Fundamentals of Data Science Assessment Sheet

Matt Prill

Note: *Version control has been used to track my workflow and all these documents are accessible through my github repository linked here (https://github.com/prillex/fundamental_assessed_exercises). It also contains images of my hand written workings out. My workflow for the R questions can be found in the 'rough_workflow.R' file. There is no onus to look at these; all relevant and neat workings can be seen in this document. I have included them in the repository to (A) track my own workflow (version control) and (B) allow you to consult them should you wish to.*

Question 1a

Initial Equations:

$$\begin{aligned} x + y + z &= 1 \\ x + 2y + 4z &= \eta \\ x + 4y + 10z &= \eta^2 \end{aligned}$$

Corresponding Coefficient Matrix:

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 4 & 10 \end{bmatrix} \begin{bmatrix} 1 \\ \eta \\ \eta^2 \end{bmatrix}$$

Calculating the Determinant of A:

To showcase that these equations lack a unique solution for any value of η, I can attempt to find the determinant. To to find a determinant of square matrix (irrespective of size), I must;

1. Pick a coefficient from the first row.
2. Delete remaining elements in coefficient's respective row and column.
3. Make a matrix of the remaining elements.
4. Find the determinant of the sub-matrix, and multiply this with the coefficient.
5. Repeat the same procedure for each element in the first row.
6. To determine the sign of each term, sum the indices of the coefficient. If it is even, the sign is positive, and if it's odd, the sign is negative.
7. Sum all terms to find the determinant.

Therefore, to find the determinant $(\det)$ of a 3×3 matrix, I can use the following formula:

$$\det \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} = a \times \det \begin{bmatrix} e & f \\ h & i \end{bmatrix} - b \times \det \begin{bmatrix} d & f \\ g & i \end{bmatrix} + c \times \det \begin{bmatrix} d & e \\ g & h \end{bmatrix}$$

The terms are added/ subtracted depending on the sum the row and column indices of the coefficient. Even = positive. Odd = negative. For example "- b" is negative because b is x_{12}, thus the sum of the row and column indices $= 1 + 2 = 3 = odd$.

Applying this to the 3×3 matrix:

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 4 & 10 \end{bmatrix} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$$

Thus,

$$\det(A) = 1 \times \det \begin{bmatrix} 2 & 4 \\ 4 & 10 \end{bmatrix} - 1 \times \det \begin{bmatrix} 1 & 4 \\ 1 & 10 \end{bmatrix} + 1 \times \det \begin{bmatrix} 1 & 2 \\ 1 & 4 \end{bmatrix}$$

To calculate the determinants of the 2×2 matrices, I can apply the formula:

$$\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc$$

Corresponding calculations for determinants of the 2×2 matrices:

$$\det \begin{bmatrix} 2 & 4 \\ 4 & 10 \end{bmatrix} = 20 - 16 = \mathbf{4}$$

$$\det \begin{bmatrix} 1 & 4 \\ 1 & 10 \end{bmatrix} = 10 - 4 = \mathbf{6}$$

$$\det \begin{bmatrix} 1 & 2 \\ 1 & 4 \end{bmatrix} = 4 - 2 = \mathbf{2}$$

Substituting these determinants into the aforementioned 3×3 matrix determinant formula gives:

$$\det(A) = (1 \times 4) - (1 \times 6) + (1 \times 2) = \mathbf{0}$$

The determinant of the coefficient matrix and therefore the matrix vector is 0. This means that the matrix is singular:

$$\det(A) = Ax = 0$$

In being singular, the matrix does not have an inverse meaning that the corresponding linear equations have either no, or unlimited solutions. Regardless, in this case, the initial equations have no unique solution for η.

Question 1b

The Set of Linear Equations:

$$\begin{aligned} x + y + z &= 1 \\ x + 2y + 4z &= \eta \\ x + 4y + 10z &= \eta^2 \end{aligned}$$

The Corresponding Augmented Matrix:

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & \eta \\ 1 & 4 & 10 & \eta^2 \end{array}\right]$$

Using Gaussian Elimination to Identify the Solutions for η:

1. $$(\text{Row } 2) - 1 \times (1) = \left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 1 & 3 & \eta - 1 \\ 1 & 4 & 10 & \eta^2 \end{array}\right]$$
2. $$(\text{Row } 3) - 1 \times (1) = \left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 1 & 3 & \eta - 1 \\ 0 & 3 & 9 & \eta^2 - 1 \end{array}\right]$$
3. $$(\text{Row } 3) - 3 \times (2) = \left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 1 & 3 & \eta - 1 \\ 0 & 0 & 0 & \eta^2 - 3\eta + 2 \end{array}\right]$$

The equation corresponding to the bottom row of the augmented matrix is 0. It is also a quadratic equation where:

$$\eta^2 - 3\eta + 2 = ax^2 + bx + c = 0$$

Now, I can either factorise:

$$\eta^2 - 3\eta + 2 = (\eta - 1)(\eta - 2) = 0$$

Or alternatively, I can use the quadratic formula to derive the solutions for η:

$$\eta = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} = \frac{-(-3) \pm \sqrt{-3^2 - 4 \times 1 \times 2}}{2 \times 1} = 1 \text{ or } 2$$

$$\eta = 1 \text{ or } 2$$

Characterising the equations for both solutions:

Where $\eta = 1$:

$$\begin{aligned} x + y + z &= 1 \\ x + 2y + 4z &= 1 \\ x + 4y + 10z &= 1^2 \end{aligned}$$

Corresponding augmented matrix:

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 1 & 3 & 1-1 \\ 0 & 0 & 0 & 1^2 - 3(1) + 2 \end{array}\right] = \left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 1 & 3 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

Thus, when $\eta = 1$:

$$\begin{aligned} x + y + z &= 1 \\ y + 3z &= 0 \end{aligned}$$

Where $\eta = 2$:

$$\begin{aligned} x + y + z &= 1 \\ x + 2y + 4z &= 2 \\ x + 4y + 10z &= 2^2 \end{aligned}$$

Corresponding augmented matrix:

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 1 & 3 & 2-1 \\ 0 & 0 & 0 & 2^2 - 3(2) + 2 \end{array}\right] = \left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 1 & 3 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

Thus when $\eta = 2$:

$$\begin{aligned} x + y + z &= 1 \\ y + 3z &= 1 \end{aligned}$$

Question 2a

$P_1 = P(\text{Alice winning when first to throw})$

The distribution shares some characteristics of the Bernoulli distribution whereby there are two outcomes (success & failure) and discrete. However, the question is asking for the probability to a given success. The balls are thrown (trials) until a success (hit). Furthermore, I have to assume that each throw is independent of the other which makes the distribution memoryless with two outcomes: success (win), & failure (miss). This is a geometric distribution.

$$\left.\begin{aligned} \text{Hit} &= \alpha \\ \text{Miss} &= 1 - \alpha \end{aligned}\right\} \text{Alice}$$

$$\left.\begin{aligned} \text{Hit} &= \beta \\ \text{Miss} &= 1 - \beta \end{aligned}\right\} \text{Ben}$$

The probability of Alice missing Ben with her first throw and going on to win:

P(Alice misses first throw) = $1 - \alpha$

P(Ben misses first throw) = $1 - \beta$

Thus,
P(Alice miss, Ben miss) = $(1 - \alpha) \times (1 - \beta)$

So,
P(Alice miss, Ben miss, Alice goes onto win) = $P_1 \times (1 - \alpha) \times (1 - \beta)$

P_1 is able to be used here because of the memoryless nature of the geometric distribution; no matter how many trials have transpired prior to a given throw of Alice's, the probability that she will win from that point onwards is still P_1.

Question 2b

Because I'm working with the Geometric distribution with parameter p, X has a probability mass function:

$$P(X = k) = p(k) = \begin{cases} (1-p)^{k-1}p, \\ 0 \end{cases}$$

This is not conditional probability as the probability is with respect to the beginning of the round.

As discussed, the probability that Alice and Ben both miss in a given round is: $(1-\alpha) \times (1-\beta)$

This could repeat n times where n can denotes any integer (theoretically): $((1-\alpha) \times (1-\beta))^n$ meaning the probability will be a summation of all the rounds where n is potentially indefinite $(\sum_{n=1}^{\infty})$

Eventual success from Alice (given that she throws first) therefore gives:

$$P(X = n) = \sum_{n=1}^{\infty}((1-\alpha) \times (1-\beta))^{n-1} \times \alpha$$

The $n-1$ term is used because Alice and Ben must miss every time until the final round (hence -1); the n-th round is the one she wins (success).

