ADVANCED DATA ANALYSIS

HW6

Fan Yang
UNI: fy2232
04/04/2018

Problem 1

(6pt) Suppose T is a life time and it satisfies

$$\log(T) = \mu + \sigma\epsilon$$

where $\epsilon \sim N(0, 1)$.

(a)

(2pt) Give the density of T. What is the name of this distribution?

T satisfies $\log(T) = \mu + \sigma\epsilon$, where $\epsilon \sim N(0, 1)$
Then $\log(T) \sim N(\mu, \sigma^2)$

$$\begin{aligned}
F_T(t) &= Pr(T \leq t) = Pr(\log(T) \leq \log(t)) \\
&= F_N(\log(t)) \\
&\text{where } F_N \text{ is the CDF of } N(\mu, \sigma^2) \\
p_T(t) &= \frac{d\,F_T(t)}{d\,t} \\
&= \frac{d\,F_T(t)}{d\,\log(t)} \frac{d\,\log(t)}{d\,t} \\
&= \frac{d\,F_N(\log(t))}{d\,\log(t)} \frac{d\,\log(t)}{d\,t} \\
&= \frac{1}{t} p_N(\log(t)) \\
&\text{where } p_N \text{ is the PDF of } N(\mu, \sigma^2) \\
p_N(\log(t)) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(\log(t) - \mu)^2}{2\sigma^2} \right\} \\
&\text{Therefore,} \\
p_T(t) &= \frac{1}{t\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(\log(t) - \mu)^2}{2\sigma^2} \right\}
\end{aligned}$$

This is called **Log-normal distribution.**

(b)

(2pt) Find E(T) and Var(T) (hint: see HW 1)

E(T)

$$
\begin{aligned}
E(T) &= \int_0^\infty \ t\, p_T(t)\, dt \\
&= \int_0^\infty \ p_N(\log(t))\, dt \\
&= \int_0^\infty \ t\, p_N(\log(t))\, d\log(t) \\
&= \int_{-\infty}^\infty \ e^x\ p_N(x)\, dx, \qquad \text{let } x = \log(t) \\
&= \int_{-\infty}^\infty \ e^x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx \\
&= E(e^x) \\
&= M(1)
\end{aligned}
$$

where $M(t)$ is the Moment generating function of $N(\mu, \sigma^2)$

$$M(t) = e^{\mu t} e^{\frac{1}{2}\sigma^2 t^2}$$

Therefore,

$$E(T) = e^{\mu} e^{\frac{1}{2}\sigma^2}$$

var(T)

$$
\begin{aligned}
E(T^2) &= \int_0^\infty \ t^2\ p_T(t)\, dt \\
&= \int_0^\infty \ t\, p_N(\log(t))\, dt \\
&= \int_0^\infty \ t^2\ p_N(\log(t))\, d\log(t) \\
&= \int_{-\infty}^\infty \ e^{2x}\ p_N(x)\, dx, \qquad \text{let } x = \log(t) \\
&= \int_{-\infty}^\infty \ e^{2x} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx \\
&= E(e^{2x}) \\
&= M(2)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
E(T^2) &= e^{2\mu} e^{2\sigma^2} \\
var(T) &= E(T^2) - E^2(T) \\
&= e^{2\mu} e^{2\sigma^2} - (e^{\mu} e^{\frac{1}{2}\sigma^2})^2 \\
&= e^{2\mu} e^{2\sigma^2} - e^{2\mu} e^{\sigma^2} \\
&= e^{2\mu} (e^{2\sigma^2} - e^{\sigma^2})
\end{aligned}
$$

(c)

(2pt) If $\mu = 4$ and $\sigma = 3$, find $P(T \leq 100)$.

$$
\begin{aligned}
P(T \leq 100) &= F_T(100) = F_N(\log(100)) \\
&= F\left(\frac{\log(100) - 4}{3}\right)
\end{aligned}
$$

$$
\begin{aligned}
&\text{where } F_N \text{ is the CDF of } N(\mu, \sigma^2) \\
&\text{and } F \text{ is the CDF of } N(0, 1)
\end{aligned}
$$

$$
\begin{aligned}
P(T \leq 100) &= F\left(\frac{\log(100) - 4}{3}\right) \\
&= F(0.20172) \\
&= 0.5799335
\end{aligned}
$$

Problem 2

(4pt) Suppose T is a life time and it satisfies

$$\log(T) = \mu + W/\alpha$$

where $\alpha > 0$ and

$$F_W(w) = 1 - e^{-e^w}$$

Show that T has Weibull distribution and specify it paramters.

