Box-Cox Verification

Show that

$$\lim_{\lambda \to 0} \frac{x^\lambda - 1}{\lambda} = \log x, \text{ (for } x \neq 0)$$

For the limit form $\frac{0}{0}$, we should use L'Hopital's rule,

$$\lim_{\lambda \to 0} \frac{x^\lambda - 1}{\lambda} = \lim_{\lambda \to 0} \frac{\frac{d}{d\lambda}(x^\lambda - 1)}{\lambda}$$

For the derivatives, since $\frac{d}{d\lambda}\lambda = 1$ and

$$\frac{d}{d\lambda}(x^\lambda - 1) = \frac{d}{d\lambda}(e^{\log x^\lambda} - 1) = \frac{d}{d\lambda}(e^{\lambda \log x} - 1) = \log x \cdot e^{\lambda \log x} = x^\lambda \log x \tag{1}$$

And we substitute (1) and $\frac{d}{d\lambda}\lambda = 1$ back into the limit,

$$\begin{aligned}
\lim_{\lambda \to 0} \frac{x^\lambda - 1}{\lambda} &= \lim_{\lambda \to 0} \frac{\frac{d}{d\lambda}(x^\lambda - 1)}{\lambda} \\
&= \lim_{\lambda \to 0} \frac{x^\lambda \log x}{1} \\
&= \log x \lim_{\lambda \to 0} x^\lambda \\
&= \log x
\end{aligned}$$

Standardisation, Z-scores, Min-Max scaling

Standard deviation:

$$\sigma_x = \sqrt{\frac{\sum_{j=1}^{n}(x_j - \bar{x})^2}{n-1}}$$

Z-scores:

$$Z_j = \frac{x_j - \bar{x}}{\sigma_x}$$

Min-Max scaling

$$x_j^* = \frac{x_j - min(x)}{max(x) - min(x)}$$

Hypothesis T-tests

R performed two t-tests, one for the intercept and one for the gradient. The hypotheses for the intercept test were

$$\begin{aligned}
H_0 &: \beta_0 = 0 \quad \text{(null hypothesis)} \\
H_\alpha &: \beta_0 \neq 0 \quad \text{(alternative hypothesis)}
\end{aligned}$$

which are tested on a t-distribution with $n - k - 1$ degrees of freedom, where n means the number of sample, k means the number of coefficients, 1 means the intercept.

Linear Model Assumptions test

Use the graphs in Figures 2–4 to determine if the assumptions for fitting a linear model are satisfied. (Make sure you refer to specific Figures in your answers.) In addition, for the independence assumption, come up with at least one reason why the assumption may not be satisfied.

i. Linearity: Satisfied. Using Figure 2 we see the red line is roughly flat and the points are roughly evenly distributed above and below it. [You might also argue that it is not satisfied, since the line is not really that flat. In this case, you should probably mention something like the data points in rows 1 and 4 might want to be investigated.]

ii. Constant variance (homoscedasticity): Satisfied. Using Figure 3 we see the red line is roughly flat and as we move from left to right the points have roughly the same distribution around the line.

iii. Normality: Satisfied. Using Figure 4 we see the bulk of the points are between-1 and 1, and in this section the points lie close to the diagonal. iv. Independence: This can' t be determined from a graph. We need more
information about the details of the dataset to know if independence is satisfied. One problem for this assumption could be the placement of the boxes —if they are too close together, then the pollen inside one box could flow to the other box and be counted twice.

Sensitivity, specificity and accuracy of confusion matrix

Sensitivity:

$$\frac{\text{True Positive}}{\text{Total Positive}}$$

Specificity:

$$\frac{\text{True Negative}}{\text{Total Negative}}$$

Accuracy:

$$\frac{\text{True Positive} + \text{True Negative}}{\text{Total}}$$

AUC-ROC

Based on your results from parts (a)–(c), would you estimate that the AUC (area under the curve) for the ROC curve is closest to:

- **between 0.65 and 0.90**
- **between 0.45 and 0.55**
- **between 0.1 and 0.35**
- **between −0.35 and −0.1**
- **between −0.9 and −0.65**

Justify your reasoning
Between 0.65 and 0.90.
An AUC less than zero is not possible, and since the sensitivity and specificity are both clearly above 0.5, (assuming a smooth curve), the ROC curve must lie well above the diagonal. This gives us an AUC close to one. **What does an AUC value close to 0 tell you about your prediction model? If this was the case, how could we improve the model?**
That the model is predicting the variables the wrong way around. To improve the model, we should switch the prediction to the other class (ie. instead of predicting a diseased patient, we should predict a healthy patient).