The α term is therefore added as it still denotes the probability that she will hit successfully.

Therefore, substituting my terms in gives:

$$P_1 = \sum_{n=1}^{\infty}((1-\alpha) \times (1-\beta))^{n-1} \times \alpha$$

In line with the geometric series, this is then equivalent to:

$$P_1 = \alpha \times \sum_{n=0}^{\infty}((1-\alpha) \times (1-\beta))^{n}$$

This holds even if n = 0 because in this instance the equation would still equate to α.

In the infinite geometric series equation $\sum_{n=0}^{\infty} r^n = \frac{1}{1-r}$ where $|r| < 1$
$(1-\alpha) \times (1-\beta)$ is the common ratio (this holds because the terms are probabilities and so < 1) Thus:

$$r = (1-\alpha) \times (1-\beta)$$

Finally, substituting the values into the geometric series equation gives:

$$P_1 = \alpha \times \frac{1}{1-(1-\alpha) \times (1-\beta)}$$

Question 2c

For Ben to throw go first but Alice still to win, there must be $1-\beta$ term.

However given the order of throws, the point at which Alice wins must be preceded by a miss from Ben $(1-\beta)$ after which she can throw (and win). Therefore,

$$P_2 \neq \sum_{n=1}^{\infty}((1-\beta) \times (1-\alpha))^{n-1} \times \alpha$$

Instead there must be a final $1-\beta$ before Alice is victorious. Again, the n-th round is the one where she wins (also the round where Ben misses a final time) Thus,

$$P_2 = \sum_{n=1}^{\infty}((1-\beta) \times (1-\alpha))^{n-1} \times (1-\beta) \times \alpha$$

By applying the geometric series steps as in the previous question, I get that:

$$P_2 = (1-\beta) \times \alpha \times \sum_{n=0}^{\infty}((1-\beta) \times (1-\alpha))^n$$

And finally:

$$P_2 = (1-\beta) \times \alpha \times \frac{1}{1-(1-\alpha) \times (1-\beta)}$$

Question 3a

Infinite geometric series for $|r| < 1$:

$$\sum_{n=0}^{\infty} r^n = \frac{1}{1-r}$$

The equation $\sum_{n=0}^{\infty} n(n-1)r^{n-2}$ has no common ratio i.e. there are extra terms: $n(n-1)$ - a polynomial.This means it is not part of the geometric series. However, I can but I can use properties of the geometric series to work out the sum of the latter. I can do so by altering the geometric series to take the same form of the equation of interest.

Differentiating the geometric series gives a similar form to the second expression (with respect to r):

$$\frac{d}{dr}\left(\sum_{n=0}^{\infty} r^n\right) = \frac{d}{dr}\left(\frac{1}{1-r}\right)$$

For the first term (equation) I apply the power rule: $\frac{d}{dx}(x^n) = nx^{n-1}$ Note, I am derivating with respect to r to match geometric series form.

$$\frac{d}{dr}\left(\sum_{n=0}^{\infty} r^n\right) = \sum_{n=0}^{\infty} nr^{n-1}$$

And for the sum I can apply the quotient rule to find the derivative: $\frac{d}{dr}\left(\frac{f}{g}\right) = \frac{g\frac{d}{dr}(f) - f\frac{d}{dr}(g)}{g^2}$ Again, I am derivating with respect to r (r = x)

$$\frac{d}{dr}\left(\frac{1}{1-r}\right) = \frac{(1-r) \times \frac{d}{dr}(1) - 1 \times \frac{d}{dr}(1-r)}{(1-r)^2} = \frac{1}{(1-r)^2}$$

Now I have:

$$\sum_{n=0}^{\infty} nr^{n-1} = \frac{1}{(1-r)^2}$$

The equation form now mirrors that of the equation of interest: $\sum_{n=0}^{\infty} n(n-1)r^{n-2}$

Next, I can take the second derivative:

The second derivative of $\sum_{n=0}^{\infty} r^n$:

$$\frac{d^2}{dr^2}\left(\sum_{n=0}^{\infty} r^n\right) = \frac{d}{dr}\sum_{n=0}^{\infty} nr^{n-1}$$

Using the power rule again:

$$\frac{d}{dr}(\sum_{n=0}^{\infty} nr^{n-1}) = \sum_{n=0}^{\infty} n(n-1)r^{n-2}$$

This is the exact same as the equation of interest. Therefore, the second derivative of the given geometric series $\left(\sum_{n=0}^{\infty} r^n\right)$ is equal to the equation of interest:

$$\frac{d2}{dr2}\left(\sum_{n=0}^{\infty} r^n\right) = \frac{d}{dr}\sum_{n=0}^{\infty} nr^{n-1} = \sum_{n=0}^{\infty} n(n-1)r^{n-2}$$

Therefore, to find its respective sum, I must find the second derivative of the initial equations sum too. I can do this again with the quotient rule.

$$\frac{d2}{dr2}\left(\frac{1}{1-r}\right) = \frac{d}{dr}\left(\frac{1}{(1-r^2)}\right) = \frac{(1-r)^2 \times \frac{d}{dr}(1) - 1 \times \frac{d}{dr}(1-r)^2}{((1-r)^2)^2}$$

Note: I now also have to apply chain rule to differentiate $(1-r)^2$. $1-r$ is the inner function, ^{2} is the outer. So $\frac{d}{dr}(1-r)^2 = -2(1-r)$ Thus, I arrive at:

$$= \frac{(1-r)^2 \times 0 - 1 \times (-2 \times (1-r))}{(1-r)^4} = \frac{2(1-r)}{(1-r)^4} = \frac{2}{(1-r)^3}$$

To conclude:

$$\sum_{n=0}^{\infty} n(n-1)r^{n-2} = \frac{2}{(1-r)^3}$$

Question 3b

A:

I need to calculate the expectation of a fine (true when weight < 420.75) per box. This is equivalent to the CDF $F(x) = P(X \leq 420.75)$. Then I can calculate the expected value of the fine/box.

The tins weight follows a normal distribution $X \sim N(\mu = 426, \sigma = 21)$. The PDF of the normal distribution is:

$$f(x \mid \mu, \sigma^2) = \frac{1}{2\pi\sigma^2}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

Substituting the parameters gives:

$$f(x \mid 426, 21^2) = \frac{1}{2\pi(21)^2}\exp\left\{-\frac{1}{2 \times (21)^2}(x-426)^2\right\}$$

The corresponding CDF cannot be denoted analytically due to but must be evaluated. The CDF $F(x) = P(X \leq 420.75)$ is the integral of the PDF. However, the CDF of a normal distribution (denoted as Φ) which means I can consult a 'Z-' probability table. To calculate the CDF corresponding z-value I can use the formula:

$$Z = \frac{X - \mu}{\sigma_{\text{sample}}}$$

Importantly, the variance value in this term is actually standard deviation of a box because it is tied to the sample size. In this case, the sample size is a whole box (100 cans), not a single tin:

$$\sigma_{\text{sample}} = \frac{\sigma}{\sqrt{n}} = \frac{21}{\sqrt{100}} = 2.1$$

This makes sense because as sample size increases, variance should decrease. In turn, by weighing a whole box, the variance of the box will decrease significantly, even if the variance of the individual tins is constant.

Substituting the parameters into the penultimate equation gives:

$$Z = \frac{420.75 - 426}{2.1} = -2.50$$

Now, I can consult the z-table:

Standard Normal Cumulative Probability Table

Cumulative probabilities for NEGATIVE z-values are shown in the following table:

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

The CDF's corresponding z-value is 0.0062. This means the probability of a box's weight/can being <420.75g is 0.62%. To work out the expectation of the fine:

$$E(\text{Fine}) = 3000 \times 0.0062 = 18.6$$

Thus for A:

$$E(\text{Fine}) = £18.60/\text{box}$$

B:

I need to calculate the CDF $F(x) = P(X \leq 421.8)$. I will use the same logic as before and calculate the z-value before consulting the table. This time, $X = 421.8$ and $n = 25$ which means that to calculate the z-value:

$$\sigma_{\text{sample}} = \frac{\sigma}{\sqrt{n}} = \frac{21}{\sqrt{25}} = 4.2$$

This higher σ compared to the previous question is to be expected given the smaller sample size. Thus:

$$Z = \frac{421.8 - 426}{4.2} = -1$$

Standard Normal Cumulative Probability Table

Cumulative probabilities for NEGATIVE z-values are shown in the following table:

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

The CDF's corresponding z-value is 0.1587. This means the probability of a box's weight/can being <421.8g is 15.87%. To work out the expectation of the fine:

$$E(\text{Fine}) = 120 \times 0.1587 = 19.044$$

Thus for B:

$$E(\text{Fine}) = £19.04/\text{box}$$

C:

The probability that X > 426 is equivalent to $1 - P(X \leq 426)$. I could do this again and consult the z-table but this is unnecessary. 426 is the mean and normally distributed so there is a 0.5 chance that a randomly sampled can will weigh more and 0.5 that it will weigh less. Note: this takes into account the fact that a PDF of a given continuous variable (in this case 426.0g) = 0.

