Applied Statistics — Assessed Coursework

06049216

Contents

1 Preface

I, 06049216, certify that this assessed coursework is my own work, unless otherwise acknowledged, and includes no plagiarism. I have not discussed my coursework with anyone else except when seeking clarification with the module lecturer via email or on MS Teams. I have not shared any code underlying my coursework with anyone else prior to submission.

Any text marked with a dagger† refers the marker to find the derivation in supplementary material.

2 Question 1

Firstly, we can consider the possibility that participants know each other in a social setting. Christakis and Fowler (2008) showed that people tend to quit smoking in groups, and the same authors found in an earlier study that people are more like to suffer from obesity if those around them are overweight (Christakis and Fowler (2007)). The assumption of independence would imply that if an individual put on weight this would have no effect on other participants but, assuming this hypothetical scenario, it would clearly be violated.

Secondly we may think about a larger scale, people who live in the same city/area. People who live in the same city could be subject to environmental changes which could affect lung capacity. As well as this, broad trends in the region could affect smoking or weight, it can certainly be argued that there is a spatial aspect to the availability of weight loss drugs or changes in wider perception of smoking. We can show how the assumption is violated through an example, say three of the survey respondents live in Los Angeles, where they are prone to wildfires. If we observe two of these said participants having a very low lung capacity, it would be clearly untrue to say this isn't at all correlated with the lung capacity of the third.

The third case we will consider is genetic. If a sampling method such as snowball sampling was used, it is entirely feasible that two of the members are directly related. The two participants will likely exhibit

dependence in lung capacity, and entirely possible the same holds for weight and cigarettes smoked.

3 Question 2

Part a.

Replacing $\eta_{1:N}$ with $S_{1:N}$ would represent a loss of information. We are losing information on the intensity of smoking, in other words the transformation of a quantity $\eta_i \in \{0, 1, \dots, \}$ into a binary $S_i \in \{0, 1\}$ is many-to-one; in the case of $S_i = 1$ the original η_i value is impossible to recover.

Part b.

We can write out the likelihood $\mathbb{P}(C_{1:N}|p, q, S_{1:N})$ as:

$$\mathbb{P}(C_{1:N}|p, q, S_{1:N}) = \prod_{n=1}^{N} \mathbb{P}(C_n|S_n) = \prod_{n=1}^{N} \delta_{S_n,S}\, \mathbb{P}(C_n|S_n = S) + \delta_{S_n,\neg S}\, \mathbb{P}(C_n|S_n = \neg S)$$

Since both can be expressed as a product of independent Bernoulli pmfs with probabilities p, q respectively:

$$\mathbb{P}(C_{1:N}|p, q, S_{1:N}) = p^{N_{s,c}}(1-p)^{N_s - N_{s,c}} \cdot q^{N_{\neg s,c}}(1-q)^{N_{\neg s} - N_{\neg s,c}}$$

Looking at the second likelihood given by the question as $\mathbb{P}(N_{s,c}, N_{\neg s,c}|p, q, N_s, N_{\neg s})$:

$$\mathbb{P}(N_{s,c}, N_{\neg s,c} \mid p, q, N_s, N_{\neg s}) = \binom{N_s}{N_{s,c}} p^{N_{s,c}}(1-p)^{N_s - N_{s,c}} \cdot \binom{N_{\neg s}}{N_{\neg s,c}} q^{N_{\neg s,c}}(1-q)^{N_{\neg s} - N_{\neg s,c}}$$

$$\frac{\mathbb{P}(N_{s,c}, N_{\neg s,c} \mid p, q, N_s, N_{\neg s})}{\mathbb{P}(C_{1:N}|p, q, S_{1:N})} = \binom{N_s}{N_{s,c}} \binom{N_{\neg s}}{N_{\neg s,c}}$$

From the fact that the two expressions are only seperated by a coefficient not involving p, q, we can conclude that all the information encoded in the full likelihood can be expressed through the jointly sufficient statistics present: $\{N_s, N_{s,c}, N_{\neg s}, N_{\neg s,c}\}$. In other words, p and q rely on the data only through these statistics.

Part c.

We can define the following uniform priors:

$$\mathbb{P}(p|M_0) = 1, \quad \mathbb{P}(q|M_0) = 1, \quad \mathbb{P}(p, q|M_0) = 1 \cdot 1$$

Here, M_0 represents the assumption that we are choosing a uniform prior.

From this, via Bayes theorem, we can define the posterior:

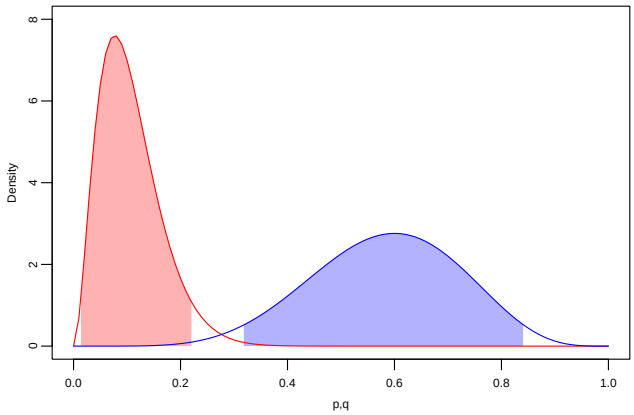
$$\mathbb{P}(p, q|C_{1:N}, S_{1:N}, M_0) \propto \mathbb{P}(C_{1:N}|p, q, S_{1:N})$$

$$\mathbb{P}(p, q|C_{1:N}, S_{1:N}, M_0) \propto p^{N_{s,c}}(1-p)^{N_s - N_{s,c}} q^{N_{\neg s,c}}(1-q)^{N_{\neg s} - N_{\neg s,c}}$$

$$\mathbb{P}(p|C_{1:N}, S_{1:N}, M_0) \propto \int_0^1 \mathbb{P}(p, q|C_{1:N}, S_{1:N}) dq, \qquad \mathbb{P}(q|C_{1:N}, S_{1:N}, M_0) \propto \int_0^1 \mathbb{P}(p, q|C_{1:N}, S_{1:N}) dp$$

$$\mathbb{P}(p|C_{1:N}, S_{1:N}, M_0) \propto p^{N_{s,c}}(1-p)^{N_s - N_{s,c}}, \qquad \mathbb{P}(q|C_{1:N}, S_{1:N}, M_0) \propto q^{N_{\neg s,c}}(1-q)^{N_{\neg s} - N_{\neg s,c}}$$

We can recognise these as $\text{Beta}(N_{s,c}+1, N_s - N_{s,c}+1)$ and $\text{Beta}(N_{\neg s,c}+1, N_{\neg s} - N_{\neg s,c}+1)$ respectively, and taking the respective parameters values from the `smoking` dataset, we can view the density of two posteriors (non-smokers in red and smokers in blue). The shaded ares are the 95% credible intervals.



