Homework 3

Due Thursday, 2/11/21 on Canvas.

Treat this problem set as a report to a client, and do NOT turn in any R, SAS or python code. For example, when reporting coefficient estimates, your answer should just be $\hat{\beta} = 2.47$ and NOT

```
> fit$coefficients[2]
[1] 2.469829453
```

It is up to you as to how many decimal places to report, but please be reasonable and consistent. Two or three decimal places is sufficient for most applications. You should only report what is necessary, i.e. estimates, confidence intervals, P values, plots, etc., in a clear and concise manner.

1. In toxicology, the LD_{50} is the dose that causes a 50% mortality rate. Experiments are often carried out at a sequence of dose levels, $x_0, x_1, x_2, \ldots$, each dose being twice the preceding dose, where we will assume for simplicity that $x_0 = 1$. The model most commonly used in toxicology is linear in log dose. Suppose that the following results have been obtained in an experiment at various multiples of the baseline dose.

$\log_2(x)$	0	1	2	3	4	5
Mortality y/m	0/7	2/9	3/8	5/7	7/9	10/11

In the above table, x is the dose and y/m is the number of deaths (y) occurring in a sample of m individuals.

logit(y/m=pi)=B0 + B1*log2(x), Yi | Xi are indie, pi(log2(Xi)) = P(Yi = 1 | log2(Xi))

(a) Consider a model in which the logit of the mortality rate is linear in log dose. Report the probability model you are assuming, and define all coefficients in your model. Remember to state which observations you are assuming are independent.

(b) Fit the linear logistic model in which the logit of the mortality rate is linear in log dose, and plot the fitted mortality rate and raw mortality fractions y/m against log dose (remember the fitted mortality rate is a continuous function). Do you think the model you assumed in part (a) is reasonable?

(c) Estimating $\gamma = \mathrm{LD}_{50}$.

(i) Obtain an estimate for γ.

(ii) Consider the null hypothesis that $\log_2(\gamma) = 4$ as a sub-model or restriction of the linear logistic model. Fit the sub-model and compute the log likelihood ratio statistic LR(4). If the null hypothesis is correct, what is the approximate distribution of LR(4)? Compute the p-value.

(iii) By plotting the restricted log likelihood against the hypothesized value of $\log_2(\gamma)$, construct a likelihood-based 95% confidence set for γ.

(d) Let $\hat{\pi}(x)$, $y(x)$ and $m(x)$ be the estimated mortality rate, measured number of deaths and number of individuals at dose x. Define

$$\hat{\sigma}^2 = \frac{1}{|\mathcal{I}| - 2} \sum_{x \in \mathcal{I}} \frac{\{y(x) - m(x)\hat{\pi}(x)\}^2}{m(x)\hat{\pi}(x)\{1 - \hat{\pi}(x)\}},$$

where $\mathcal{I}$ is the set of doses considered in the experiment.

(i) Show that $(|\mathcal{I}| - 2)\,\hat{\sigma}^2 = X^2$, where X^2 is Pearson's statistic.

gamma, check transformatiton between gamma and chi squared

(ii) Assuming the model from part (a) is correct, what is the asymptotic distribution of $\hat{\sigma}^2$ as $m(x) \to \infty$ for each $x \in \mathcal{I}$, and what the asymptotic distribution's mean? What in the data from the above table suggest that the sample size may be too small to assume the asymptotic distribution is approximately correct?

sigma hat sqr = variance based on sample datta -emperical variance var(y(x)) - theoretical variance

(iii) Suppose that the model for $\pi(x)$ you considered in part (a) is correct and that $y(x)$ is independent of $y(x')$ for $x \neq x'$. If $\hat{\sigma}^2$ is large, what does this tell you about $\mathrm{Var}\{y(x)\}$ with respect to that which you assumed in part (a)? What could be causing this difference in variance?

(iv) Do you think the confidence interval computed in part (c) would be too narrow or too wide if $\hat{\sigma}^2$ were large? What if $\hat{\sigma}^2$ were small?

(v) What is $\hat{\sigma}^2$ in these data? Assuming the asymptotic distribution from part (ii) is approximately correct, is this value consistent with what you would expect by chance if the model you assumed in part (a) were correct? Explain.