Here, each sampling event is either success $(X > 426)$ or failure $(X \leq 426)$ where n could take any number. This is characteristic of the geometric formula. As discussed, the finite geometric series is:

$$\sum_{n=0}^{\infty} r^n = \frac{1}{1-r}$$

Since the probability of success = 0.5:

$$\sum_{n=0}^{\infty} r^n = \frac{1}{1-0.5} = 2$$

As expected, it takes two trials until a success is expected. In turn:\ (turn:\)

$$E(n) = 2$$

N itself however will have its own distribution which must be considered before plugging into the fine equation. The variance of n can be calculated using the variance formula for geometric distributions:

$$Var(n) = \frac{1-p}{p^2}$$

Thus:

$$Var(n) = \frac{1-0.5}{0.5^2} = 2$$

To calculate the expectation of the fine I need to use the formula,

$$E(n^2) = \mathrm{Var}(n) + (E(n))^2$$

I have the variance and expected no. samples until success but i also need the expectation of n^2 because $n(n-1) = n^2 - n$. Therefore to calculate the expectation of n^2:

$$E(n)^2 = 4 + 2^2 = 6$$

Now I can calculate the E(n(n-1)):

$$E(n(n-1)) = E(n^2) - E(n) = 6 - 2 = 4$$

Now I have $E(n(n-1))$, I can plug it into the equation:

$$E(\mathrm{Fine}) = 5 \times E(n(n-1)) = 5 \times 4 = 20$$

In conclusion:

$$E(\mathrm{Fine}) = £20/\mathrm{box}$$

Question 4a

Introducing the random variables:

Individually, $X_i^{(1)}$ and $X_i^{(2)}$ random variables follow Bernoulli distributions because its success/failure (outcomes sum to 1), each trial is independent and n is 1.

$$X_i^{(1)} = \begin{cases} 1, & \text{if B supporter number } i \text{ still supports B at the end of day 1,} \\ 0, & \text{if B supporter number } i \text{ changes to an M supporter at the end of day 1.} \end{cases}$$

For $i = 1..., 175$ this is equivalent to:

$$X_i^{(1)} = \begin{cases} k = 1,\ p = 1 - 0.004 = 0.996 \\ k = 0,\ 1 - p = 1 - 0.996 = 0.004 \end{cases}$$

In summary, each random variables can be denoted as $X_i^{(1)} \sim \text{Bernoulli}(1, p)$ for $i = 1..., 175$.

For $X^{(2)}$:

$$X_i^{(2)} = \begin{cases} 1, & \text{if M supporter number } i \text{ changes to a B supporter at the end of day 1,} \\ 0, & \text{if M supporter number } i \text{ still supports M at the end of day 1.} \end{cases}$$

for $i = 1..., 186$, this is equivalent to:

$$X_i^{(2)} = \begin{cases} k = 1,\ p = 0.005 \\ k = 0,\ 1 - p = 1 - 0.005 = 0.996 \end{cases}$$

Again, each random variable trial can be denoted as $X_i^{(2)} \sim \text{Bernoulli}(1, p)$ for $i = 1..., 186$.

Expressing the number of B supporters at the end of the first day:

Every trial is independent and peoples allegiance has no bearing on other people. Crucially, distinct Bernoulli trials are repeated; not the same Bernoulli trial. This is because every X is unique within a given day i.e. unique random variables (prospective voters). This precludes the use of the Binomial distribution characteristics; using binomial formulas would imply that the Bernoulli trials are repeated on single person n within a single day. This is not the case.

The number of B supporters at the end of day 1 will be the B's Initial supporters - those that switch allegiance to support M + those that switch allegiance to support B:

$$\text{B supporters after 1 day} = \text{B supporters at the start of day} - \text{Supporters who switch to M} + \text{Supporters who switch to B}$$

Here I will denote 'B supporters at the start of the day' as 'B' and M supporters at the start of the day as 'M' which are both integers. Using the introduced Random distributions this gives:

$$\text{B supporters after 1 day} = B - (B - \sum_{i=1}^{B} X_i^{(1)}) + \sum_{i=1}^{M} X_i^{(2)}$$

The $(B - \sum_{i=1}^{B} X_i^{(1)})$ is included as it is all the failed trials for $X_i^{(1)}$ minus the number of B supporters who switch. The sum symbol is sufficient here because if a trial ends in success,the output is 1. This represents a person so the sum of these will equate to the number of successes. Therefore, probabilities such as $'p'$ and $'p - 1'$ do not need to be expressed in the equation above.

Expected number of B supporters at the end of the first day:

Using the formula, I can calculate the E(B supporters after one day). I must first calculate the sum of the expected occurrences in the two $X_i^{(1)}$ and $X_i^{(2)}$ terms.

I can use the properties of the expectation to calculate these. Because I'm not working with a binomial distribution I will not use its expectation formula. Instead, I will use the expectation formula for the Bernoulli: E(X) = p. Therefore p can be multiplied for the number of Bernoulli trials in each term:

expectations For $E(\sum_{i=1}^{175} X_i^{(1)}) = 175 \times 0.996 = 174.3 = 174$

For $E(\sum_{i=1}^{186} X_i^{(2)}) = 186 \times 0.005 = 0.93 = 1$

I have rounded these numbers to the nearest integer as people can cast either 1 or 0 votes.

Our terms are:

- $B = 175$
- $M = 186$
- $E(\sum_{i=1}^{175} X_i^{(1)}) = 174$
- $E(\sum_{i=1}^{186} X_i^{(2)}) = 1$

Substituting these values gives:

$$\text{B supporters after 1 day} = 175 - (175 - 174) + 1$$

Thus:

$$\text{B supporters after 1 day} = 175$$

Question 4b

Number of M supporters at end of first day:

$$\text{M supporters after 1 day} = \text{M supporters at the start of day} + \text{Supporters who switch to M} - \text{Supporters who switch to B}$$

Therefore:

$$\text{M supporters after 1 day} = M + (B - \sum_{i=1}^{B} X_i^{(1)}) - \sum_{i=1}^{M} X_i^{(2)}$$

Substituting the same values from Question 4a gives:

$$\text{M supporters after 1 day} = 186 + (175 - 174) - 1$$

Thus:

