# SMI Assignment 5

Andrew Martin

October 20, 2017

1. Exponential distribution
   $y_i$ independent exponential observations with PDF

$$f(y;\theta) = \frac{1}{\theta}e^{-y/\theta} \quad y \geq 0, \quad \theta > 0$$

   (a) Write down the log-likelihood $\ell(\theta;\mathbf{y})$
   **Solution**

$$\begin{aligned}
\ell(\theta; y) &= \log(L(\theta; y)) \\
&= \log(f(y;\theta)) \\
&= \log\left(\prod_{i=1}^{n} \frac{1}{\theta}e^{-y_i/\theta}\right) \\
&= \log\left(\frac{1}{\theta^n}e^{\sum_{i=1}^{n} -y_i/\theta}\right) \\
&= \log\frac{1}{\theta^n} + \log(e^{-n\bar{y}/\theta}) \\
&= \log\frac{1}{\theta^n} + \frac{-n\bar{y}}{\theta} \\
&= \frac{-n\bar{y}}{\theta} - n\log\theta
\end{aligned}$$

**...**

   (b) Hence find the maximum likelihood estimate $\hat{\theta}$
   **Solution**

$\theta$ which maximises $\ell(\theta; y)$ Maximised when:

$$
\begin{aligned}
\frac{\partial \ell}{\partial \theta} &= 0 \\
&= \frac{n\bar{y}}{\theta^2} - \frac{n}{\theta} \\
&= \frac{n\bar{y} - n\theta}{\theta^2} \\
\implies \theta &= \bar{y}
\end{aligned}
$$

Check it is a maximum:

$$
\begin{aligned}
\frac{\partial^2 \ell}{\partial \theta^2} &= \frac{n\theta - 2ny}{\theta^3} \\
\text{So when } \theta &= y \\
\implies \frac{n\theta - 2n\bar{y}}{\theta^3} &= -yn/y^3 < 0 \\
\implies &\textit{decreasing} \\
\implies &\textit{maximum}
\end{aligned}
$$

Maximised when $\theta = y$ So

$$\hat{\theta} = y$$

...

(c) Find the Fisher information $I_\theta$
**Solution**

$$
\begin{aligned}
I_\theta &= -E\left[\frac{\partial^2 \ell}{\partial \theta^2}\right] \\
&= -E\left[\frac{n\theta - 2n\bar{y}}{\theta^3}\right] \\
&= -\left[\frac{n\theta - 2nE[\bar{y}]}{\theta^3}\right] \\
&= -\left[\frac{n\theta - 2n\theta}{\theta^3}\right] \\
&= -\frac{-n\theta}{\theta^3} \\
&= \frac{n}{\theta^2}
\end{aligned}
$$

...

2. Binomial distribution
   Consider a single binomial observation $x$ from $\text{Bin}(n, \theta)$ with $n$ trials and probability of success $p$

   (a) Find the log-likelihood $\ell(\theta; x)$

   **Solution**

   The $x$ observation is Bernoulli distributed, so the probability mass function is:

$$f(x, p) = \begin{cases} p, & x = 1 \\ 1 - p, & x = 0 \end{cases}$$

   Or alternatively:

$$f(x, p) = \binom{n}{1} p^x (1 - p)^{n-x}$$

   Discard the $\binom{n}{1}$ as it is not relevant.

$$f(\mathbf{x}, \theta) = \theta^x (1 - \theta)^{n-x}$$

$$\begin{aligned} \ell(\theta; x) = \log L(\theta; x) &= \log \left(f(x; \theta)\right) \\ &= \log \left(\theta^x (1 - \theta)^{n-x}\right) \\ &= \log(\theta^x) + \log((1 - \theta)^{n-x}) \\ \ell(\theta; x) &= x \log \theta + (n - x) \log(1 - \theta) \end{aligned}$$

...