We can see no overlap whatsoever between the two 95% credible intervals, indicating strong evidence that $p > q$, that smoking is linked to lung cancer. More quantitatively, since we can actually sample from each posterior, we can get a Monte-Carlo approximation to the posterior probability $\mathbb{P}(p > q|N_{s,c}, N_s, N_{\neg s,c}, N_{\neg s})$ by taking 100,000 samples of each Beta distribution and calculating the proportion. This gives $\mathbb{P}(p > q|N_{s,c}, N_s, N_{\neg s,c}, N_{\neg s}, M_0) \approx 0.999^{\dagger}$. Given our data, there is very strong evidence that $p > q$.

Part d.

First considering M_1:

$$\mathbb{P}(N_{s,c},\, N_{\neg s,c} \mid N_s,\, N_{\neg s},\, \mathcal{M}_1) = \int_P \int_Q \mathbb{P}(N_{s,c}, N_{\neg s,c} \mid p, q, N_s, N_{\neg s}) \mathbb{P}(p, q|M_1)\, dq\, dp$$

$$\mathbb{P}(N_{s,c},\, N_{\neg s,c} \mid N_s,\, N_{\neg s},\, \mathcal{M}_1) = 2\binom{N_s}{N_{s,c}}\binom{N_{\neg s}}{N_{\neg s,c}} \int_0^1 \int_q^1 p^{N_{s,c}}(1-p)^{N_s - N_{s,c}} q^{N_{\neg s,c}}(1-q)^{N_{\neg s} - N_{\neg s,c}}\, dp\, dq$$

For M_2:

$$\mathbb{P}(N_{s,c},\, N_{\neg s,c} \mid N_s,\, N_{\neg s},\, \mathcal{M}_2) = \binom{N_s}{N_{s,c}}\binom{N_{\neg s}}{N_{\neg s,c}} \int_0^1 \int_0^1 p^{N_{s,c}}(1-p)^{N_s - N_{s,c}} q^{N_{\neg s,c}}(1-q)^{N_{\neg s} - N_{\neg s,c}} \delta(p-q)\, dp\, dq$$

Since $\int_0^1 f(x)\delta(y-x)dx = f(y)$:

$$\mathbb{P}(N_{s,c},\, N_{\neg s,c} \mid N_s,\, N_{\neg s},\, \mathcal{M}_2) = \binom{N_s}{N_{s,c}}\binom{N_{\neg s}}{N_{\neg s,c}} \int_0^1 q^{N_{s,c}}(1-q)^{N_s - N_{s,c}} q^{N_{\neg s,c}}(1-q)^{N_{\neg s} - N_{\neg s,c}}\, dq$$

We can use a similar methodology as we did in M_1 for M_3:

$$\mathbb{P}(N_{s,c},\, N_{\neg s,c} \mid N_s,\, N_{\neg s},\, \mathcal{M}_3) = 2\binom{N_s}{N_{s,c}}\binom{N_{\neg s}}{N_{\neg s,c}} \int_0^1 \int_0^q p^{N_{s,c}}(1-p)^{N_s - N_{s,c}} q^{N_{\neg s,c}}(1-q)^{N_{\neg s} - N_{\neg s,c}}\, dp\, dq$$

We can use R to evaluate these integrals numerically. Below is a table showing the evaluation of the integrals, as well as the Bayes factors $B_{1,2}$ and $B_{1,3}$:

i	Likelihood under model i	$B_{1,i}$
1	0.006729	1.000
2	0.000061	110.389
3	0.000005	1353.592

As per Jeffrey's scale, the data gives strong evidence that we should prefer M_1 over M_2 or M_3. In other words, the data implies there is strong evidence of a positive link between smoking and lung cancer.

4 Question 3

Part a.

Though a reasonable assumption is that $L^* \in [0, \infty]$, we cannot make this assumption about L, as under this model it could be impacted by noise which pushes it into the negative domain. Therefore, we take the support of L to be over the reals. We will start by calculating the normalising constant, then go into constraints on β:

$$\mathbb{P}(L|L^*, \Delta, \sigma, \beta) = \frac{1}{C}\left\{1 + \frac{[L - (L^* + \Delta)]^2}{\sigma^2}\right\}^{-\beta}, \qquad \int_{-\infty}^{\infty}\left\{1 + \frac{[L - (L^* + \Delta)]^2}{\sigma^2}\right\}^{-\beta} dL = C$$

Substituting $\xi = [L - (L^* + \Delta)]/\sigma$, and noticing $dL = \sigma d\xi$:

$$\sigma \int_{-\infty}^{\infty} (1 + \xi^2)^{-\beta}\, d\xi = C$$

Using the identity $\int_{-\infty}^{\infty}(1 + x^2)^{-y}dx = \dfrac{\sqrt{\pi}\,\Gamma(y - \frac{1}{2})}{\Gamma(y)}$:

$$C = \frac{\sigma\sqrt{\pi}\,\Gamma\left(\beta - \frac{1}{2}\right)}{\Gamma(\beta)}, \quad \mathbb{P}(L|L^*, \Delta, \sigma, \beta) = \frac{\Gamma(\beta)}{\sigma\sqrt{\pi}\,\Gamma\left(\beta - \frac{1}{2}\right)}\left\{1 + \frac{[L - (L^* + \Delta)]^2}{\sigma^2}\right\}^{-\beta}$$

Now looking at whether the integral will evaluate for all β, we can re-introduce ξ as before, and we will need the following to be positive and finite (we are now dropping the normalisation constant and working proportionally with respect to L):

$$\int_{-\infty}^{\infty} (1 + \xi^2)^{-\beta} d\xi$$

There should be no singularities, as when L is around 0 the term inside the integrand should be constant. So we need to consider the tails. For large L, $1 + \xi^2 \to L^2/\sigma^2$. Similarly for small L, $1 + \xi^2 \to L^2/\sigma^2$. So in the tails, we have an integral of the form:

$$\int_{-\infty}^{\infty} L^{-2\beta}\, dL$$

For large L:

$$\int_{c}^{\infty} L^{-2\beta} dL = \left[\frac{L^{-2\beta+1}}{-2\beta + 1}\right]_c^{\infty}$$

Which will only be defined for $\beta > 1/2$. Simiarly, for small L via substitution:

$$\int_{-\infty}^{-c} (-L)^{-2\beta}\, dL = \int_{c}^{\infty} u^{-2\beta}\, du = \left[\frac{u^{-2\beta+1}}{-2\beta + 1}\right]_c^{\infty}$$

We get the same condition. Therefore, for the distribution to be proper and normalisable we need $\beta > 1/2$.