$$\text{M supporters after 1 day} = 186$$

Question 4c

```
# Creating the loops for one day first:

# For the X1 outcomes ----

X1 <- 175  # This will become no. permutations for X1
X1_trials <- numeric(175)  # Where outcomes of each Bernoulli trial will be stored
X1_probs <- c(0.996, 0.004)  # The two probabilities for each X2 trial
results <- c(1, 0) # The corresponding outputs for the probabilities above
no._X1_successes <- 0  # Initial number of successes (Will grow with loops)


for(i in 1:X1){  # Sampling everyone
  X1_trials[i] <- sample(results,
                      size = 1, # 1 Bernoulli per person
                      replace = F,  # Only one decision per person (all people must be sampled)
                      prob = X1_probs)  # Pre-specified probabilities
  if (X1_trials[i] == 1) {  # 'If current trial is a success...'
    no._X1_successes <- no._X1_successes + 1  # '...Increase the number of success by 1
  }
}

print(no._X1_successes)




# For X2 outcomes ----

X2 <- 186  # This will become no. permutations for X1
X2_trials <- numeric(186)  # Where outcomes of each Bernoulli trial will be stored
X2_probs <- c(0.005, 0.996)  # The two probabilities for each X2 trial
results <- c(1, 0) # The corresponding outputs for the probabilities above
no._X2_successes <- 0  # Initial number of successes (Will grow with loops)


for(i in 1:X2){  # Sampling everyone
  X2_trials[i] <- sample(results,
                      size = 1, # 1 Bernoulli per person
                      replace = F,  # Only one decision per person (all people must be sampled)
                      prob = X2_probs)  # Pre-specificed probabilities
  if (X2_trials[i] == 1) {  # 'If current trial is a success...'
    no._X2_successes <- no._X2_successes + 1  # '...Increase the number of success by 1
  }
}


print(no._X2_successes)

No._B <- X1 - (X1 - no._X1_successes) + no._X2_successes  # No. B supporters after Day 1
No._M <- 361 - No._B  # No. M supporters after Day 1









# Now I need to have this iterated 14 times but have 'No._B' and 'No._M' values
# assigned to the 'X1' and 'X1_trials' and 'X2' and 'X2_trials' respectively

# For the 14 days:
no._iterations <- 14  # No. days


X1 <- 175  # This will become no. permutations for X1
X1_trials <- numeric(175)  # Where outcomes of each Bernoulli trial will be stored
no._X1_successes <- 0  # Initial number of successes (Will grow with loops)

X2 <- 186  # This will become no. permutations for X1
X2_trials <- numeric(186)  # Where outcomes of each Bernoulli trial will be stored
no._X2_successes <- 0  # Initial number of successes (Will grow with loops)



for (day in 1:no._iterations) { # Loop for each day (n = 14)
  X1_trials <- numeric(X1)  # Store outcomes for X1 trials
  no._X1_successes <- 0  # Reset success count for X1

  for (i in 1:X1) {
```

```
    X1_trials[i] <- sample(results, size = 1, replace = FALSE, prob = X1_probs)
    if (X1_trials[i] == 1) {
      no._X1_successes <- no._X1_successes + 1
    }
  }

  # For X2 outcomes
  X2_trials <- numeric(X2)  # Store outcomes for X2 trials
  no._X2_successes <- 0  # Reset success count for X2

  for (i in 1:X2) {
    X2_trials[i] <- sample(results, size = 1, replace = FALSE, prob = X2_probs)
    if (X2_trials[i] == 1) {
      no._X2_successes <- no._X2_successes + 1
    }
  }

  # Calculate No._B and No._M
  No._B <- X1 - (X1 - no._X1_successes) + no._X2_successes
  No._M <- 361 - No._B

  # Update X1 and X2 for the next iteration
  X1 <- No._B
  X2 <- No._M
}

print(X1)  # No. B supporters after 14 days
print(X2)  # No M supporters after 14 days









# Now to work out the probability that No._B > No._M, I will nest one more loop
# I will repeat the 14 days experiment thousands of times and
# calculate what proportion of the outputs have No._B > No._M


# Final loop iterations ----
simulations <- 10000  # This sample size willallow the true probability to be inferred confidently
B_more_than_M <- 0  # Intial No. where B > M


X1 <- 175  # This will become no. permutations for X1
X1_trials <- numeric(175)  # Where outcomes of each Bernoulli trial will be stored
no._X1_successes <- 0  # Initial number of successes (Will grow with loops)

X2 <- 186  # This will become no. permutations for X1
X2_trials <- numeric(186)  # Where outcomes of each Bernoulli trial will be stored
no._X2_successes <- 0  # Initial number of successes (Will grow with loops)



# Run the simulation 10000 times
for (sim in 1:simulations) {

  # Run the 14-iteration loop
  for (day in 1:no._iterations) {
    # For X1 outcomes
    X1_trials <- numeric(X1)
    no._X1_successes <- 0

    for (i in 1:X1) {
      X1_trials[i] <- sample(results, size = 1, replace = FALSE, prob = X1_probs)
      if (X1_trials[i] == 1) {
        no._X1_successes <- no._X1_successes + 1
      }
    }

    # For X2 outcomes
    X2_trials <- numeric(X2)
    no._X2_successes <- 0

    for (i in 1:X2) {
      X2_trials[i] <- sample(results, size = 1, replace = FALSE, prob = X2_probs)
```

```
      if (X2_trials[i] == 1) {
        no._X2_successes <- no._X2_successes + 1
      }
    }

    # Calculate No._B and No._M
    No._B <- X1 - (X1 - no._X1_successes) + no._X2_successes
    No._M <- 361 - No._B

    # Update X1 and X2 for the next iteration
    X1 <- No._B
    X2 <- No._M
  }

  # Check if No._B > No._M in the final result and count it
  if (No._B > No._M) {
    B_more_than_M <- B_more_than_M + 1
  }
}

answer <- B_more_than_M/ 10000  # Average number of times B has the majority

print(answer)  # = 0.098
```

In conclusion, the probability that in this election B would hold the majority of the votes is 0.098 (9.8%).

Question 4d

```
# Repeat but change no. iterations to 60
no._iterations_60 <- 60  # No. days

# Final loop iterations ----
simulations <- 10000  # This sample size willallow the true probability to be inferred confidently
B_more_than_M <- 0  # Intial No. where B > M


X1 <- 175  # This will become no. permutations for X1
X1_trials <- numeric(175)  # Where outcomes of each Bernoulli trial will be stored
no._X1_successes <- 0  # Initial number of successes (Will grow with loops)

X2 <- 186  # This will become no. permutations for X1
X2_trials <- numeric(186)  # Where outcomes of each Bernoulli trial will be stored
no._X2_successes <- 0  # Initial number of successes (Will grow with loops)



# Run the simulation 10000 times
for (sim in 1:simulations) {

  # Run the 14-iteration loop
  for (day in 1:no._iterations_60) {
    # For X1 outcomes
    X1_trials <- numeric(X1)
    no._X1_successes <- 0

    for (i in 1:X1) {
      X1_trials[i] <- sample(results, size = 1, replace = FALSE, prob = X1_probs)
      if (X1_trials[i] == 1) {
        no._X1_successes <- no._X1_successes + 1
      }
    }

    # For X2 outcomes
    X2_trials <- numeric(X2)
    no._X2_successes <- 0

    for (i in 1:X2) {
      X2_trials[i] <- sample(results, size = 1, replace = FALSE, prob = X2_probs)
      if (X2_trials[i] == 1) {
        no._X2_successes <- no._X2_successes + 1
      }
    }

    # Calculate No._B and No._M
    No._B <- X1 - (X1 - no._X1_successes) + no._X2_successes
    No._M <- 361 - No._B

    # Update X1 and X2 for the next iteration
    X1 <- No._B
    X2 <- No._M
  }

  # Check if No._B > No._M in the final result and count it
  if (No._B > No._M) {
    B_more_than_M <- B_more_than_M + 1
  }
}

answer <- B_more_than_M/ 10000  # Average number of times B has the majority

print(answer)  # = 0.981
```

Conclusion

P(B>M after 60 days) = 0.981 = 98.1%.

The difference in probability of B winning with the delay is:

$$0.981 - 0.098 = 0.883$$

Therefore the increase in B probability of winning following the delay is:

$$= +88.3\%$$

Question 5a

One important feature of the poisson distribution is that lambda (λ) represents the mean and the sample variance as they should be approximately equal values. Therefore I can consider both as foundations for calculating point estimation. To calculate the mean no. strikes I can do (3sf):

$$\bar{y} = \frac{\sum_{i=1}^{182} yi}{n} = \frac{60}{182} = 0.330$$

For the sample variance (3sf):

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} y_i = \frac{1}{182-1} \times 60 = 0.331$$

I have proven that there can be 2 estimates for lambda. However, the mean is regarded as a better foundation for an estimator as it tends to be more concentrated around the true value of lambda than variance. Therefore, I will continue with the first lambda estimate.

Now I can scrutinise whether my suggested estimator is bias or not:

$\hat{\lambda} = \frac{\sum_{i=1}^{n} yi}{n} = \bar{y}$ Bias: $\text{Bias}(\lambda) = E(\lambda) - \lambda$ $E(\lambda) = E(\frac{\sum_{i=1}^{n} yi}{n}) = \frac{\sum_{i=1}^{n} E(yi)}{n}$ $E(yi) = \lambda$

Therefore: $E(\lambda) = \frac{n \times \lambda}{n} = \lambda$

In turn: $\text{Bias}(\lambda) = E(\lambda) - \lambda = \lambda - \lambda = 0$

I have proved that the estimator is not bias. Furthermore, the sample mean is known to be a consistent estimator of lambda. Now I can calculate the standard error for the estimator.

Standard Error

Standard error = sd of the sample mean:

$$se = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{1}{n}\text{Var}(X)}$$

As previously calculated, the sample variance = 0.331. Therefore :

$$se\ (3\text{sf}) = \sqrt{\frac{1}{182} \times 0.331} = 0.0426$$

Calculating 95% confidence intervals for λ

While the data does not have a normal distribution, I can apply the central limit theorem. The sample size is large (n > 30) so the sample mean will be approximately normal ($\bar{y}$ ~ $N(\lambda, \frac{\lambda}{n})$). For a poisson distribution with a large sample size, we can use the following formula for a approximate confidence interval:

$$95\%\ CI = \left[\bar{y} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\ \bar{y} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$$

In this case α 0.05 but by consulting the z-table, I know that z_1 for a 95% CI is 1.96. Therefore:

$$95\%\ CI = [0.330 - 1.96 \times 0.0426,\ 0.330 + 1.96 \times 0.0426]$$

$$95\%\ CI = [0.247,\ 0.414]$$

Does the confidence interval support the hypothesis at the 5%-level that λ = 1?

