mu - \mu_0)^2\right) \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \mu)^2\right)\end{aligned}$$

Now, we can write

$$\begin{aligned}
\sum_{i=1}^n (Y_i - \mu)^2 &= \sum_{i=1}^n [(Y_i - \bar{Y}) - (\mu - \bar{Y})]^2 \\
&= \sum_{i=1}^n (Y_i - \bar{Y})^2 - 2 \sum_{i=1}^n (Y_i - \bar{Y})(\mu - \bar{Y}) + n(\mu - \bar{Y})^2 \\
&= \sum_{i=1}^n (Y_i - \bar{Y})^2 - 2(\mu - \bar{Y}) \sum_{i=1}^n (Y_i - \bar{Y}) + n(\mu - \bar{Y})^2 \\
&= \sum_{i=1}^n (Y_i - \bar{Y})^2 + n(\mu - \bar{Y})^2 \\
&= C + n(\mu - \bar{Y})^2
\end{aligned}$$

where in the second to last step we used the fact that $\sum_{i=1}^n (Y_i - \bar{Y}) = 0$. We can thus write

$$\begin{aligned}
\pi(\mu) f_{Y|\mu}(Y|\mu) &= C \exp\left(-\frac{1}{2\tau^2}(\mu - \mu_0)^2\right) \exp\left(-\frac{n}{2\sigma^2}(\bar{Y} - \mu)^2\right) \\
&= C \exp\left[-\frac{1}{2}\left(\frac{1}{\tau^2}(\mu - \mu_0)^2 + \frac{n}{\sigma^2}(\bar{Y} - \mu)^2\right)\right]
\end{aligned}$$

We now expand the term in parentheses in the exponent:

$$\begin{aligned}
\frac{1}{\tau^2}(\mu - \mu_0)^2 + \frac{n}{\sigma^2}(\bar{Y} - \mu)^2 &= \frac{1}{\tau^2}(\mu^2 - 2\mu\mu_0 + \mu_0^2) + \frac{n}{\sigma^2}(\bar{Y}^2 - 2\bar{Y}\mu + \mu^2) \\
&= C + \left(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right)\mu^2 - 2\left(\frac{1}{\tau^2}\mu_0 + \frac{n}{\sigma^2}\bar{Y}\right)\mu \\
&= C + \left(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right)\left[\mu^2 - 2\frac{\left(\frac{1}{\tau^2}\mu_0 + \frac{n}{\sigma^2}\bar{Y}\right)}{\left(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right)}\mu\right] \\
&= C + v^{-1}(\mu^2 - 2\bar{\mu}\mu) \\
&= C + v^{-1}(\mu^2 - 2\bar{\mu}\mu + \bar{\mu}^2) \\
&= C + v^{-1}(\mu - \bar{\mu})^2
\end{aligned}$$

Putting this back into the exponent, we get

$$\pi(\mu) f_{Y|\mu}(Y|\mu) = C \exp\left(-\frac{1}{2v}(\mu - \bar{\mu})^2\right)$$

This is equal, up to a constant factor, to the $N(\bar{\mu}, v)$ density, so we have now arrived at the desired conclusion.

3. Let Y_1, \dots, Y_n be iid $Ber(\theta)$. In class we discussed Bayesian inference for θ using the Beta prior $\pi(\theta) = B(\alpha, \beta)^{-1}\theta^{\alpha-1}(1-\theta)^{\beta-1}$. Here, we will consider an alternate prior, under which $\theta = g(W)$, where $W \sim N(\mu, \sigma^2)$ and $g(w) = e^w/(e^w + 1)$. Using the formula for the density of a transformed random variable with a monotone transformation, the prior density is given by

$$\pi(\theta) = f_W(h(\theta))|h'(\theta)|$$

where $h(\theta) = \log \theta - \log(1 - \theta)$ is the inverse of g , and we obtain

$$\pi(\theta) = \begin{cases} \sigma^{-1}[\theta(1-\theta)]^{-1}\phi(\sigma^{-1}(h(\theta) - \mu)) & \theta \in (0, 1) \\ 0 & \theta = 0 \text{ or } \theta = 1 \end{cases}$$

where ϕ is the $N(0, 1)$ density and $h(\theta) = \log \theta - \log(1 - \theta)$. Suppose $n = 20$ and $\sum_{i=1}^n Y_i = 4$. Take $\mu = 0.25$ and $\sigma^2 = 1$.

- Produce a plot of the posterior density.
- Compute $P(\theta \leq 0.35|Y)$.
- Using the Newton-Raphson algorithm, find the value θ_L that satisfies $P(\theta \leq \theta_L|Y) = 0.05$ and the value θ_U such that $P(\theta \geq \theta_U|Y) = 0.05$. This yields a 90% equal-tail (ET) Bayesian credible interval for θ .
- Compute $E[\theta|Y]$, which is the Bayes estimate of θ under squared error loss.

Note: In this question you will need to use numerical integration to compute integrals of the form

$$\int_0^a \psi(\theta') d\theta', \quad a \in [0, 1]$$

where ψ is a complicated function. The topic of numerical integration is a very big topic, but for purposes of this question, you should use the R function `integrate`.

