Quiz #2

Jan. 27, 2017

20 pts

12

Below, we fit a simple liner regression **model1**:

$$\textbf{model1}:\ y_i = \beta_0 + \beta_1 x_i + \varepsilon_i\text{, where } \varepsilon_i \sim normal(0,\ \sigma^2)$$

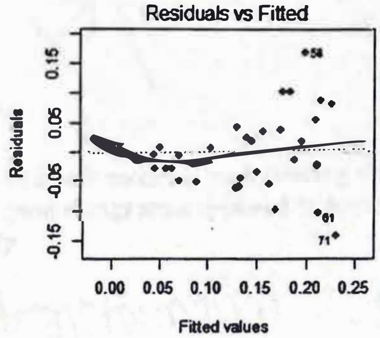
```
> data1 <- read.csv("U:\\STAT510\\lizards.csv")
> dim(data1)
[1] 80 3
> head(data1)
       x      y
1   0.3880 0.1073
2   0.4003 0.0700
3   0.3233 0.0655
4   0.3316 0.0716
5   0.3254 0.1703
6   0.3331 0.0885
> attach(data1)
> model1 <- lm(y~x)
> summary(model1)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.04158    0.00806   -5.159  < 0.001 ***
JAWL         0.52456    0.03109   16.872  < 0.001 ***
---
Residual SE: 0.04285  on  78  degrees of freedom
R-squared: 0.7849, Adjusted R-squared: 0.7822
F-statistic:284.7 on 1 and 78 DF, p-value:< 0.001

> anova(model1)
Analysis of Variance Table

Response: BVOL
           Df   Sum Sq   Mean Sq  F value   Pr(>F)
JAWL        1  0.52278   0.52278  284.65  < 0.001 ***
Residuals  78  0.14325   0.00184
TOTAL      79  0.66603  #This was added by the instructor.
> plot(model1)    #This creates 4 plots. Only the first one is shown.
```

Residuals vs Fitted



(4) (a) In the diagnostic plot shown above, *y*-axis = residuals, *x*-axis = predicted values. <u>*Does the plot look satisfactory*</u>? If NOT, explain briefly what the problem is.

No, there is a megaphone shape, meaning that it doesnt meet the assumption of homogeneity of variance. ✓

```
> logLik(model1)
'log Lik.' 139.4926 (df=3)
```

(4) (b) You must have an assumption about the distribution of the errors in order to calculate (log) Likelihood. What is the default distribution that's being assumed when **R** calculates the (log) likelihood?

Normal distribution ✓

(4) (c) Calculate the **AIC** value of `model1`. Showing **"set up"** (with *numbers*) is enough.
<u>**Hint**</u>: $\mathrm{AIC} = -2 \cdot \log Likelihood + 2(p+1)$.

$-2 \cdot (139.4926) + 2(\cancel{3}+1)$

p=2 not 3

-3

(4) (d) Calculate the **AIC**c value of `model1`. Showing **"set up"** (with *numbers*) is enough.
<u>**Hint**</u>: $\mathrm{AICc} = AIC + \dfrac{2(p+1)(p+2)}{n-p}$.

$$-2 \cdot (139.4926) + 2(3+1) + \frac{2(3+1)(3+2)}{80-\cancel{3}}$$

-3

(4) (e) Theory says to use AICc when $\dfrac{n}{p+1} < 40$. With **model1**, which one should we use: AIC or AICc?

$\dfrac{80}{\cancel{3}+1} < 40$

-2

$\dfrac{80}{4} < 40$

20

Quiz #2 (Keys)

(a) NO it's NOT satisfactory. There is a heterogeneous variance!

(* This can be fixed by the Box-Cox- transformation).