For a distribution to have a defined mean it must have a finite first moment. Again, for small L the terms inside the integrand will reduce to a constant, but for large L we will have:

$$\int_{-\infty}^{\infty} L \cdot L^{-2\beta} dL = \int_{-\infty}^{\infty} L^{-2\beta+1} dL$$

So we require $\beta > 1$ for a defined mean.

Similarly, for a probability distribution to possess a defined variance, it must have a finite second moment. Again, we witness the same behaviour, for small $|L|$ there is no misbehaviour. However, for large L we have:

$$\int_{-\infty}^{\infty} L^{-2\beta+2}$$

So we require $\beta > 3/2$ for a defined variance.

A condition of a valid probability distribution is its ability to be normalised, so any valid noise model must be normalisable. A model is of little use if we cannot sample from it.

For noise models in general, we assume samples to be symmetric around zero. In fact, we can notice that if we substitute $(L^* + \Delta + c)$ $(L^* + \Delta - c)$ into the density, we get:

$$\tilde{\mathbb{P}}_L[(L^*+\Delta+c)|L^*,\Delta,\sigma,\beta] = \left\{1 + \frac{[(L^*+\Delta+c)-(L^*+\Delta)]^2}{\sigma^2}\right\}^{-\beta} = \left\{1+\frac{c^2}{\sigma^2}\right\} = \tilde{\mathbb{P}}_L[(L^*+\Delta-c)|L^*,\Delta,\sigma,\beta]$$

This shows the distribution is symmetric around $(L^* + \Delta)$. This property is unaffected by the value of β.

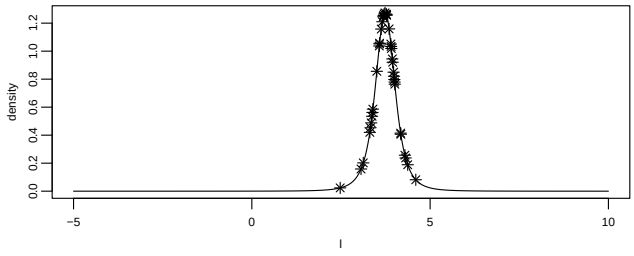
For a noise model to be valid, it is logical to impose two conditions as absolutely necessary: symmetry and normalisability. This is satisified even for $\beta \in (0.5, 1]$, and therefore we can argue that whilst likely useful to have in a statistical setting, there is nothing about having an undefined mean or undefined variance which would make this noise model invalid.

Part b.

The fact that our model has negative support may seem incorrect in principle, but it completely depends on the nature of the machine taking the observations. Perhaps there is a large enough noise to allow for negative values, or perhaps the machine has an algorithm which rejects readings below a certain "impossibly low" lung capacity.

In terms of the impact on inference, if we're trying to analyse the effect of lung capacity on cancer risk, a negative value doesn't pose any great problems, we can just treat it as a low lung capacity (as the corresponding L^* is likely low).

Looking at $L_{1:M}$, we do not observe any negative values. Estimating $\hat{\Delta} = \bar{L} - L^*$ where $\bar{L} = \frac{1}{m}\sum_{m=1}^{M} L_m$ and taking $\sigma = 0.5$, $\beta = 2$, we can plot what an example of our function could look like with known parameters. Note that given these parameters and this participant's lung capacity, we would expect a negative value roughly every 2000† samples. This will increase for smaller L^* values. However, based on our $L_{1:M}$, a negative value is seemingly unlikely.



Part c.

$$\mathbb{P}(\Delta,\sigma,\beta|L_{1:M},L^*) = \mathbb{P}(\Delta,\sigma,\beta|L_{1:M})\mathbb{P}(L_{1:M}|\Delta,\sigma,\beta,L^*)$$

$$\mathbb{P}(\Delta,\sigma,\beta|L_{1:M},L^*) = \Theta(\sigma)\Theta(\beta-2)\prod_{m=1}^{M}\frac{\Gamma(\beta)}{\sigma\sqrt{\pi}\,\Gamma\left(\beta-\frac{1}{2}\right)}\left(1+\frac{[L_m-(L^*+\Delta)]^2}{\sigma^2}\right)^{-\beta}$$

Evaluating this integral completely is challenging. An easier alternative is to try and show a region of (Δ, σ, β) where this integral diverges. Proving one region is infinite is sufficient to prove the distribution is improper, since we do not have any negative mass.

$$\int_0^\infty \int_2^\infty \int_{-\infty}^\infty \left(\frac{\Gamma(\beta)}{\sigma\sqrt{\pi}\,\Gamma\left(\beta - \frac{1}{2}\right)} \right)^M \prod_{m=1}^M \left(1 + \frac{\left[L_m - (L^* + \Delta)\right]^2}{\sigma^2} \right)^{-\beta} d\Delta\, d\beta\, d\sigma$$

Let's consider large β and large σ. For large β, $\Gamma(\beta) \approx \beta!$, so for large β, $\dfrac{\Gamma(\beta)}{\Gamma\left(\beta - \frac{1}{2}\right)} \approx \beta^{1/2}$. We can therefore re-write the normalising constant for large β as $\dfrac{\beta^{M/2}}{\sigma^M \pi^{M/2}}$. So for large β:

$$\propto \int_0^\infty \int_2^\infty \int_{-\infty}^\infty \beta^{M/2} \sigma^{-M} \prod_{m=1}^M \left(1 + \frac{\left[L_m - (L^* + \Delta)\right]^2}{\sigma^2} \right)^{-\beta} d\Delta\, d\beta\, d\sigma$$