At the 5% level, the confidence intervals do not span 1. Therefore the parameter lambda is estimated to fall within this interval 95% of the time. As a result, the intervals do not support the hypothesis at the 5% level. Additionally, it does span any value ≥ 0.5 so if we were to round to the nearest integer, this would be 0, not 1.

Question 5b

I want to find the log of the likelihood function for this discrete random variable. First I will identify the the likelihood of the given probability mass function. Therefore:

$$L(y;\lambda) = L(\frac{e^{-\lambda}\lambda^{y}}{(1-e^{-\lambda})y!}) = \prod_{i=1}^{n}\frac{e^{-\lambda}\lambda^{yi}}{(1-e^{-\lambda})yi!}$$

Now I can discern the log likelihood (including constants):

$$LL(y;\lambda) = ln(L(\lambda)) = ln(\prod_{i=1}^{n}\frac{e^{-\lambda}\lambda^{y_i}}{(1-e^{-\lambda})y_i!}) = \sum_{i=1}^{n} ln(\frac{e^{-\lambda}\lambda^{y_i}}{(1-e^{-\lambda})y_i!})$$

I can use the rule whereby $ln(ab) = ln(a) + ln(b)$. This also means that:

$$LL(y;\lambda) = \sum_{i=1}^{n}(\ln(e^{-\lambda}) + \ln(\lambda y_i) - \ln(1-e^{-\lambda}) - \ln(y_i!))$$

Now I will take the sum of the log of each term ($\sum_{i=1}^{n} ln(x)$) and then sum them together using the $ln(ab) = ln(a) + ln(b)$. This will allow me to decipher which terms are constants and rearrange accordingly afterwards.

$$\sum_{i=1}^{n} ln(e^{-\lambda}) = -n\lambda$$

Using the property: $ln(a^b) = ln(a) \times b$:

$$\sum_{i=1}^{n} ln(\lambda^{yi}) = (\sum_{i=1}^{n} y_i) \times ln(\lambda)$$

$$\sum_{i=1}^{n} ln(1-e^{-\lambda}) = n \times ln(1-e^{-\lambda})$$

$$\sum_{i=1}^{n} ln(yi!) = \sum_{i=1}^{n} ln\,(yi!)$$

Given the aforementioned rule whereby ln(ab) = ln(a) + ln(b), this gives:

$$LL(y;\lambda) = -n\lambda + \sum_{i=1}^{n} y_i \ln\lambda - n\ln(1-e^{-\lambda}) - \sum_{i=1}^{n} ln\,(y_i!)$$

To find the maximum likelihood estimate, I can now differentiate this equation with respect to lambda. I will use partial derivatives because one term depends on both y_i and lambda

$$\frac{d(LL)}{d\lambda} = \frac{d}{d\lambda}(-n\lambda + \sum_{i=1}^{n} y_i \ln\lambda - n\ln(1-e^{-\lambda}) - \sum_{i=1}^{n} ln\,(y_i!))$$

Derivative of each term:

Power rule:

$$\frac{d}{d\lambda}(-n\lambda) = -n$$

Chain rule: ($\sum_{i=1}^{n} y_i$) = constant, outer function = $f(x)$ = ln, inner function = $g(x)$ = lambda:

$$\frac{d}{d\lambda}(\left(\sum_{i=1}^{n} y_i\right)\ln\lambda) = \frac{\sum_{i=1}^{n} y_i}{\lambda}$$

Chain rule: $-n$ = constant, outer function = $f(x)$ = ln, inner function = $g(x) = 1-e^{-\lambda}$:

$$\frac{d}{d\lambda}(-n\ln(1-e^{-\lambda})) = -\frac{ne^{-\lambda}}{1-e^{-\lambda}}$$

Derivating a constant (no dependency on lambda):

$$\frac{d}{d\lambda}(\sum_{i=1}^{n} ln\,(y_i!)) = 0$$

Therefore:

$$\frac{d(LL)}{d\lambda} = -n + \frac{\sum_{i=1}^{n} y_i}{\lambda} - \frac{ne^{-\lambda}}{1-e^{-\lambda}}$$

Because I'm maximizing the log likelihood model, I now equate this to 0:

$$\frac{d(LL)}{d\lambda} = -n + \frac{\sum_{i=1}^{n} y_i}{\lambda} - \frac{ne^{-\lambda}}{1-e^{-\lambda}} = 0$$

The lambda value that maximises the likelihood must satisfy this equation. Now I can rearrange the equation to get the desired form from the question ($\frac{\hat{\lambda}}{1-e^{-\hat{\lambda}}} = \frac{4}{3}$):

$$-n + \frac{\sum_{i=1}^{n} y_i}{\lambda} = \frac{ne^{-\lambda}}{1-e^{-\lambda}}$$

Now I can clear the denominators by multiplying by $\lambda(1-e^{-\lambda})$:

$$-n\lambda(1-e^{-\lambda}) + \sum_{i=1}^{n} y_i(1-e^{-\lambda}) = \lambda ne^{-\lambda}$$

Expanding the brackets gives:

$$-n\lambda + n\lambda e^{-\lambda} + \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} y_i e^{-\lambda} = \lambda ne^{-\lambda}$$

Cancel out $\lambda ne^{-\lambda}$:

$$-n\lambda + \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} y_i e^{-\lambda} = 0$$

A property of the poisson random variable is that the sample mean is given by $\frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$. Therefore, $\sum_{i=1}^{n} y_i = n\bar{y}$. Applying this to the equation gives:

$$-n\lambda + n\bar{y} - n\bar{y}e^{-\lambda} = 0$$

Each term has n in common so factorising gives:

$$-n(\lambda - \bar{y} + \bar{y}e^{-\lambda}) = 0$$

As stated in the question, n > 0 and a must be an integer. Because n does not equal 0, I can divide by -n on both sides:

$$\lambda - \bar{y} + \bar{y}e^{-\lambda} = 0$$

Rearranging for lambda:

$$-\bar{y} + \bar{y}e^{-\lambda} = \lambda$$

Matching the form from the question:

$$y(-1 + e^{-\lambda}) = y(1 - e^{-\lambda}) = \lambda$$

Finally, dividing both side by $(1 - e^{-\lambda})$ isolates the mean:

$$y = \frac{\lambda}{1 - e^{-\lambda}}$$

Now this matches the form of the equation for $\bar{y}$ in the question.