(b) Normal distribution

(c) AIC = $(-2 \times 139.4926) + (2 \times 3) = -272.9852$

(d) $AIC_c = -272.9851 + ((2 \times 3 \times 4)/(80-2)) = -272.67$

(e) AIC_c because $80 \div 3 < 40$

Quiz #3

Feb. 3, 2017

(20) pts

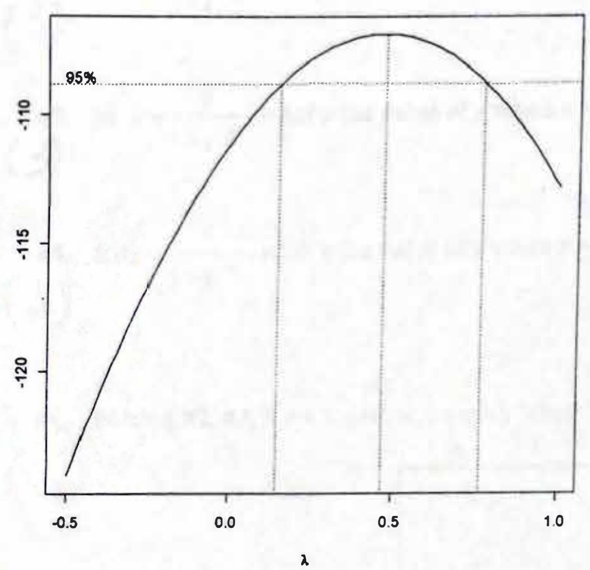
12/20

==

Below, we fit a simple liner regression **model1**:

model1: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where $\varepsilon_i \sim normal(0, \sigma^2)$

```
> data1 <- read.csv("U:\\STAT510\\sampledata1.csv")
> attach(data1)
> model1 <- lm(y~x)
> boxcox(model1)
```



(4) (a) According to the Box-Cox plot, what kind of "action" is appropriate here?

~~log~~ -4

(4) (b) If you fit a new model after following the advice in (a), what good things are supposed to happen? Explain briefly.

the transformation should ~~normalize the~~ stabilize resids. (homogeneity of variance)

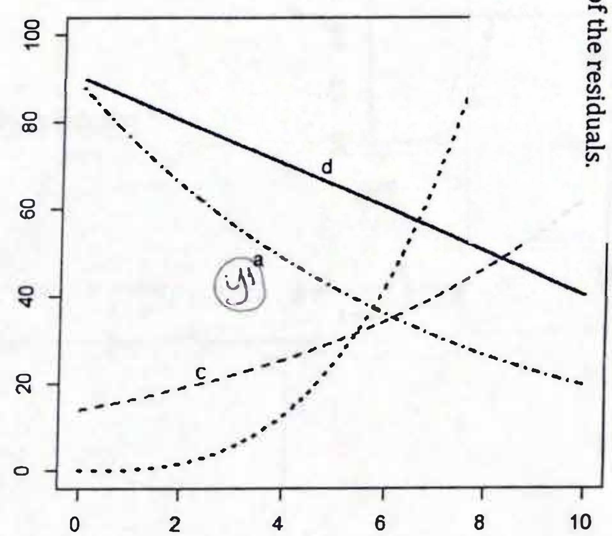
The following four curves are plotted on a single graphing window as shown below.

a $y_1 = 88e^{-0.15x}$

d $y_2 = 90 - 5x$

b $y_3 = 0.2x^3$

c $y_4 = 14e^{0.15x}$



(4) (c) $\lim_{x\to\infty} y_1 = ?$ That is, as $x \to \infty$, where does $88e^{-0.15x}$ go to?

~~20~~ "0" -4

(4) (d) Continuing from (c), which of the four plots would be the graph of y_1? Choose one from a, b, c & d.

a ✓

(4) (e) Which of the four plots would be the graph of y_3? Why? Choose one from a, b, c & d.

b, because the line starts at 0,0 & follows a x^3 curve

Quiz #3 (Keys)

(a) Transform y by square root, i.e., $\sqrt{y}$

(b) Stabilizes the variance of the residuals.

(c) 0

(d) a

(e) b because at x=0, y=0.

12/20

Quiz #3 (Keys)

(a) Transform *y* by square root, i.e., $\sqrt{y}$

(b) Stabilizes the variance of the residuals.

(c) 0

(d) a

(e) b because at x=0, y=0.