Now if we are considering large σ, the term inside the product will be of the form $(1 + x^2)^{-\beta}$ where x is small. Since we have small x, we can use the identity $(1 + x)^y \approx \exp(xy)$ for small x. Reformulating the integral accordingly:

$$\propto \int_0^\infty \int_2^\infty \int_{-\infty}^\infty \beta^{M/2} \sigma^{-M} \prod_{m=1}^M \exp\left(\frac{-\beta \left[L_m - (L^* + \Delta)\right]^2}{\sigma^2} \right) d\Delta\, d\beta\, d\sigma$$

$$\propto \int_0^\infty \int_2^\infty \int_{-\infty}^\infty \beta^{M/2} \sigma^{-M} \exp\left(\frac{-\beta}{\sigma^2} \sum_{m=1}^M (L_m - L^* - \Delta)^2 \right) d\Delta\, d\beta\, d\sigma$$

Let $y_m = L_m - L^*$. The summation term then becomes $(y_m - \Delta)^2$. Using standard trickery like that seen in MLE and inference on Gaussians, we can re-order the integral as follows:

$$\propto \int_0^\infty \int_2^\infty \beta^{M/2} \sigma^{-M} \exp\left(-\frac{\beta \cdot (\sum y_m^2 - M\bar{y}^2)}{\sigma^2} \right) \left[\int_{-\infty}^\infty \exp\left(-\frac{\beta M}{\sigma^2} (\Delta - \bar{y})^2 \right) d\Delta \right] d\beta\, d\sigma$$

The Δ integrand is a Gaussian integral:

$$\propto \int_0^\infty \int_2^\infty \beta^{M/2} \sigma^{-M} \exp\left(-\frac{\beta \cdot (\sum_M y_m^2 - M\bar{y}^2)}{\sigma^2} \right) \sigma \sqrt{\frac{\pi}{\beta M}} d\beta\, d\sigma$$

Substituting $A = (\sum_M y_m^2 - M\bar{y}^2)/\sigma^2$:

$$\propto \int_0^\infty \sigma^{1-M} \left[\int_2^\infty \beta^{(M-1)/2} e^{-A\beta} d\beta \right] d\sigma$$

Let $\alpha = A\beta$, $d\beta = A^{-1} d\alpha$

$$\propto \int_0^\infty \sigma^{1-M} \int_{2A}^\infty (\alpha/A)^{(M-1)/2} e^{-\alpha} A^{-1} d\alpha\, d\sigma$$

$$\propto \int_0^\infty \sigma^{1-M} A^{-(M+1)/2} \int_{2A}^\infty \alpha^{(M-1)/2} e^{-\alpha} d\alpha\, d\sigma$$

The integral with respect to α is an incomplete Gamma function, and will converge to $\Gamma((M+1)/2)$, which we can factor out. Additionally, when expanding A, we are only interested in the σ terms.

$$\propto \int_0^\infty \sigma^{1-M} \cdot \sigma^{M+1} d\sigma \qquad \propto \int_0^\infty \sigma^2 d\sigma$$

This integral clearly diverges for large σ, proving the posterior is improper.

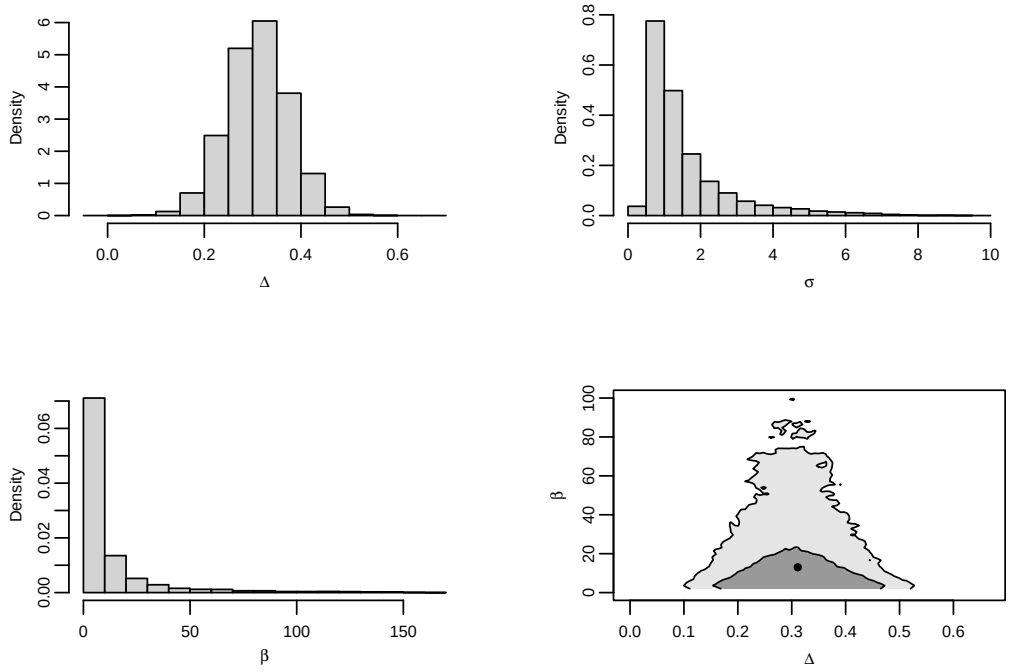
Part d.

Since we have a large product, it will be easier to work with the log-posterior:

$$\ell(\Delta, \sigma, \beta | L_{1:M}, L^*) = \log \left\{ \Theta(\sigma)\Theta(\beta - 2)\beta^{-2} \left(\frac{\Gamma(\beta)}{\sigma\sqrt{\pi}\,\Gamma\left(\beta - \frac{1}{2}\right)} \right)^M \prod_{m=1}^{M} \left(1 + \frac{[L_m - (L^* + \Delta)]^2}{\sigma^2} \right)^{-\beta} \right\}$$

$$\ell(\Delta, \sigma, \beta | L_{1:M}, L^*) = -2\log\beta + M\left(\log(\Gamma(\beta)) - \log\sigma - \frac{1}{2}\log\pi - \log\left(\Gamma(\beta - \frac{1}{2})\right)\right) - \sum_{m=1}^{M} \beta \log\left(1 + \frac{[L_m - (L^* + \Delta)]^2}{\sigma^2} \right)$$

We run the MH algorithm with this target distribution for 1,000,000 iterations, discarding the first 2,000 samples as burn-in, will generate us 998,000 samples of (Δ, β, σ). The proposal scale of the MH algorithm was tuned to match an acceptance rate in the range 0.3-0.4, found to be optimal for mixing in 3 dimensions by Rosenthal (2011). The marginal posteriors for (Δ, σ, β) are below, followed by the 80% and 95% credible interval of the joint posterior of (Δ, β):



From the contour plot, we can note there is large uncertainty on the true value of β. Looking at its marginal distribution too, we can see a peak at small values, but a very slow decay, which leaves us unsure about the true value. Whilst there is a lot of mass near $\beta < 10$ which would imply that our noise model is heavy-tailed, there is still substantial mass at $\beta > 50$, meaning we can't reject the idea of a Gaussian-esque tail.