Question 5c

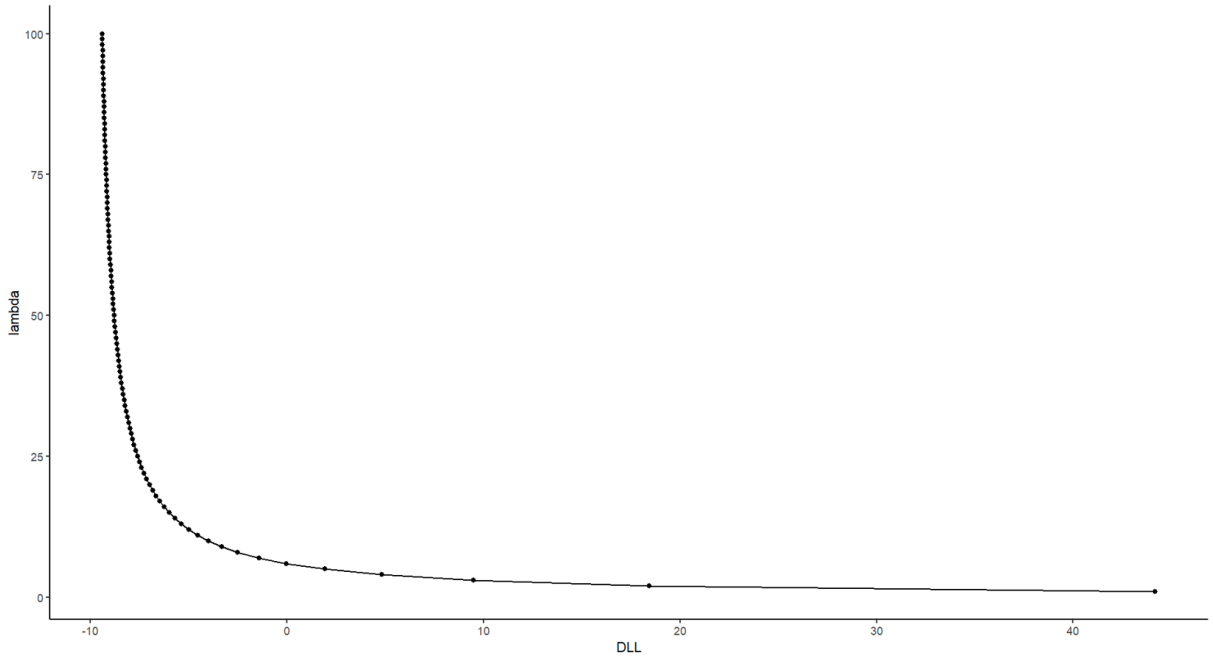
To plot the $-n + \frac{\sum_{i=1}^{n} y_i}{\lambda} - \frac{ne^{-\lambda}}{1-e^{-\lambda}}$:

```
lambdas <- seq(1, 100, length.out = 100) # Creating values from 1-100 increasing by intervals of 1 to represent hypothetical lambda values
n <- 182  # Using given data from the question (n)
sumyi <- 60 # Using data from the question (sum of ys)
DLL_values <-  -n + (sumyi/lambdas) - ((n * exp(-lambdas)) / (1 - exp(-lambdas)))  # Converting the derivative of the LL equation into R


plotting_data <- data.frame(lambda = lambdas, DLL = DLL_values)  # Creating a dataframe of the lambdas and DDL vlaues


plotting_data %>%  # Specifying the dataframe
  ggplot(aes(x = DLL, y = lambda)) +  # Specifying the data
  geom_point() +  # Plotting the  data
  geom_line() +  # Fitting a line that goes through every points
  theme_classic()  # Removing the background
```

This gives:



I could determine the interval (length 1) where DLL becomes positive visually (root) but I will use R to be sure:

```
plotting_data %>%  # Specifying the dataframe
  ggplot(aes(x = DLL, y = lambda)) +  # Specifying the data
  geom_point() +  # Plotting the  data
  geom_line() +  # Fitting a line that goes through every points
  theme_classic()  # Removing the background

plotting_data %>%  # Specifying the data frame
  filter(DLL >=  0) %>%  # Removing all rows where the DLL is > 0
  row_number()  # Printing remaing rows. Now I can workout tje interval spaning 0 (containing the root)
```

The final row number (lambda value) where the DLL is <= 0 is row 5. Therefore, it becomes positive at row 6 meaning that the interval (length 1) containing the root of the function is (5, 6).

Question 5d

```
# Creating the function
help(optimise)  # Can see that a function ('f') is needed

# Creating a function from the absolute value of the function in Q5b
f <- function(lambda){(lambda/1-(exp(-lambda)))- (4/3)}  # The (2) equation function (to be minimised)

# Defining the interval
interval  <- c(5, 6)  # Interval identified in 5c

minimised <- optimise(f,  # Optimising the pre-defined function
         lower = min(interval),  upper = max(interval),
         maximum = F)  # False because I am minimising the function


minimised %>%  # The minimised function
  .$minimum %>%  # The minimimised function minimum
  round(3)  # Rounding to 3 dp
```

Therefore, the MLE $\hat{\lambda}$ is 5.000.

Question 6a

A characteristic of the binomial is that its theoretical mean $= n \times p$ where p is the probability of success.

Therefore, $E(Y) = n \times p$

According to methods of moments, the aforementioned theoretical mean can be equated to the sample mean (y) which is $\frac{\text{No. successes}}{n} = n \times p$

Therefore, $E(Y) = y = n \times p$

Since $y = n \times p$ can be rearranged to give $p = \frac{y}{n}$:

$$E(Y) = \frac{y}{n} = \hat{p}$$

Question 6b

Mean:

Given the methods of moments estimator $T = T(Y) = \frac{Y}{n}$:

$E(T) = E(\frac{Y}{n})$

Referring back to the equation for the theoretical mean of the binomial $(p \times n)$, the mean of $E(\frac{Y}{n})$, and therefore T is:

$$\text{mean of } E(\frac{Y}{n}) = \frac{Y}{n} \times n \times p = Y \times p$$

$Y \times p$ is simply p because Y ($Y = y_i$) must sum to 1 (p + (1-p) = 1). Therefore:

$$E(T) = \bar{T} = p$$

Variance:

The equation for the variance of a binomial is Var(Y) = $n \times p(1-p)$

Method of moments states that $E(Y) = E(Y_i)$.

In turn $Var(Y) = Var(Y_i)$.

and that: $\frac{Var(Y)}{n} = \frac{1}{n^2} \sum_{i=1}^{n} Var(X_i)$

Given this equation and the fact that $E(Y) = E(Y_i)$ this means that:

$$Var(\frac{Y}{n}) = \frac{1}{n^2} Var(Y)$$

Subsisting the aforementioned equation for a binomials variance:

$$Var(\frac{Y}{n}) = \frac{1}{n^2} \times n \times p(1-p)$$

This is equivalent to:

$$Var(\frac{Y}{n}) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

In conclusion:

$$Var(T) = \frac{p(1-p)}{n}$$

The bias of T as an estimator of p

$E(\hat{T})$ should equal T if it is an unbiased estimator.

As previously defined, the $E(T) = p$. T itself is also p since $\frac{Y}{n} = \frac{y}{n} = p$ and therefore the difference between E(T) and T itself is 0 (p-p). Therefore, T is an unbiased estimator (bias = 0).

The mean square error of T as an estimator of p

$MSE(T) = Var(T) + \text{Bias}^2(T)$

A stated, the estimator is unbiased. Therefore, the bias = 0 and the MSE is simply Var(T).

$MSE = Var(T) = Var(\frac{Y}{n}) = \frac{p(1-p)}{n}$

Question 6c

If S is unbiased, $E(S) = E(\theta) = E(p^2)$

This would mean that $E(\frac{Y^2}{n^2}) = p^2$

I can recall the first theoretical moments of a binomial random variable:

1. $E(Y) = n \times p$
2. $E(Y^2) = Var(Y) + (n \times p)^2$

The second moment is denoted as $\frac{1}{n}\sum_{i=1}^{n} Y_i = n \times p(1-p) + n^2p^2$

Since $Y_i = Y$ in method of moments, subsisting gives:

$$E(\frac{Y^2}{n^2}) = \frac{1}{n^2}(n \times p(1-p) + n^2 \times p^2) = \frac{np(1-p) + n^2p^2}{n^2} = \frac{p(1-p)}{n} + p^2$$

The expectation of the parameter ($E(\theta)$) is $E(p^2) = p^2$ / There is a clear difference between the expectation of the parameter and the estimator:

$$\frac{p(1-p)}{n} + p^2 \neq p^2$$

The bias is:

$$\frac{p(1-p)}{n} + p^2 - p^2 = \frac{p(1-p)}{n}$$

$$\text{In conclusion, the bias of the estimator S for } \theta \text{ is } \frac{p(1-p)}{n}.$$

Question 6d

For aS + bT to give an unbiased estimate for θ, E(aS + bT) must equal p^2.

Currently, I know that:

$$ E(T) = p\[10pt]

E(S) = + p^2

$$Therefore, currently:$$

aS + bT = a(+ p^2) + b(p) = a() + a(p^2) + b(p)

$$I want this to sum to p^2 instead since this would match the expectation of \theta.$$

a() + a(p^2) + b(p) = p^2

$$Since E(S) already has a term that's p^2, and I can't split up the terms belonging to S, I can alter the b multiplier of T to cancel out the rest of the bia$$

+ p^2 + b(p) = p^2

$$Isolating b(p) and then the multiplier:$$

b(p) = -p^2 - + p^2

b = - p +

b = -p - p + p

b = - p = - = -

$$This proves that the multiplier b acting on T(p) that makes $aS + bT$'s expectation = p^2 (when a = 1) is $ - \frac{1-p}{n}$. In turn:$$

aS + bT a = 1 b = -= p^2
$$

Question 7a

Exploratory Data Analysis

```
ozone <- read.csv("data/ozone.csv")
summary(ozone)  # Summary statistics
```

This shows me the mean, quartiles and the minima and maxima from each variable.