   (b) Find the Score function, the Fisher information and the MLE, $\hat{\theta}$

   **Solution**

$$
\begin{aligned}
\text{Score function, } S &= \frac{\partial \ell}{\partial \theta} \\
&= \frac{x}{\theta} + (n-x) * (-\frac{1}{1-\theta}) \\
&= \frac{x}{\theta} + \frac{x-n}{1-\theta} \\
\text{Fisher information} &= -E\left[\frac{\partial^2 \ell}{\partial \theta}\right] \\
&= -E\left[\frac{\partial}{\partial \theta}\left(\frac{x}{\theta} + \frac{x-n}{1-\theta}\right)\right] \\
&= -E\left[\frac{-x}{\theta^2} + \frac{x-n}{(1-\theta)^2}\right] \\
&= -E\left[\frac{-x}{\theta^2}\right] - E\left[\frac{x-n}{(1-\theta)^2}\right] \\
&= \frac{x}{E\left[\theta^2\right]} - \frac{x-n}{E[(1-\theta)^2]} \\
\hat{\theta} &= \theta | S = 0 \\
\frac{x}{\theta} + \frac{x-n}{1-\theta} &= 0 \\
\frac{x}{\theta} &= -\frac{x-n}{1-\theta} \\
x(1-\theta) &= -\theta(x-n) \\
x - x\theta + x\theta - n\theta &= 0 \\
\hat{\theta} &= \frac{n}{x}
\end{aligned}
$$

...

(c) Find expressions for the log-likelihood ratio test statistic, $G^2$, and the score test statistic, $U$, for testing the null hypothesis $H_0 : \theta = \theta_0$ versus $H_A : \theta \neq \theta_0$

**Solution**

Log likelihood-ratio test statistic is given by

$$
\begin{aligned}
G^2 &= -2(\ell(\theta_0; Y) - \ell(\hat{\theta}; Y)) \\
&= -2\left(x \log \theta_0 + (n-x)\log(1-\theta_0) - (x \log \hat{\theta} + (n-x)\log(1-\hat{\theta}))\right) \\
&= -2\left(x \log(\theta_0/\hat{\theta}) + (n-x)\log \frac{1-\theta_0}{1-\hat{\theta}}\right)
\end{aligned}
$$

$$\begin{aligned} U &= \frac{S}{\sqrt{l_{\theta_0}}} \\ &= \frac{\frac{x}{\theta} + \frac{x-n}{1-\theta}}{\sqrt{\frac{x}{E[\theta^2]} - \frac{x-n}{E[(1-\theta)^2]}}} \end{aligned}$$

...

(d) State the asymptotic distributions of $G^2$ and $U$ respectively under $H_0$

**Solution**

If $H_0$ is true U converges to $N(0,1)$

And $G^2$ converges to $\chi_1^2$ ...

3. Poisson distribution
$y_i$ independent $Po(\lambda)$ observations
You may use the log-likelihood, score function and Fisher information for $\lambda$ and the MLE $\hat{\lambda}$ from the lecture notes

(a) State an approximate $100(1-\alpha)\%$ confidence interval for $\lambda$
**Solution**
The Confidence interval is:

$$\left(\hat{\lambda} - z_{\alpha/2}\sqrt{\hat{\lambda/n}}, \quad \hat{\lambda} + z_{\alpha/2}\sqrt{\hat{\lambda/n}}\right)$$

...

(b) Let $\phi = \log \lambda$

i. Write down the log-likelihood $\ell_\phi(\phi; y)$
**Solution**

$$\ell_\phi(\phi; \mathbf{y}) = -ne^\phi + \sum_{i=1}^{n} y_i\phi + \log\left(\prod_{i=1}^{n} \frac{1}{y_i!}\right)$$

...

ii. Hence find the maximum likelihood estimate $\hat{\phi}$ and the Fisher information $i_n(\phi)$
**Solution**

$$\begin{aligned} S(\phi, \mathbf{y}) &= \frac{\partial l}{\partial \phi} \\ &= -ne^\phi + \sum_{i=1}^{n} y_i \end{aligned}$$