There is however strong evidence suggesting that there is a systemic bias, under a non-biased distribution we would expect approximately 50% of the mass of the Δ posterior to be in the negative domain, but we find that there is almost zero† mass from our samples in the negative domain. So, our data displays strong evidence to suggest that the machine is overestimating lung capacity.

5 Question 4

Part a.

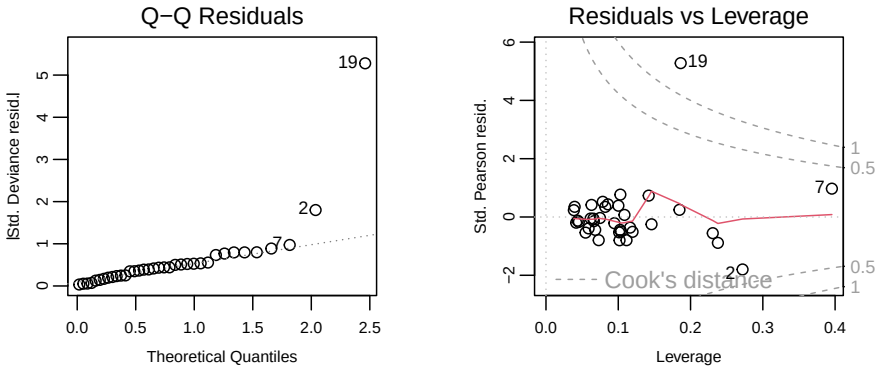
Here, we are assuming $L_n \sim \mathrm{N}(L_n^*, 0.3^2)$. Inputting this information into the `glm` function, we can observe the corresponding p_i, z_i and β_i values for each covariate in the table below:

Parameter	Estimate	z	p
β_0	-2.905	-3.304	9.54e-04
β_1	1.465	7.839	4.54e-15
β_2	0.058	0.317	7.51e-01
β_3	-0.053	-8.726	2.64e-18

Each p_i value can be interpreted as $p_i = 2(\Phi(-|z_i|)$ where Φ is the standard normal cdf and $z_i = \beta_i / SE(\beta_i)$. The standard error for each covariate is based on our inputted standard deviation rather than the observed scatter.

We have strong evidence to suggest that there is a negative relationship between smoking and lung capacity, as well as a positive link between age and lung capacity. However, there is no suggestion that there is a link between weight and lung capacity.

We can further explore the impact of outliers on the gaussian noise assummption, as well as their leverage. Below are the Pearson residuals plotted against the leverage of each point, as well as a Q-Q plot of the residuals:

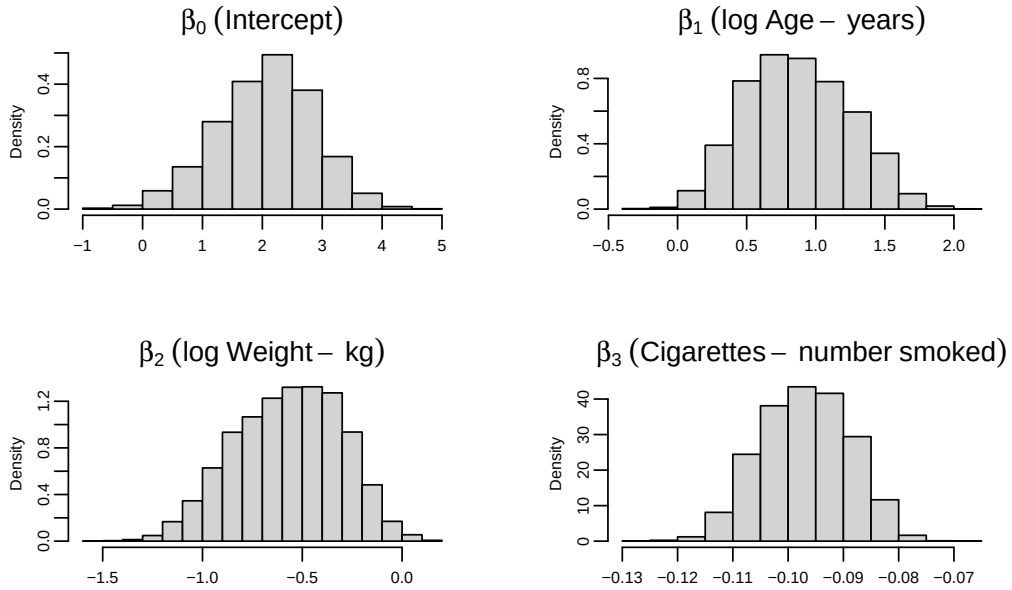


The Q-Q plot is largely what we expect to see from a well fitted model, but the at the tail the extreme residuals show huge deviation from the normality assumption. We can see from the leverage plot that the covariates of participant $n = 7$ is having the largest leverage on the model for any participant. As W_7 is an extremely low level, we will explore what impact this has in the following part.

Part b.

We take the logarithm of the noise model described in Question 3. As we're sampling with fixed β, σ, we do not need to include the proportionality constant as we're sampling with respect to L. We run 4 chains of original length 5000, discarding the first 1000 samples of each, giving us 16,000 total samples. Below are the marginal posteriors for each β_i, as well as the summary of the model output.

Parameter	Estimate	SD	2.5%	97.5%	$\hat{R}$
β_0	2.052	0.809	0.365	3.549	1.000
β_1	0.872	0.375	0.200	1.593	1.001
β_2	-0.571	0.265	-1.096	-0.106	1.001
β_3	-0.097	0.008	-0.112	-0.082	1.001



Instantly, we have a very different output to the normal noise `glm` model. Firstly, we can notice that for all of the covariates we have no crossing of 0 by the 95% credible interval, indicating evidence that all of the covariates have a link with lung capacity, which marks a change from the normal model finding no relevance for weight. We still observe a positive link between log(age) and lung capacity as well as a negative link between smoking and lung capacity as in the normal model.