$$\log(T) = \mu + W/\alpha \implies W = \alpha(\log(T) - \mu)$$

$$\begin{aligned}
S(t) &= Pr(T > t) = 1 - Pr(T \le t) = 1 - Pr(\log(T) \le \log(t)) \\
&= 1 - Pr\left[\alpha(\log(T) - \mu) \le \alpha(\log(t) - \mu)\right] \\
&= 1 - Pr\left[W \le \alpha(\log(t) - \mu)\right] \\
&= 1 - F_W\left[\alpha(\log(t) - \mu)\right] \\
&= \exp\left\{-e^{\alpha(\log(t) - \mu)}\right\} \\
&= \exp\left\{-t^{\alpha} e^{-\alpha\mu}\right\} \\
&= \exp\left\{-(te^{-\mu})^{\alpha}\right\}, \quad t > 0
\end{aligned}$$

While Weibull distributionhas the survival function

$$S(t) = e^{-(\lambda t)^{\alpha}}, \quad t > 0$$

Let $\lambda = e^{-\mu}$ and $\alpha = \alpha$, then it becomes

$$S(t) = e^{-(e^{-\mu}t)^{\alpha}}, \quad t > 0$$

Therefore T has Weibull distribution with $\lambda = e^{-\mu}$ and $\alpha = \alpha$.

Problem 3

(10pt) Suppose that T has a Weibull distribution with a survival function is given by

$$S(t) = e^{-(\alpha t)^{\beta}}$$

where $\alpha > 0$ and $\beta > 0$. ***(Hint: compute*** $P(T \le t)$***)***

(a)

(2pt) Find the density, $f_T(t)$ ***of*** T

$$\begin{aligned}
S(t) &= Pr(T > t) = e^{-(\alpha t)^{\beta}} \\
\text{then } P_T(T \le t) &= 1 - Pr(T > t) = 1 - e^{-(\alpha t)^{\beta}} \\
f_T(t) &= \frac{d\, P_T(T \le t)}{d\, t} = \frac{d\,[1 - e^{-(\alpha t)^{\beta}}]}{d\, t} \\
&= \frac{1}{t}\left(\beta e^{-(\alpha t)^{\beta}}(\alpha t)^{\beta}\right), \quad t > 0
\end{aligned}$$

(b)

(2pt) Find the hazard function $\lambda(t)$ of T

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{\frac{1}{t}\left(\beta e^{-(\alpha t)^\beta}(\alpha t)^\beta\right)}{e^{-(\alpha t)^\beta}} = \frac{1}{t}\left(\beta(\alpha t)^\beta\right) = \alpha\beta(\alpha t)^{\beta-1}$$

(c)

(2pt) Show that

$$\log(-\log(S(t))) = \beta\log(\alpha) + \beta\log(t)$$

Based on this, describe a graphical method for checking whether or not the data is from a Weibull distribution.

$$\begin{aligned} S(t) &= e^{-(\alpha t)^\beta} \\ \log(S(t)) &= \log(e^{-(\alpha t)^\beta}) = -(\alpha t)^\beta \\ \log(-\log(S(t))) &= \log((\alpha t)^\beta) = \beta\log(\alpha t) \\ &= \beta\log(\alpha) + \beta\log(t) \end{aligned}$$

We can first compute $\log(-\log(S(t)))$, denote it as y; compute $\log(t)$, denote it as x. Then draw the plot y against x.
If y against x follows certain linear pattern, i.e. $y - x$ is a linear line approximately. Then we can conclude the data is from a Weibull distribution.

(d)