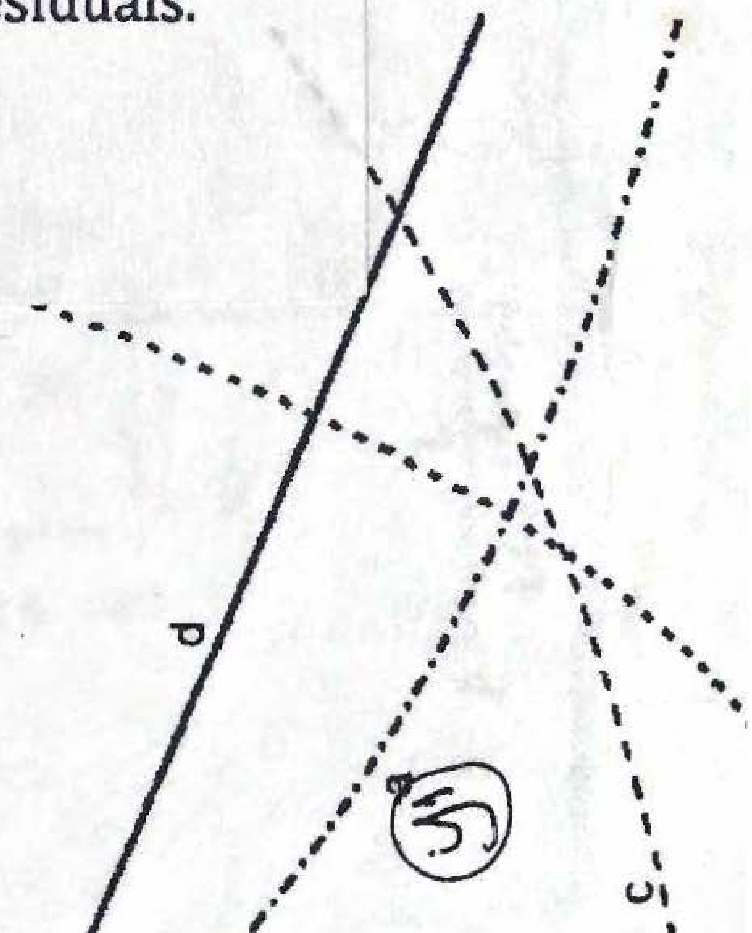
...rves are plotted on a single graphing ...ow.

$y_1 = 88e^{-0.15x}$

$y_2 = 90 - 5x$

$y_3 = 0.2x^3$

$y_4 = 14e^{0.15x}$



Quiz #4

Feb. 10, 2017

(20) pts

U M

===

#1. Which of the following is the same expression as $y = \dfrac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$?

(3)

(a) $y = \dfrac{1}{1+e^{\beta_0+\beta_1 x}}$ (b) $y = \dfrac{1}{1+e^{-(\beta_0+\beta_1 x)}}$ ✓ (c) $y = \dfrac{1}{e^{\beta_0+\beta_1 x}-1}$ (d) $y = \dfrac{1}{e^{-(\beta_0+\beta_1 x)}-1}$

#2. In $y = \dfrac{1}{1+e^{-x}}$, what's the value of y when $x = 0$? $\underline{1/2}$ ✓

(3)

#3. In $y = \dfrac{1}{1+e^{-x}}$, what's the value of y when $x = \infty$? $\underline{1}$ ✓

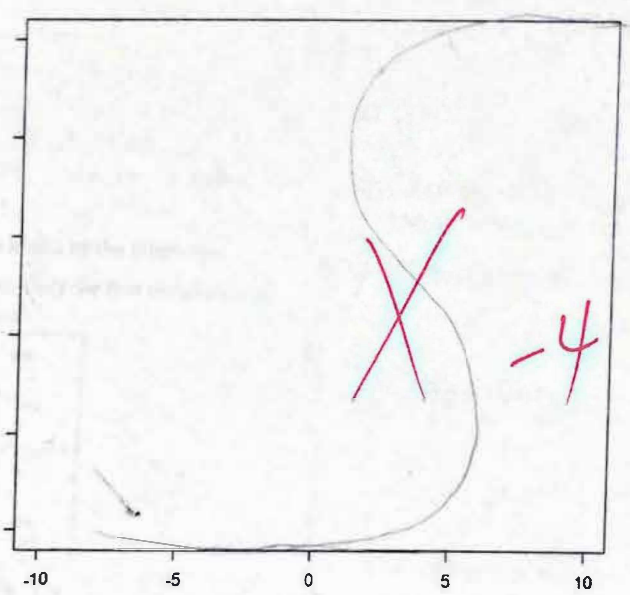
(3)

#4. In $y = \dfrac{1}{1+e^{-x}}$, what's the value of y when $x = -\infty$ (i.e., minus infinity)? $\underline{0}$ ✓

(3)

#5. Putting #2, #3, & #4 together, roughly "sketch" the graph of $y = \dfrac{1}{1+e^{-x}}$.