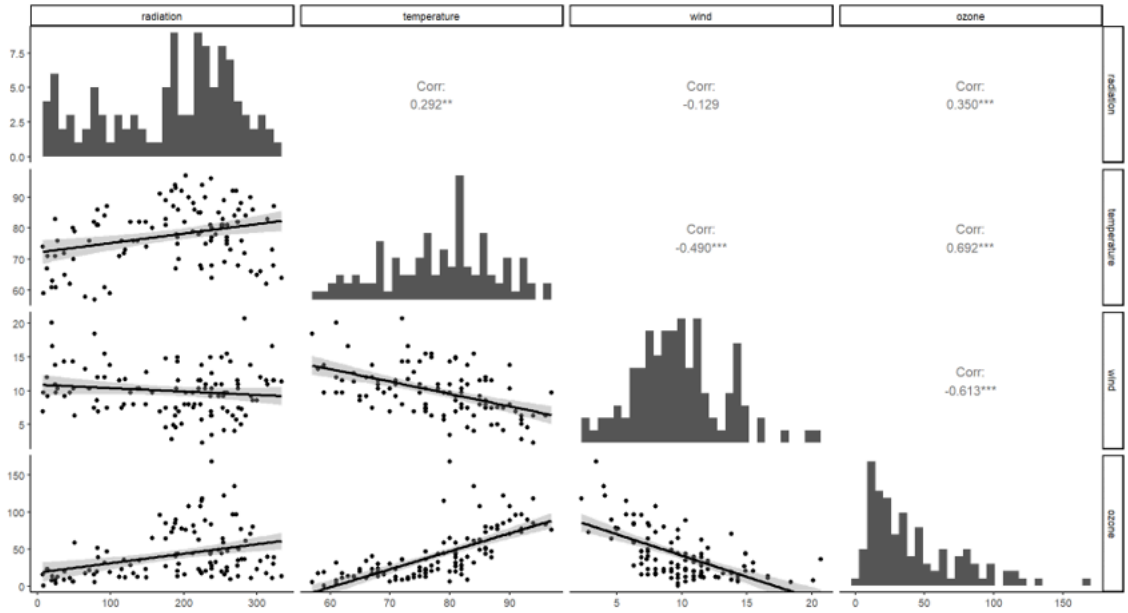
Correlations

```
cor(ozone)  # Correlation coefficients between variables
```

This allows me to see the correlations between the raw data which I visualise in the following scatterplot matrix:

```
ggpairs(ozone,  # Full matrix
        lower = list(continuous = "smooth"),  # Line OBF
        diag = list(continuous = "barDiag"),  # Order
        axisLabels = "show") +  # Axes
  theme_classic()  # Blank background
```

This outputs:



Multiplot Visualising ozone as the response:

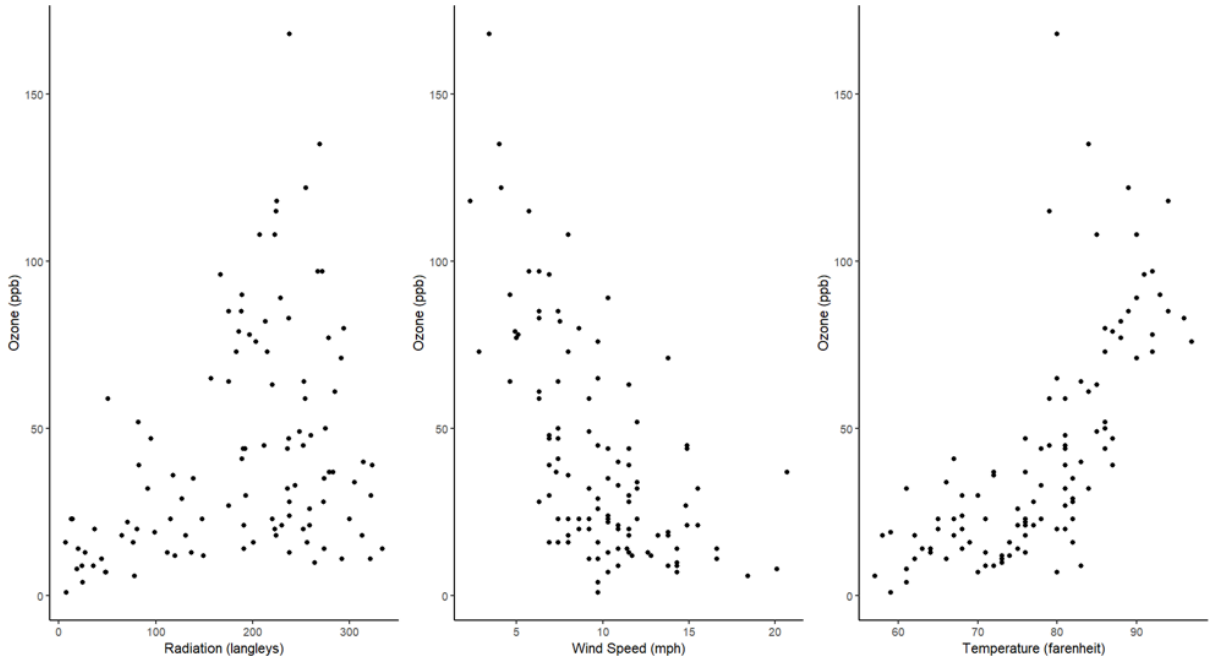
```
# Radiation
radiation_plot <- ozone %>%  # Choosing data frame and creating a plot with the pipe
  ggplot(aes(x = radiation, y = ozone)) +  # Setting variables for axis
  geom_point() +  # Scatterplot
  labs(x = "Radiation (langleys)", y = "Ozone (ppb)") + # Axis Labels
  theme_classic()  # Make background white

# Temperature
temperature_plot <- ozone %>%  # Choosing data frame and creating a plot with the pipe
  ggplot(aes(x = temperature, y = ozone)) +  # Setting variables for axis
  geom_point() +  # Scatterplot
  labs(x = "Temperature (farenheit)", y = "Ozone (ppb)") + # Axis Labels
  theme_classic()  # Make background white

# Wind speed
wind_plot <- ozone %>%  # Choosing data frame and creating a plot with the pipe
  ggplot(aes(x = wind, y = ozone)) +  # Setting variables for axis
  geom_point() +  # Scatterplot
  labs(x = "Wind Speed (mph)", y = "Ozone (ppb)") + # Axis Labels
  theme_classic()  # Make background white


multiplot(radiation_plot, wind_plot, temperature_plot, cols = 3)
```

The output is:



Discussion of findings:

Starting with the relationship between radiation and ozone, there appears to be a slight positive correlation. This is supported by the correlation coefficient of 0.35 seen in the scatter matrix which confirms a weak positive correlation. Furthermore, through visual assessment, it appears that the effect of radiation on ozone appears to be more volatile at higher radiations and suggest that any linearity could be heteroscedsatic with more variation in residuals at higher radiations.

Moving onto the relationship between ozone and windspeed, there appears to be a negative correlation (-0.613). However, the strongest correlation is between temperature and ozone which is positive and has a correlation coefficient of 0.692.

Judging by the correlation coefficients, there does not appear to be any extreme co-linearity between any of the variables despite the fact that wind speed and temperature are almost certainly non-independent.

Question 7b

Creating the linear model

```
model <- lm (ozone ~ temperature + wind + radiation, data = ozone)  # Response and fixed effects
summary(model)
```

The summary of the model which including the three fixed explanatory effects highlights the extent of their relationship with ozone (ppb). Firstly, the effect of temperature is positive as expected; the estimate is 1.6 which means that for every increase in 1 degrees Fahrenheit there is an associated increase in ozone of 1.6 ppb. This supports the initial expectations. This relationship is statistically significant at the 0.05 significance threshold (used for all future analyses), (p <0.0001). This was the most significant result which I also predicted given the initial visualisation.

For wind speed, the estimate of -3.39 also supports the negative correlations and coefficients observed din question 7a. The p-value here is also very significant (p = <0.0001).

Finally, and also unexpectedly, the effect of radiation on ozone was also significant (p < 0.05). While I had identified the weak positive relationship beforehand, which is compounded by the estimate for radiation's effect of 0.06, I had not expected this weak relationship to render a statistically significant finding.