We have included the Gelman-Rubin statistic $\hat{R}$ to show adequate convergence has taken place.

From this model, we have evidence to suggest that lung capacity has a positive link with age, and a negative link with smoking and weight.

As for the reliability of the two models, the main difference comes in how they treat extreme noise. If we analyse the loss functions with respect to the negative log-likelihood, the normal model will be of the form:

$$\mathrm{L}_1 \propto \frac{(L - L^*)^2}{2\sigma^2}$$

Whilst the loss function with respect to the negative log-likelihood of the Q3 noise model will be of the form:

$$\mathrm{L}_2 \propto \beta \log\left(1 + \frac{(L - L^*)^2}{\sigma^2}\right)$$

Both `rstan` and `glm` work by seeking to reduce these loss functions, and since L_1 grows quadratically, it punishes outliers far more than the logarithmic growth of L_2.

It is very likely that it is as a direct result of the extremely low value of W_7 that the normal model fails to observe the negative link between weight and lung capacity, but the heavy tailed model does. As we have changed our loss function from quadratic to logarithmic, the leverage of these individually extreme values will be greatly reduced

Part c.

We first have to consider if we are to view lung capacity and whether an individual has lung cancer as independent. If we do decide to assume independence, we can form a joint model where: $\mathbb{P}(L, C|\Delta, \sigma, \beta) = \mathbb{P}(L|\Delta, \sigma, \beta)\mathbb{P}(C|\Delta, \sigma, \beta)$

As having or not having cancer is a question of binary classifcation, we can use a logistic classification model. Taking the same fixed $\beta = 3, \sigma = 0.3$ as before and combining these two models, we would input the following log-posterior to `rstan` sample from:

$$\log \mathbb{P}(\beta, \gamma \mid L_{1:N}, C_{1:N}) \propto \sum_{i=1}^{N} \ell_{L_n} + \ell_{C_n}$$

Where:

$$\ell_{L_n} \propto -3 \log\left(1 + \frac{(L_n - L_n^*)^2}{0.3^2}\right), \quad \ell_{C_n} = C_n \log(p_n) + (1 - C_n)\log(1 - p_n)$$

$$L_n^* = \beta_0 + \beta_1 \log(A_n) + \beta_2 \log(W_n) + \beta_3 \eta_n$$

$$\log\left(\frac{p_i}{1 - p_i}\right) = \gamma_0 + \gamma_1 \log(A_i) + \gamma_2 \log(W_i) + \gamma_3 \eta_i$$

Here p_i is the predicted underlying probability that individual i has lung cancer.

By default `rstan` will default to the following improper prior as in the previous part, which works for us:

$$\gamma, \beta \sim \text{Uniform}(-\infty, \infty)$$

However, this is likely a naïve approach. It is very possible that lung capacity and lung cancer are linked, and this model will completely miss this dependence, leading to bias. A correlation between the two makes a binary classification model much more complex, and it would be easier to model a propensity term ω_i such that:

$$C_n = \begin{cases} 1 & \text{if } \omega_n > 0 \\ 0 & \text{if } \omega_n \leq 0 \end{cases}$$

ω_n will be sampled from a negative support for $C_n = 0$, and positive for $C_n = 1$. The higher the absolute value the more the sampler's predictors also agree with the actual outcome, and a lower absolute is likely an effect of a sampler trying to model the other binary outcome. We can ensure this happens with the following addition to the log-posterior: $\log\{\Theta[(C_n - 0.5)\,\omega_n]\}$ where Θ is the Heaviside step function.

We then need to define a joint distribution. We could simplify to student-t distribution, as for low β this will be very similar to our current noise model. However, we will suggest generalising the existing model as follows:

$$\frac{(L - L^*)^2}{\sigma^2} = (L - L^*) \cdot \frac{1}{\sigma^2} \cdot (L - L^*) \to (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

Additionally define:

$$\mathbf{x}_n = (L_n, \omega_n)^T, \quad \boldsymbol{\mu}_n = \begin{pmatrix} \beta_0 + \beta_1 \log(A_n) + \beta_2 \log(W_n) + \beta_3 \eta_n \\ \gamma_0 + \gamma_1 \log(A_n) + \gamma_2 \log(W_n) + \gamma_3 \eta_n \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} 0.3^2 & 0.3\rho \\ 0.3\rho & 1 \end{pmatrix}$$

We have fixed the variance for ω_n, as it is largely inconsequential in the binary classification setup, a negative value will predict a 0 no matter the scale of the negativity. Adding it as an unknown parameter would likely cause `rstan` serious convergence issues.

Assuming our uniform priors, this gives us an overall log-posterior of:

$$\log \mathbb{P}(\beta, \gamma, \rho, \boldsymbol{\omega} \mid L_{1:N}, C_{1:N}) \propto \sum_{n=1}^{N} \left[-3 \log\left(1 + (\mathbf{x}_n - \boldsymbol{\mu}_n)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_n)\right) + \log\{\Theta[(C_n - 0.5)\,\omega_n]\}\right]$$

This would be ready to input to `rstan` with a similar framework as that of Q4(b).

6 Additional notes and Supplementary Material

Additional notes

For Q1 for the 1st and 3rd point I assume some poor sampling technique, especially for such a small survey. I took the view that this is valid as we are given no information on how the sampling is performed, so we cannot say for definite that the survey was not haphazard in its sampling methodology.

For Q2 (b), in the second expression of the first likelihood it seemed intuitive to put just S_i and $\neg S_i$ instead of $S_i = S$ and $S_i = \neg S$, but this choice of notation would have implied that the likelihood was conditioning on all N participants being smokers. I don't believe this was the question's intention.

For Q3,Q4 there is a mixing of notation, with β sometimes meaning the shape parameter for the noise model and sometimes meaning the covariates for predicting lung capacity. This is a bit sloppy, but I hope it's always clear which I mean.

For Q4(b) the iterations on the stan model are fairly low. I don't know if it's because stan is poorly setup on my computer or I badly wrote the model but it took me a lot of time to run, so I had to keep iterations low. Since the $\hat{R}$ was very low I saw this as being sufficient, and when I tried higher number of iterations the standard errors changed little.