(4)



#6. Consider $\log\left(\dfrac{p}{1-p}\right) = \beta_0 + \beta_1 x$. Rewrite this expression in terms of p.

(4)

$\frac{p}{1-p} = \frac{\beta_0+\beta_1 x}{\log}$

-4 $\hat{p} = \dfrac{1}{1+e^{-(B_0+B_1x)}}$

Quiz #4

Feb. 10, 2017

20 pts

12/20

===

#1. Which of the following is the same expression as $y = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$?

(3)

(a) $y = \frac{1}{1+e^{\beta_0+\beta_1 x}}$ (b) $y = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$ ✓ (c) $y = \frac{1}{e^{\beta_0+\beta_1 x}}$

#2. In $y = \frac{1}{1+e^{-x}}$, what's the value of y when $x = 0$? _____

(3)

#3. In $y = \frac{1}{1+e^{-x}}$, what's the value of y when $x = \infty$? _____

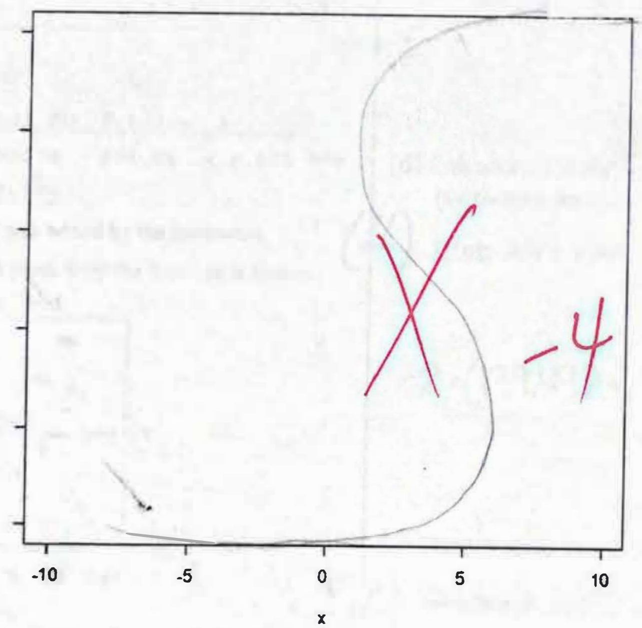
(3)

#4. In $y = \frac{1}{1+e^{-x}}$, what's the value of y when $x = -\infty$ (i.e., minus infinity)? _____

(3)

#5. Putting #2, #3, & #4 together, roughly "sketch" the graph of $y = \frac{1}{1+e^{-x}}$.

(4)



#6. Consider $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$. Rewrite this expression in terms of p.

(4)

$\frac{p}{1-p} = \frac{\beta_0 + \beta_1 x}{\log}$

~~-4~~ $\hat{p} = \frac{1}{1+e^{-(B_0+B_1 x)}}$

Quiz #4 (Keys)

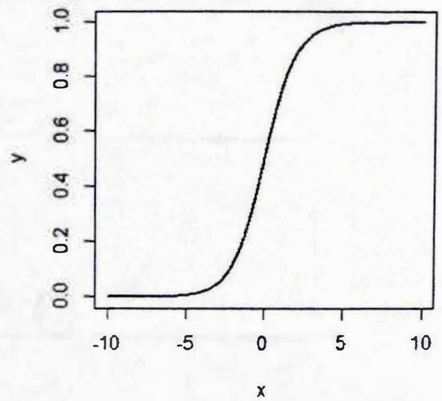
#1 (b) Divide both top and bottom by $e^{\beta_0+\beta_1 x}$