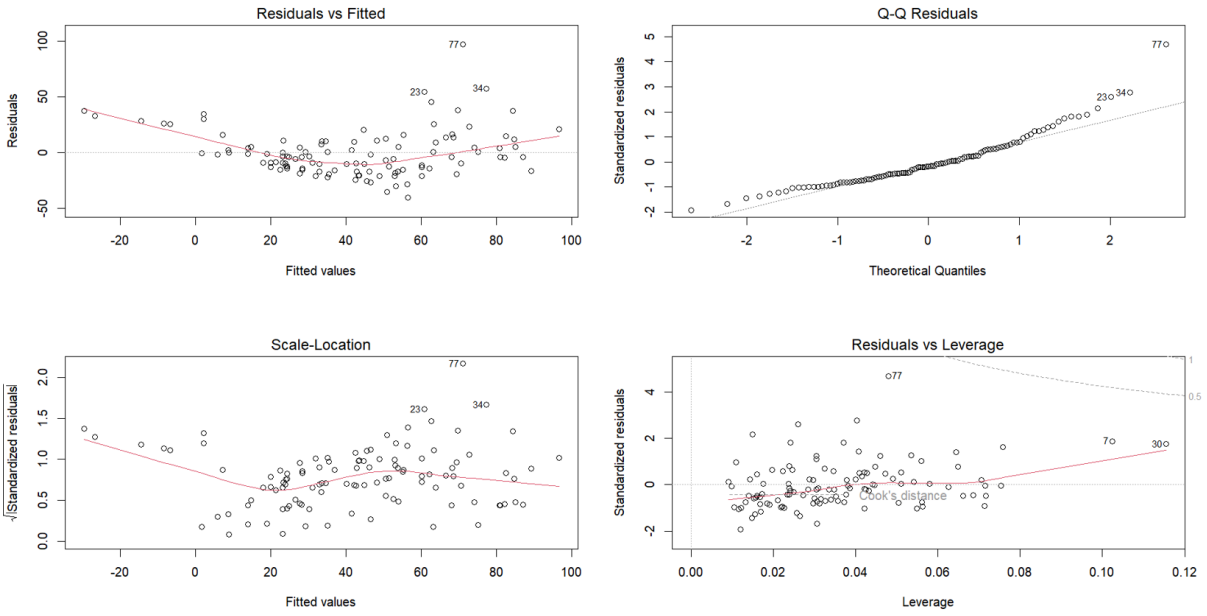
It should be noted that I have not used a linear mixed effects model. However, it could be argued that wind temperature and radiation's effect on the ozone could depend on each variables levels, thus making the current model too simple. Furthermore to test whether the models assumptions have been violated or not, I will check the diagnostic plots.

Model Validity Checks

I generate the diagnostic plots by:

```
par(mfrow = c(2, 2))
plot(model)
```

This gives:



The diagnostic plot give a host of insights into whether the model assumptions have been violated. Firstly, the 'Residuals vs Fitted' values reveal that there is a level of heteroscedasticity in the data. It shows that the spread of residuals across different levels of the fitted values is inconsistent. Therefore the assumption that they are even (the assumption of homoscedasticity) has been violated. In particular, there appears to be a more erratic spread of residuals at the extremes of the fitted values. A similar interpretation can be made for the 'Scale-Location' plot.

Next, the assumption of normality in the residuals distribution has also been violated. This can be seen in the Q-Q plot whereby a number of residuals deviate from the expected normal distribution (along the dotted line). This is most apparent when looking at the data point from the 77th day.

Finally, the 'Residuals vs Leverage' diagnostic plot shows that the there are no residuals that have a extreme and noteworthy effect on the conclusions. Generally, a Cook's score that exceeds 1 is regarded as extreme leverage and yet none exceed 0.5. In turn, I can conclude that there are no residuals with extreme leverage that could have a disproportionate influence on the model outputs

In conclusion, the model's reliability is limited in some respect. Namely, the lack of homoscedasticity and normality in the residuals limits the validity of inferring the results of the model.

Question 7c

```
model_2 <- lm (log(ozone) ~ log(temperature) + log(wind) + log(radiation), data = ozone)  # Log-transform variables
summary(model_2) # Model summary
```

Starting with the model outputs, it is clear to see that log-transforming the response and explanatory effects has affected the coefficients and the corresponding p-values:

Initial Model:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -61.66314   22.97055  -2.684  0.00842 **
temperature   1.62246    0.25295   6.414 3.94e-09 ***
wind         -3.39249    0.65132  -5.209 9.29e-07 ***
radiation     0.06094    0.02318   2.629  0.00982 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

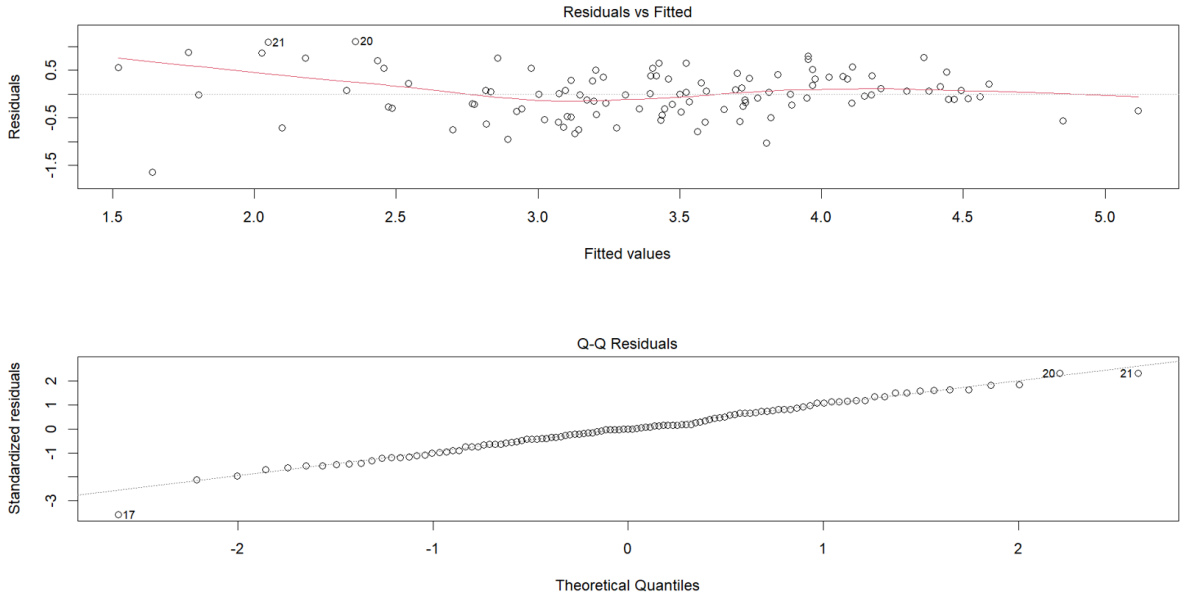
Second Model:

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -10.32605    2.08167  -4.960 2.66e-06 ***
log(temperature)  3.15419    0.45909   6.871 4.40e-10 ***
log(wind)        -0.67650    0.13697  -4.939 2.91e-06 ***
log(radiation)    0.30919    0.05872   5.265 7.29e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Firstly, the modelled effect of wind speed on ozone has been greatly reduced. The coefficent is -0.68 which means that the second model predicts for every 1 mph increase in wind speed, there is a ~2.27 smaller change in the ozone compared to the first model. This is undoubtedly caused by the log transformation. Earlier, I identified data point 77 as a clear outlier. Upon looking at the data, I can see that this data point contains the highest ozone recording (168 ppb) and the second lowest wind speed (3.4 mph). The log transformation would have made this raw data less extreme relative to the rest of the data and therefore the modelled effect of wind speed on ozone has been largely suppressed hence the less extreme coefficient and p-value (2.91e-06 compared to the first models p value of 9.29-07).

In contrast, the log transformation has heightened the perceived impact of radiation and temperature in comparison to the first model with coefficients coefficients increasing from 0.06 to 0.31 and 1.62 to 3.15, respectively. In turn, the model successfully explains more of the random varaition in the data (noise) compared to the previous model. The most likely explanation for this is that the transformation has made the data more linear; it has made the extreme residual values (discussed in the previous question) less extreme relative to the rest of the data. This is evidenced by assessing the diagnostic plots:

```
par(mfrow = c(2, 1))
plot(model_2)
```



In comparison to the former models diagnostic plot, this ‘Residuals vs Fitted’ plot highlights a huge amelioration in the extent of heteroscedasticity in the data. Now the residuals are much more evenly distributed around the flat x axis line. This means that the variance of the residuals is more consistent and that the assumption of homoscedasticity has not been violated to the same extent as before. Furthermore, the linearity of the residuals is much better as can be seen in the Q-Q plot; this means that the residuals follow the normal distribution assumed of linear models. While I have not used statistical tests to confirm whether the model assumptions have been met such as the Bartlett test for homoscedasticity and the Shapiro Wilks for normality, I can at least confirm through visual assessment of the diagnostic plot that the transformations have made the data in the second model more appropriate for a linear model.