Figures were not labelled, as in the model solutions this was also the case.

Supplementary Material

Q2(c):

```
q_samples <- rbeta(100000, N_ns_c + 1, N_ns - N_ns_c + 1)
p_samples <- rbeta(100000,N_s_c + 1, N_s - N_s_c + 1)
mean(p_samples > q_samples)
#> [1] 0.99909
```

Q3(b):

```
1/integrate(f, -Inf, 0)$value
#> [1] 2013.47
```

Q3(d):

```
mean(samples[,1]>0)
#> [1] 0.999999
```

7 Code Appendix

```
knitr::opts_chunk$set(
  collapse = TRUE,
  comment = "#>"
)
include_solutions <- TRUE
require(rmarkdown)
require(knitr)
require(kableExtra)
require(mcmc)
require(MASS)
require(rstan)
# initalise
set.seed(06049216) #CID
```

```r
smokingdata <- read.csv("smoking.csv")
calibrationdata <- read.csv("calibration.csv")

N <- nrow(smokingdata)

# Smokers partitioned
N_s <- sum(smokingdata$cigarettes > 0)
N_s_c <- sum(smokingdata$cigarettes > 0 & smokingdata$cancer == T )

# Non-smokers partitioned
N_ns <- N - N_s
N_ns_c <- sum(smokingdata$cigarettes == 0 & smokingdata$cancer == T)

par(mar=c(4, 3.5, 2, 2), cex.lab=0.7, cex.axis=0.7, mgp=c(1.5, 0.5, 0))
# 2c plots
hpd <- function(a, b) {
  x <- sort(rbeta(1e5, a, b))
  gap <- round(0.95 * length(x))
  i <- which.min(x[(gap + 1):length(x)] - x[1:(length(x) - gap)])
  c(x[i], x[i + gap])
}

h_q <- hpd(N_ns_c + 1, N_ns - N_ns_c + 1)
h_p <- hpd(N_s_c + 1, N_s - N_s_c + 1)

plot(NULL, xlim = c(0, 1), ylim = c(0, 8), xlab="p,q",ylab="Density")

x_q <- seq(h_q[1], h_q[2], length = 100)
polygon(c(h_q[1], x_q, h_q[2]), c(0, dbeta(x_q, N_ns_c + 1, N_ns - N_ns_c + 1), 0),
col = rgb(1, 0, 0, 0.3), border = NA)

x_p <- seq(h_p[1], h_p[2], length = 100)
polygon(c(h_p[1], x_p, h_p[2]), c(0, dbeta(x_p, N_s_c + 1, N_s - N_s_c + 1), 0),
col = rgb(0, 0, 1, 0.3), border = NA)

curve(dbeta(x, N_ns_c + 1, N_ns - N_ns_c + 1), col = "red", add = TRUE)
curve(dbeta(x, N_s_c + 1, N_s - N_s_c + 1 ), col = "blue", add = TRUE)

#2d integrals and BF
f <- function(p, q) p^N_s_c * (1 - p)^(N_s - N_s_c) * q^N_ns_c * (1 - q)^(N_ns - N_ns_c)
integ_m1 <- integrate(function(q) {sapply(q, function(qn) {
        integrate(function(p) f(p, qn), lower = qn, upper = 1)$value
      })},lower = 0, upper = 1)$value
m1 <- 2 * choose(N_s, N_s_c) * choose(N_ns, N_ns_c) * integ_m1
f_2 <- function(q) q^(N_s_c + N_ns_c) * (1 - q)^(N - N_s_c - N_ns_c)
integ_m2 <- integrate(function(q) f_2(q), lower = 0, upper = 1)$value
m2 <- choose(N_s, N_s_c) * choose(N_ns, N_ns_c) * integ_m2
integ_m3 <- integrate(function(q) {sapply(q, function(qn) {
        integrate(function(p) f(p, qn), lower = 0, upper = qn)$value
      })},lower = 0, upper = 1)$value
m3 <- 2 * choose(N_s, N_s_c) * choose(N_ns, N_ns_c) * integ_m3
kab <- data.frame(
  I = c(1,2,3),
```

```
  M = c(m1, m2, m3),
  B = c(1, m1/m2,m1/m3)
)


kabl <- kable(kab, digits = 6,
              col.names = c("$i$","Likelihood under model $i$", "$B_{1,i}$"),
              format = "latex", escape = F)
kable_styling(kabl, latex_options = "HOLD_position")
#3b star plot
par(mar=c(4, 3.5, 2, 2), cex.lab=0.7, cex.axis=0.7, mgp=c(1.5, 0.5, 0))
lstar <- 3.44
sigma <- 0.5
beta <- 2
L <- calibrationdata$X3.419
delta <- mean(calibrationdata$X3.419) - lstar

f <- function(l) {
 gamma(beta) / (sigma * sqrt(pi) * gamma(beta - 0.5)) *
    (1 + ((l - (lstar + delta))^2) / sigma^2)^(-beta)
}

l <- seq(-5, 10, length.out = 1000)
y <- f(l)

plot(l, y, type = "l", ylab="density")
points(L, f(L), pch = 8)
#3d mcmc
log_posterior <- function(theta, dataset) {
  delta <- theta[1]
  sigma <- theta[2]
  beta  <- theta[3]

  if (sigma <= 0 || beta <= 2) return(-Inf)

  log_post <- -2 * log(beta)

  M <- length(dataset)

  log_post <- log_post + M * (log(gamma(beta)) - log(sigma) -
  0.5 * log(pi) - log(gamma(beta - 0.5))) -
  beta * sum(  log(1 + ((dataset - (lstar + delta))^2) / sigma^2))

  return(log_post)
}

mcmc_samples <- metrop(
  obj =  log_posterior,
  initial = c(delta = 0.3, sigma = 1, beta = 3),
  nbatch = 1000000,
  scale = c(0.1, 0.1, 0.7), # tuned to get decent acceptance rate
  dataset = L,
)
```

```r
samples <- mcmc_samples$batch[-(1:2000),] # remove first 2000 observations for each param
colnames(samples) <- c("delta", "sigma", "beta")


#3d contour plot
par(mar=c(4, 3.5, 2, 2), cex.lab=0.7, cex.axis=0.7, mgp=c(1.5, 0.5, 0), mfrow = c(2, 2))
frac1 <- 0.80
frac2 <- 0.95

dens <- kde2d(samples[, "delta"], samples[, "beta"], n = 100) #2d density from MASS

# Module provided code slightly modified
densities <- dens$z
x_grid <- dens$x
y_grid <- dens$y

densities_sorted <- sort(densities, decreasing = TRUE)