#2 ½ Because $e^0 = 1$

#3 1 Because $e^{-\infty} = 0$

#4 0 Because $e^{\infty} = \infty$

#5



#6 $p = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$

Quiz #5

Feb. 17, 2017

20 pts

Consider the following R codes and printout.

```
> library(faraway)
> data(bliss)
> bliss
  dead alive conc
1    2    28    0
2    8    22    1
3   15    15    2
4   23     7    3
5   27     3    4
> attach(bliss)
> Y <- cbind(dead, alive)
> model1 <- glm(Y~conc, family=binomial(link=logit))
> summary(model1)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.3238     0.4179  -5.561 2.69e-08
conc          1.1619     0.1814   6.405 1.51e-10
---
Dispersion parameter for binomial family taken to be 1
    Null deviance: 64.76327  on 4  degrees of freedom
Residual deviance:  0.37875  on 3  degrees of freedom
AIC: 20.854

> p1 = (dead/(dead+alive))
> p2 = predict(model1,type="response")
> r1 = p1-p2
> round(cbind(p1,p2,r1),4)
      p1     p2      r1
1 0.0667 0.0892 -0.0225
2 0.2667 0.2383  0.0283
3 0.5000 0.5000  0.0000
4 0.7667 0.7617  0.0050
5 0.9000 0.9108 -0.0108
```

#1. Look at the *p*-value for "**conc**", i.e., $\boxed{\texttt{1.51e-10}}$. Explain briefly what this *p*-value means in *plain* terms. No points if you just write "reject H_0" or "do not reject H_0".

(3) The p-value of 1.51e^{-10} for conc. means that the parameter 'conc.' significantly predicts 'y' & should therefore be kept in the model.

#2. Explain what this calculation $e^{1.1619} = 3.196$ tells you. Oh, 1.1619 is the coefficient of "**conc**" in the printout.

(4) It tells you the odds ratio ~~between~~ between what? -2

See Ans. Key

as conc. ↑ by 1 unit, odds for dead become 3.196 x higher

#3. According to the printout, how would you estimate the "**odds**" for "dead" at **conc**=1? Just write the set up, you do NOT need to finish calculation.

(4) R code

exp(-2.3238+1.1619) ✓

#4. According to the printout, how would you estimate the "**probability**" for "dead" at **conc**=1? Just write the set up, you do NOT need to finish calculation.

(3) 1/(1+exp$^{-(-2.3238+1.1619)}$) ✓

#5. What do the two "**deviance**" numbers (i.e., 64.76327 & 0.37875) tell you? Explain briefly where/how these two numbers are used.

(3) These #'s are used to compare the deviance between the null model & the actual ~~obs.~~ deviance

null dev. − resid dev. OK

#6. Shown under **p1**, **p2** and **r1** are "actual" probability, "predicted" probability and "residual", respectively. Which of the following is the most ideal case for the "***residuals***"?

(3)

(a) mostly *positive* residuals
(b) mostly *negative* residuals
(c) random mixture of positive and negative residuals ✓
(d) residuals that change signs constantly, i.e., +, -, +, -, +, -, etc

logit))

z|)
-08
-10

n to be 1
freedom
freedom

xplain
s. No points
o".

conc.
c.
uld
del.

tells you. Oh,
tout.
both what?

-2

#3. According to the printout, how
"**odds**" for "dead" at **conc**=1?
do NOT need to finish calculat

(4)

R code

exp(-2.3238+1.161

#4. According to the printout, how
"**probability**" for "dead" at **co**
up, you do NOT need to finish c

(3)

1/(1+exp^-(-2.323

#5. What do the two "**deviance**" nu
0.37875) tell you? Explain brie
numbers are used.

(3)

These #'s are
the deviance betwe
model & the act
$\boxed{\text{null dev. - resid dev}}$

#6. Shown under **p1, p2** and **r1** are
"predicted" probability and "res
Which of the following is the mo
"*residuals*"?

(3)

(a) mostly *positive* residuals
(b) mostly *negative* residuals

Quiz #5 (Keys)

#1 "**conc**" is highly significant, i.e., it's a very significant variable in modeling the probability of "dead".

#2 It's the odds ratio for 1 unit increase of **conc**, i.e., as **conc** increases by 1 unit, odds for dead become 3.196 times bigger.

#3 $e^{-2.3238+(1.1619\times1)} = 0.3128911$

#4 $\frac{1}{1+e^{-\{-2.3238+(1.1619\times1)\}}} = 0.238322$

#5 They are used to test if the model is valid, i.e., $(64.76327-0.37875)\sim \chi^2_{df=1}$

#6 c

Quiz #6

Feb. 24, 2017

20 pts

Let X be the number of "failures" before the rth "success" in binomial trials. We say X has a <u>*negative binomial*</u> distribution with (r, p) parameters. The pdf of X is given by

$$f(x) = \binom{x+r-1}{r-1} p^r (1-p)^x, \text{ where } x = 0, 1, 2, 3, \dots$$

The mean & variance of X are $\mu = \dfrac{r(1-p)}{p}$, $\sigma^2 = \dfrac{r(1-p)}{p^2}$.