And $\hat{\phi}$:

$$\begin{aligned}
\hat{\phi} &= \phi | S = 0 \\
-ne^{\phi} + \sum_{i=1}^{n} y_i &= 0 \\
e^{\phi} &= \frac{\sum_{i=1}^{n} y_i}{n} \\
\phi &= \log(y_i)
\end{aligned}$$

Fisher info:

$$\begin{aligned}
i_n \phi &= -E[\ell''] \\
&= -E[S'] \\
&= -E[-ne^{\phi}] \\
&= ne^{\phi}
\end{aligned}$$

**...**

iii. It was shown that the test statistic to test $H_0 : \lambda = \lambda_0$ is

$$U = \frac{\bar{Y} - \lambda_0}{\sqrt{\lambda_0 / n}}$$

Let $\phi_0 = \log \lambda_0$. Using (ii), show $U'$, the score test for testing $H_0 : \phi = \phi_0$ is the same as $U$

**Solution**

$$\begin{aligned}
U' &= \frac{S}{\sqrt{i_n \phi}} \\
&= \frac{-ne_0^{\phi} + \sum_{i=1}^{n} y_i}{\sqrt{ne_0^{\phi}}} \\
&= \frac{-ne_0^{\phi} + \bar{y}}{\sqrt{ne_0^{\phi}}} \\
&= \frac{\bar{y} - e_0^{\phi}}{\sqrt{\frac{e_0^{\phi}}{n}}}
\end{aligned}$$

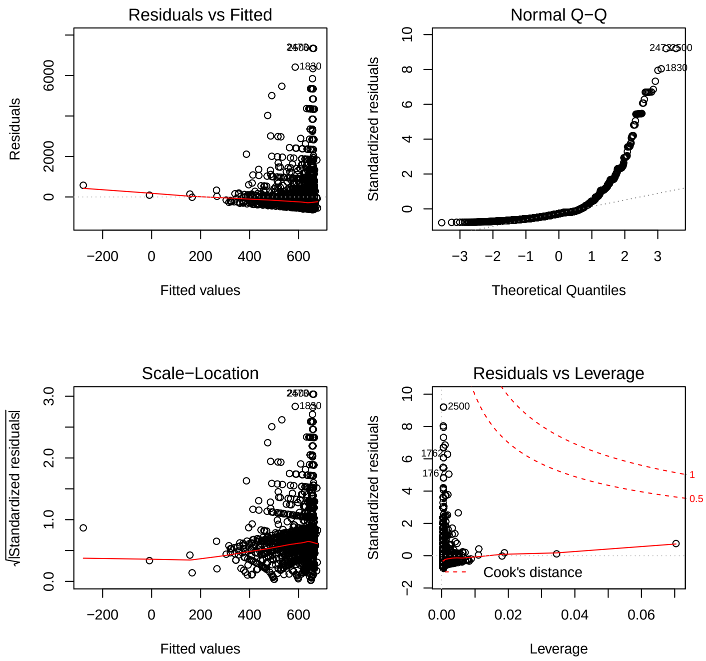


Figure 1: Model Checks for model 1

And since $\phi = \log(\lambda)$

$$\implies U' = \frac{\bar{y} - \lambda_0}{\sqrt{\lambda_0/n}}$$

...

4. Linear modelling

   (a) Load in 2.gumtree.rds

   (b) Only considering dogs that have been sold. I.e. only dogs with price larger than \$10. Filter the data to only include these.