Solution

See the R code `ex12q3.r`

- Suppose that Y_1, \dots, Y_n are iid $N(0, \theta)$.
- Derive the Jeffreys prior for θ .
- Using the Jeffreys prior, derive an expression for the posterior distribution of θ .

Solution

The Jeffreys prior is given by $\pi(\theta) = c\mathcal{I}(\theta)^{1/2}$, where $\mathcal{I}(\theta)$ is the Fisher information. We have

$$\begin{aligned} \log f(Y|\theta) &= -\frac{1}{2} \log 2\pi - \frac{1}{2} \log \theta - \frac{1}{2\theta} Y^2 \\ \frac{d}{d\theta} \log f(Y|\theta) &= -\frac{1}{2\theta} + \frac{1}{2\theta^2} Y^2 \\ \frac{d^2}{d\theta^2} \log f(Y|\theta) &= \frac{1}{2\theta^2} - \frac{1}{\theta^3} Y^2 \\ \mathcal{I}(\theta) &= -E \left[\frac{d^2}{d\theta^2} \log f(Y|\theta) \right] \\ &= -E \left[\frac{1}{2\theta^2} - \frac{1}{\theta^3} Y^2 \right] \\ &= -\frac{1}{2\theta^2} + \frac{1}{\theta^3} E[Y^2] \\ &= -\frac{1}{2\theta^2} + \frac{1}{\theta^3} \theta \\ &= \frac{1}{2\theta^2} \end{aligned}$$

So the Jeffreys prior is $\pi(\theta) = c'\theta^{-1}$. This is an improper prior because

$$\int_0^\infty \theta^{-1} d\theta = \log \theta \Big|_0^\infty = \infty$$

However, as explained in class, we can proceed, and we will be OK provided that the posterior density is a legitimate probability density.

Define $\mathcal{Y} = (Y_1, \dots, Y_n)$. We have

$$f(\mathcal{Y}|\theta) = (2\pi\theta)^{-n/2} \exp\left(-\frac{1}{2\theta} \sum_{i=1}^n Y_i^2\right) = (2\pi\theta)^{-n/2} \exp\left(-\frac{S}{2\theta}\right)$$

where

$$S = \sum_{i=1}^n Y_i^2$$

So we have

$$\pi(\theta)f(\mathcal{Y}|\theta) = c''\theta^{-(n/2+1)} \exp\left(-\frac{S}{2\theta}\right)$$

Thus, the posterior density is given by

$$p(\theta|\mathcal{Y}) = \theta^{-(n/2+1)} \exp\left(-\frac{S}{2\theta}\right) \Big/ \int_0^\infty \tilde{\theta}^{-(n/2+1)} \exp\left(-\frac{S}{2\tilde{\theta}}\right) d\tilde{\theta}$$

The integral in the denominator does not have an analytic solution, but it can be evaluated numerically for given n and S .

5. Consider the setup where we have Y_1, \dots, Y_n are iid $f_{Y|\theta}(y|\theta)$. The Bernstein - von Mises theorem says that, under suitable conditions, the posterior distribution of θ can be approximated for large n by the distribution $N(\hat{\theta}_n^{(MLE)}, n^{-1}\mathcal{I}(\hat{\theta}_n^{(MLE)})^{-1})$, where $\hat{\theta}_n^{(MLE)}$ is the maximum likelihood estimator of θ and $\mathcal{I}(\theta)$ is the Fisher information. A consequence is that for large n the $100(1 - \gamma)\%$ Bayesian credible interval is approximately the same as the $100(1 - \gamma)\%$ confidence interval. For the case where $Y_i \sim Ber(\theta)$ and the prior distribution is the $Beta(\alpha, \beta)$ distribution, we saw in class that the posterior distribution of θ is the $Beta(\alpha + n\bar{Y}, \beta + n(1 - \bar{Y}))$ distribution. Taking $\bar{Y} = \hat{\theta}_n^{(MLE)} = 0.35$, plot together on the same graph the $Beta(1 + n\bar{Y}, 1 + n(1 - \bar{Y}))$ density and the $N(\hat{\theta}_n^{(MLE)}, n^{-1}\mathcal{I}(\hat{\theta}_n^{(MLE)})^{-1})$ density for $n = 10$, $n = 20$, $n = 30$, and $n = 40$. The Fisher information for the Bernoulli distribution is derived on page 130 of Wasserman's book. Comment on the results.

Solution

Below is the R code for carrying out the calculation. We see that as n increases, the degree of similarity between the two curves increases.

```
ybar = 0.35
n = 10
alf = 1 + n*ybar
bet = 1 + n*(1-ybar)
se = sqrt(ybar*(1-ybar)/n)
```

```
th = seq(0,1,0.01)
d1 = dbeta(th,alf,bet)
d2 = dnorm(th,ybar,se)
plot(th,d1,type='l',col='red')
lines(th,d2,col='green')
title('Posterior Density Plot - Red is Exact and Green is Asymptotic
Approximation')
```