#1. Let X ~ negative binomial (r=3, p=0.5). Find the probability that X=1. Set up is enough, you do NOT need to finish calculation.

(4)

$P(X=1) = (\cancel{3})(1/2)^3(1/2)^1 =$ -4

$f(1) = \begin{bmatrix} 1+3-1 \\ 3-1 \end{bmatrix} 0.5^3(1-0.5)^1$

#2. Let X ~ negative binomial (r=3, p=0.5). Find the ***mean*** and the ***variance*** of X. Finish calculation!

(4)

$\mu = \dfrac{3(1-0.5)}{0.5} = 3$ ✓

$\sigma^2 = \dfrac{3(1-0.5)}{0.5^2} = 6$ ✓

#3. Let X ~ negative binomial (r=3, p=0.5). The **R** command for the negative binomial distribution is `nbinom`. Write **R** codes to compute the probability that X=1.

(3)

pnbinom(1, size=3, prob=1/2) ✓

#4. Let X ~ negative binomial (r=3, p=0.5). Write **R** codes to simulate 1,000 random numbers from a negative binomial distribution with (r=3, p=0.5).

(3)

rnbinom(1000, size=3, prob=1/2) ✓

#5. Suppose we have a dataset with "survival" time for men and women. The three variables used are:

(3)

- `time` = time until death
- `status` = alive(1), dead (2)
- `sex` = male (1), female (2)

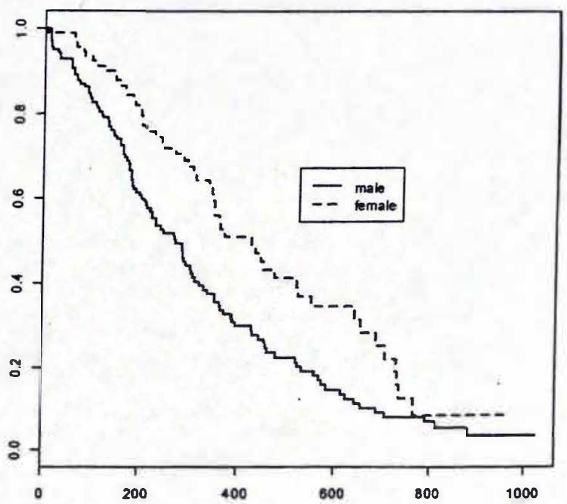
Which of the following are the commonly used distributions for "`time`"? Just <u>***circle*** all the popular distribution name(s)</u> that can handle such a variable as "`time`" here.

(a) lognormal distribution
(b) exponential distribution ✓
(c) Poisson distribution
(d) Weibull distribution

#6. (*Continued* from #5.)

R codes and plot:

```
> surv_time <- Surv(time,status==2)
> model1 <- survfit(surv_time~sex)
> plot(model1,col=c(2,4),lwd=2,lty=1:2)
> legend(locator(1),c("male","female"),lwd=2,
  col=c(2,4),lty=1:2)
```



According to the plot, **whose median survival time is greater: *male* or *female*?** Just circle your answer -- answer alone is enough. ✓

(3)

Quiz #6 (Keys)

#1 $f(1)=\binom{1+3-1}{3-1}0.5^3(1-0.5)^1=\binom{3}{2}0.5^4=\frac{3}{16}=0.1875$

#2 mean=3; var=6

#3 `> dnbinom(1,3,0.5)`

#4 `> rnbinom(1000,3,0.5)`

#5 a, b, d

#6 female

ave a dataset with
he three variables
ne until death
alive(1), dead (2)
le (1), female (2)
following are the c
for "`time`"? Just *cir*
name(s) that can h

mal distribution
ential distribution
on distribution
ll distribution

rom #5.)

lot:

```
<- Surv(time,stat
survfit(surv_time
1,col=c(2,4),lwd=
ator(1),c("male",
,lty=1:2)
```