   (c) Fit a model with age as a predictor and price as response (Model 1). Are the assumptions of the linear model reasonable (include plots)

   **Solution**

   The linear model assumptions are: homoscedasticity, normality, linearity and independence.

   i. homoscedasticity - from the scale-location plot, the variance of the data does not appear equal

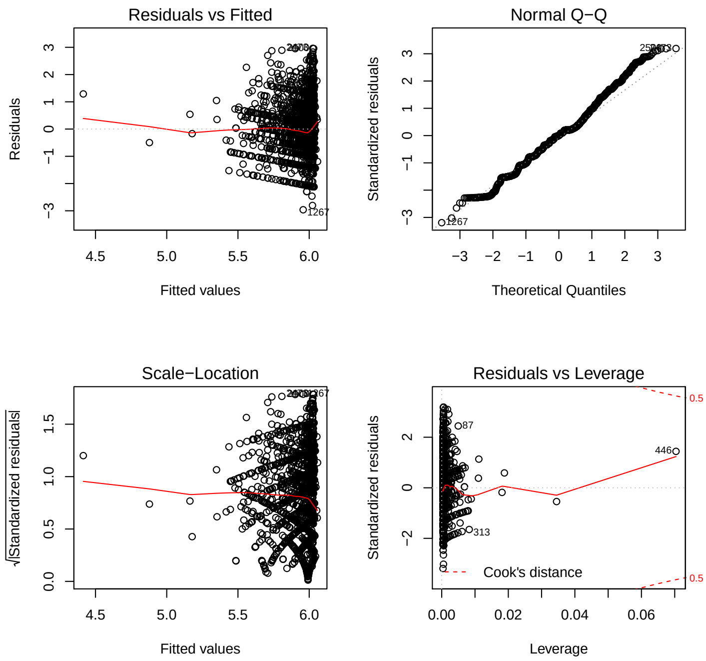


Figure 2: Model Checks for model 2

ii. normality - the residuals vs fitted shows the data is not normal, the red line has negative slope rather than being constant = 0.

iii. linearity - the normal q-q plot shows a very non-linear trend in the data, as it grows somewhat exponentially .

iv. independence - design assumption - assume to be true

From this the assumptions are *not* reasonable. **...**

(d) Fit a model with age as a predictor and log(price) as the response (Model 2). Are the assumptions of the linear model reasonable (include plots)

**Solution**

i. Homoscedasticity: Shown in the scale-location plot, the data appears to have constant variance.

ii. Normality: Shown in the residuals vs fitted plot, the trend is still not a constant zero and appears to 'jump' somewhat - so it is not quite normal.

iii. Linearity: Shown in the normal-q-q plot. The data for this model seems much closer to a linear trend but is not quite perfect.

iv. Independence: this is a key part of the model and must be assumed to be true.

From this the assumptions are *not* reasonable. **...**

(e) Using model 2, calculate a 95% prediction interval for a dog with an age of 1 year

**Solution**

The prediction gives:

| fit | lwr | upr |
|---|---|---|
| 5.987069 | 4.166801 | 7.807337 |

I.e. the prediction interval is

$$(4.166801, 7.807337)$$

**...**

(f) Fit a model with state as a predictor and log(price) as the response variable (Model 3). Are the assumptions of the linear model reasonable (include plots).

**Solution**

Model 3 gives summary statistics (these will be referred to later):

Coefficients :

| | Estimate | Std. Error | **t** value | Pr(>\|**t**\|) | |
|---|---|---|---|---|---|
| (Intercept) | 6.08462 | 0.20212 | 30.103 | <2e−16 | *** |
| stateNA | −0.19459 | 0.20407 | −0.954 | 0.3404 | |
| stateNSW | −0.17601 | 0.20618 | −0.854 | 0.3934 | |
| stateNT | −0.13748 | 0.22741 | −0.605 | 0.5455 | |
| stateQLD | −0.09484 | 0.20700 | −0.458 | 0.6469 | |
| stateSA | −0.11727 | 0.22253 | −0.527 | 0.5983 | |
| stateTAS | −0.34207 | 0.24016 | −1.424 | 0.1545 | |
| stateVIC | 0.43588 | 0.21870 | 1.993 | 0.0464 | * |
| stateWA | −0.07604 | 0.21181 | −0.359 | 0.7196 | |

−−−

i. Homoscedasticity: Shown in the scale-location plot, the data appears to have reasonably constant variance for the different levels.

ii. Normality: Shown in the residuals vs fitted plot, the trend is very close to 0, implying normality.

iii. Linearity: Shown in the normal-q-q plot. The plot appears very similar to model 2, so it is close to linear.

iv. Independence: this is a key part of the model and must be assumed to be true.