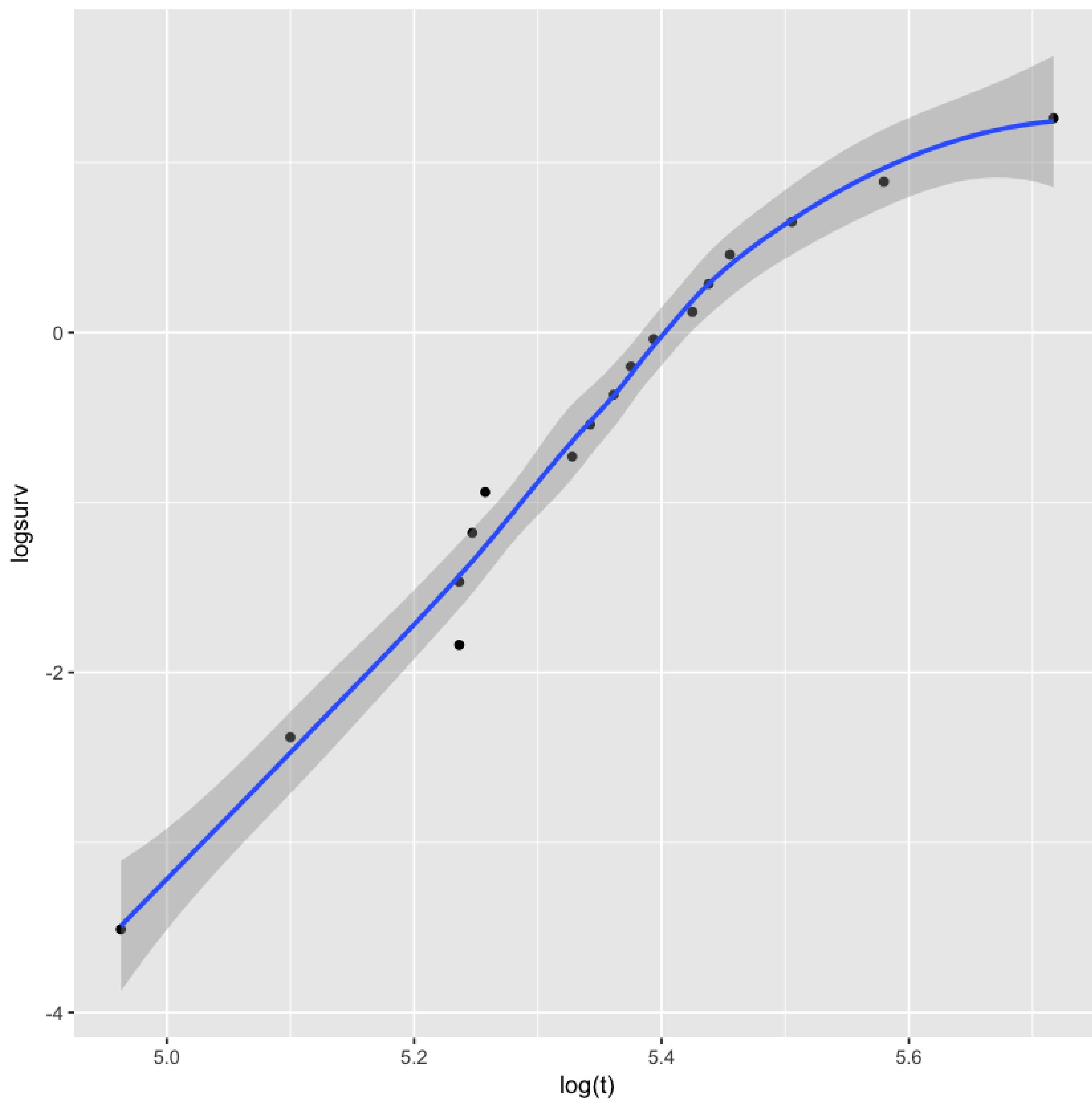
(2pt) Consider the following data
143, 164, 188, 188, 190, 192, 206, 209, 213, 216, 220, 227, 230, 234, 246, 265, 304
and use as an estimate of $S(t(i))$

$$\hat{S}(t(i)) = 1 - (i - 0.5)/n$$

where $t(i)$ is the ith ordered value and n is the sample size. Use the graphical technique in the previous question to check if a Weibull distribution is appropriate for these data

In [1]:
```
library(ggplot2)
data = c(143, 164, 188, 188, 190,
         192, 206, 209, 213, 216,
         220, 227, 230, 234, 246, 265, 304)
surv = 1-(order(data)-0.5)/length(data)
logsurv = log(-log(surv))
ggplot() +
    geom_point(aes(x=log(data),y=logsurv)) +
    geom_smooth(aes(x=log(data),y=logsurv)) +
    xlab("log(t)")
```

```
`geom_smooth()` using method = 'loess'
```



From the plot above, we find that the plot is approximately a linear line, therefore we can conclude that a Weibull distribution is appropriate for these data.

(e)

(2pt) Assume that the Weibull distribution is a good fit, use least squares approach to estimate its parameters.

In [2]:
```
y = logsurv
x = log(data)
fit <- lm(y~x)
summary(fit)
```

```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.68997 -0.12226  0.09174  0.19153  0.30116 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -37.2330     2.1806  -17.07 3.08e-11 ***
x             6.8538     0.4073   16.83 3.80e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2871 on 15 degrees of freedom
Multiple R-squared:  0.9497,	Adjusted R-squared:  0.9463 
F-statistic: 283.1 on 1 and 15 DF,  p-value: 3.796e-11
```

$$y = -37.2330 + 6.8538x$$
$$\text{While } \log(-\log(S(t))) = \beta\log(\alpha) + \beta\log(t)$$

$$\begin{aligned} \text{so } \beta\log(\alpha) &= -37.2330 \\ \text{and } \beta &= 6.8538 \\ \Rightarrow \alpha &= e^{-37.2330/\beta} = e^{-37.2330/6.8538} \\ &= 0.00437 \end{aligned}$$

In summary, $\alpha = 0.00437$ and $\beta = 6.8538$