mass_tot <- sum(densities_sorted)
mass_cumulative <- 0
found1 <- FALSE
found2 <- FALSE
level1 <- 0
level2 <- 0

for (i in seq_len(length(densities_sorted))) {
  mass_cumulative <- mass_cumulative + densities_sorted[i]

  if (mass_cumulative >= frac1 * mass_tot && found1 == FALSE) {
    level1 <- densities_sorted[i]
    found1 <- TRUE
  }
  if (mass_cumulative >= frac2 * mass_tot && found2 == FALSE) {
    level2 <- densities_sorted[i]
    found2 <- TRUE
  }
}

hist(samples[, "delta"], xlab = expression(Delta), prob = TRUE, main="")
hist(samples[, "sigma"], xlab = expression(sigma), prob = TRUE, main="")
hist(samples[, "beta"], xlab = expression(beta),  prob = TRUE,main="")

plot(samples[, "delta"], samples[, "beta"], type = "n",
     xlab = expression(Delta), ylab = expression(beta), ylim = c(0,100))

level_max <- max(densities)

.filled.contour(x_grid, y_grid, z = densities,
                levels = c(level2, level1, level_max), col = c("grey90", "grey60"))

contour(x_grid, y_grid, z = densities, levels = c(level1, level2),
        add = TRUE, drawlabels = FALSE, lwd = 1)

points(mean(samples[, "delta"]), mean(samples[, "beta"]), pch = 16, cex = 0.8)
```

```r
#4a glm
fit <- glm(capacity ~ log(age) + log(weight) + cigarettes, data = smokingdata,
           family = gaussian())
s <- summary(fit, dispersion = 0.3^2)

param_names <- c("$\\beta_0$", "$\\beta_1$", "$\\beta_2$", "$\\beta_3$")

tr <- data.frame(
  Estimate = s$coefficients[, 1],
  Z_value  = s$coefficients[, 3],
  P_value  = formatC(s$coefficients[, 4], format = "e", digits = 2)
)

trkabl <- kable(
  cbind(Parameter = param_names, tr),
  digits = 3,
  col.names = c("Parameter", "Estimate", "$z$", "$p$"),
  format = "latex",
  escape = FALSE,
  row.names = FALSE
)

kable_styling(trkabl, latex_options = "HOLD_position")
par(mar=c(4, 3.5, 2, 2), cex.lab=0.7, cex.axis=0.7, mgp=c(1.5, 0.5, 0),mfrow = c(1, 2))
plot(fit, which = c(5, 2))
#4b stan model
rstan_options(auto_write = TRUE)
dataset_stan <- list(
  N = N,
  L = smokingdata$capacity,
  log_age = log(smokingdata$age),
  log_weight = log(smokingdata$weight),
  cigs = smokingdata$cigarettes,
  sigma_fixed = 0.3,
  beta_fixed = 3.0
)

stan_program <- "
data {
  int<lower=0> N;
  vector[N] L;
  vector[N] log_age;
  vector[N] log_weight;
  vector[N] cigs;
  real sigma_fixed;
  real beta_fixed;
}

parameters {
  real b0;
  real b1;
  real b2;
  real b3;
```

```
}

model {
  vector[N] mu;

  mu = b0 + b1 * log_age + b2 * log_weight + b3 * cigs;

  for (n in 1:N) {
    target += -beta_fixed * log(1 + square((L[n] - mu[n]) / sigma_fixed));
  }
}
"

stanfit <- stan(
  model_code = stan_program,
  data = dataset_stan,
  chains = 4,
  iter = 5000,
  refresh = 0
)



stansamples <- extract(stanfit)
#4b stan kable
st <- summary(stanfit, pars = c("b0", "b1", "b2", "b3"))$summary

kab_st <- data.frame(
  Estimate = st[, "mean"],
  SD       = st[, "sd"],
  Q2.5     = st[, "2.5%"],
  Q97.5    = st[, "97.5%"],
  Rhat     = st[, "Rhat"]
)

kablstan <- kable(
  cbind(Parameter = param_names, kab_st),
  digits = 3,
  col.names = c("Parameter", "Estimate", "SD", "2.5\\%", "97.5\\%", "$\\hat{R}$"),
  format = "latex",
  escape = FALSE,
  row.names = FALSE
)

kable_styling(kablstan, latex_options = "HOLD_position")
#4b stan plots
par(mar=c(4, 3.5, 2, 2), cex.lab=0.7, cex.axis=0.7, mgp=c(1.5, 0.5, 0),mfrow = c(2, 2))

hist(stansamples$b0, main = expression(beta[0] ~ (Intercept)), freq = F, xlab = "")
hist(stansamples$b1, main = expression(beta[1] ~ (log~Age--years)), freq = F, xlab = "")
hist(stansamples$b2, main = expression(beta[2] ~ (log~Weight--kg)), freq = F, xlab = "")
hist(stansamples$b3,
     main = expression(beta[3] ~ (Cigarettes--number~smoked)), freq = F, xlab = "")
```

```r
q_samples <- rbeta(100000, N_ns_c + 1, N_ns - N_ns_c + 1)
p_samples <- rbeta(100000,N_s_c + 1, N_s - N_s_c + 1)
mean(p_samples > q_samples)
1/integrate(f, -Inf, 0)$value
mean(samples[,1]>0)
```

References

Christakis, Nicholas A., and James H. Fowler. 2007. "The Spread of Obesity in a Large Social Network over 32 Years." *New England Journal of Medicine* 357 (4): 370–79. https://doi.org/10.1056/NEJMsa066082.

———. 2008. "The Collective Dynamics of Smoking in a Large Social Network." *New England Journal of Medicine* 358 (21): 2249–58. https://doi.org/10.1056/NEJMsa0706154.

Rosenthal, Jeffrey S. 2011. "Optimal Proposal Distributions and Adaptive MCMC." In *Handbook of Markov Chain Monte Carlo*, edited by S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng. Chapman & Hall/CRC.