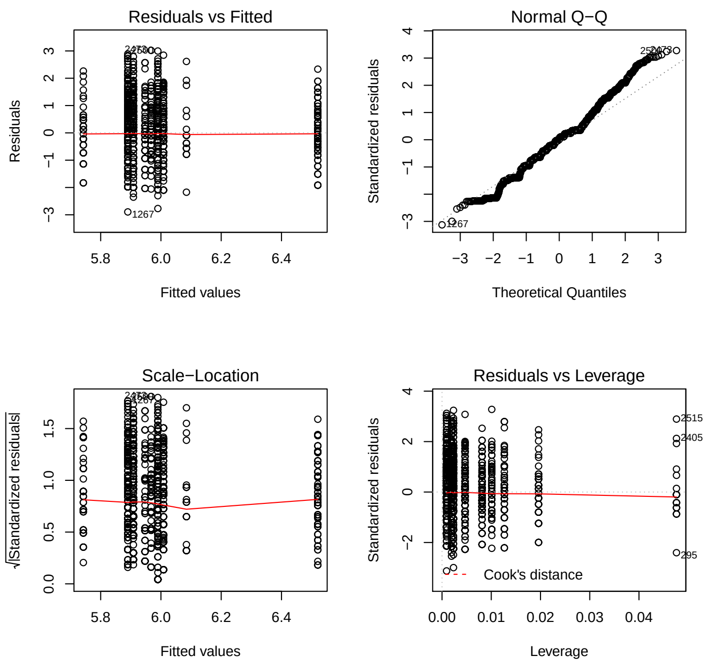


Figure 3: Model Checks for model 3

The assumptions are *reasonable*, though the model will not be perfect.

**...**

**Leave the observations with state NA in the dataset for the model**

(g) In model 3, which state is used as the reference level

**Solution**

ACT is used as the intercept.

**...**

(h) In model 3, which state is significantly different to the reference level at the 5% significance level

**Solution**

Victoria is, with a probability of 0.0464. As is shown in the summary statistics **...**

(i) Using model 3, predict the price of a dog from South Australia. Also give the 95% prediction interval for the price of a dog from South Australia

**Solution**

The R output is:

```
     fit      lwr      upr
1 5.96735 4.14195 7.792751
```

I.e. the predicted value is `5.96735` and the confidence interval is `(4.14195, 7.792751)` ...

The Code used for section 4 is below:

```
library(tidyverse)
library(broom)
setwd("D:/Documents/Uni/Smi")
pdf(file="Graphs.pdf")
##Read in the data
gumtree = readRDS("2.gumtree.rds")

##Filter to only 'sold' dogs
gumtree = gumtree %>%
  filter (price> 10)

##Model 1
model1lm = lm(price~age,data=gumtree)
tmp = par(mfrow = c(2,2))
plot(model1lm)
par(tmp)

##Model 2
model2lm = lm(log(price)~age,data=gumtree)
tmp = par(mfrow = c(2,2))
plot(model2lm)
par(tmp)
#assuming age is in years
newdat = data.frame(age=1)
predict(model2lm, newdata=newdat, interval="prediction")
##Model 3
model3lm = lm(log(price)~state,data=gumtree)
tmp = par(mfrow = c(2,2))
plot(model3lm)
par(tmp)
dev.off()
#price of dog from SA
summary(model3lm)
#95% prediction interval
newdat = data.frame(state="SA")
predict(model3lm, newdata=newdat, interval="prediction")
```