HW2

Orly Olbum

"Enzyme.txt" contains the data set for problems 1 and 2. If ZinR is a random variable, the notation Z ~ (mu, v) is such that E(Z) = mu and Var(Z) = v.

Problem 1

In an enzyme kinetics study the velocity of a reaction (Y) is expected to be related to the concentration (X) as follows:

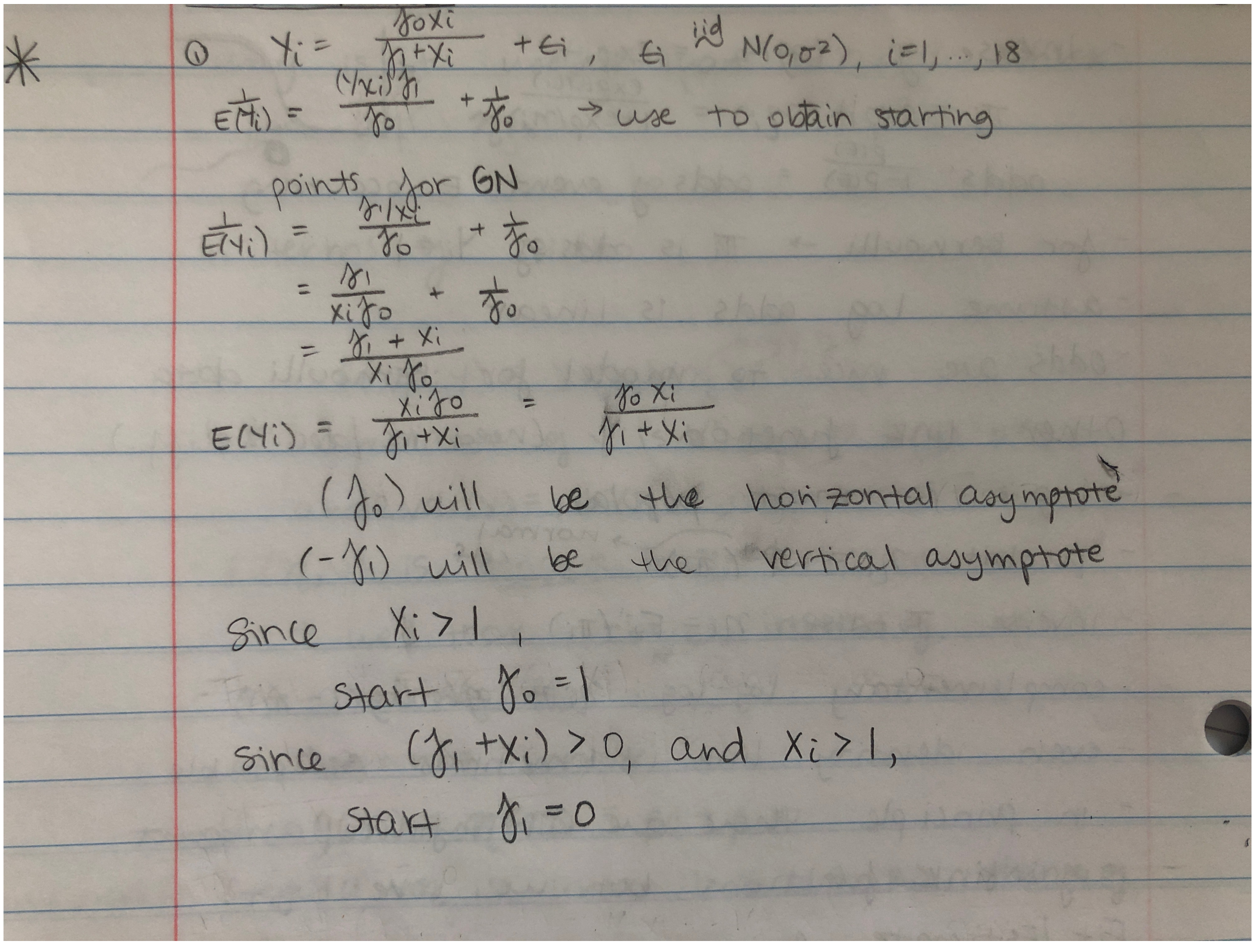
$$Y_i = \frac{\gamma_0 X_i}{\gamma_1 + X_i} + \epsilon_i, \quad \epsilon_i \overset{i.i.d}{\sim} (0, \sigma^2), \quad i = 1, \ldots, n = 18.$$

(a) We must first obtain starting points for Gauss-Newton to be able to estimate gamma0 and gamma1. Observe that

$$1/\mathbb{E}(Y_i) = (1/X_i)\gamma_1/\gamma_0 + 1/\gamma_0.$$

Use this to obtain starting points for Gauss-Newton.

With some fenegling of the function presented above we can see that gamma0 will be our scaling factor and -gamma1 will be our vertical asymptote.



* ① $Y_i = \frac{\gamma_0 X_i}{\gamma_1 + X_i} + \epsilon_i$, $\epsilon_i \overset{iid}{\sim} N(0,\sigma^2)$, $i=1,\dots,18$

$\frac{1}{E(Y_i)} = \frac{(1/X_i)\gamma_1}{\gamma_0} + \frac{1}{\gamma_0}$ → use to obtain starting points for GN

$$\frac{1}{E(Y_i)} = \frac{\gamma_1/X_i}{\gamma_0} + \frac{1}{\gamma_0}$$

$$= \frac{\gamma_1}{X_i\gamma_0} + \frac{1}{\gamma_0}$$

$$= \frac{\gamma_1 + X_i}{X_i\gamma_0}$$

$$E(Y_i) = \frac{X_i\gamma_0}{\gamma_1 + X_i} = \frac{\gamma_0 X_i}{\gamma_1 + X_i}$$

(γ_0) will be the horizontal asymptote

($-\gamma_1$) will be the vertical asymptote

Since $X_i > 1$,

start $\gamma_0 = 1$

Since $(\gamma_1 + X_i) > 0$, and $X_i > 1$,

start $\gamma_1 = 0$

(b) Estimate gamma0 and gamma1 using the starting points obtained in part (a).

Since as X -> infinity Y -> 1/infinity, we see that gamma0 will approach the max value of Y so gamma0 should start at max(Y) = 21.6. The gamma0 starting point can be 0 because the denominator cannot be 0, and since all of Xi is positive, gamma0 has to be > -Xi.

```
model.start = nls(Y ~ gamma0*X / (gamma1 + X), data = enzyme,
                  start = list(gamma0 = 21.6, gamma1 = 0))
gamma.hat = model.start$m$getAllPars()
gamma.hat
```

```
##   gamma0   gamma1
## 28.13704 12.57444
```

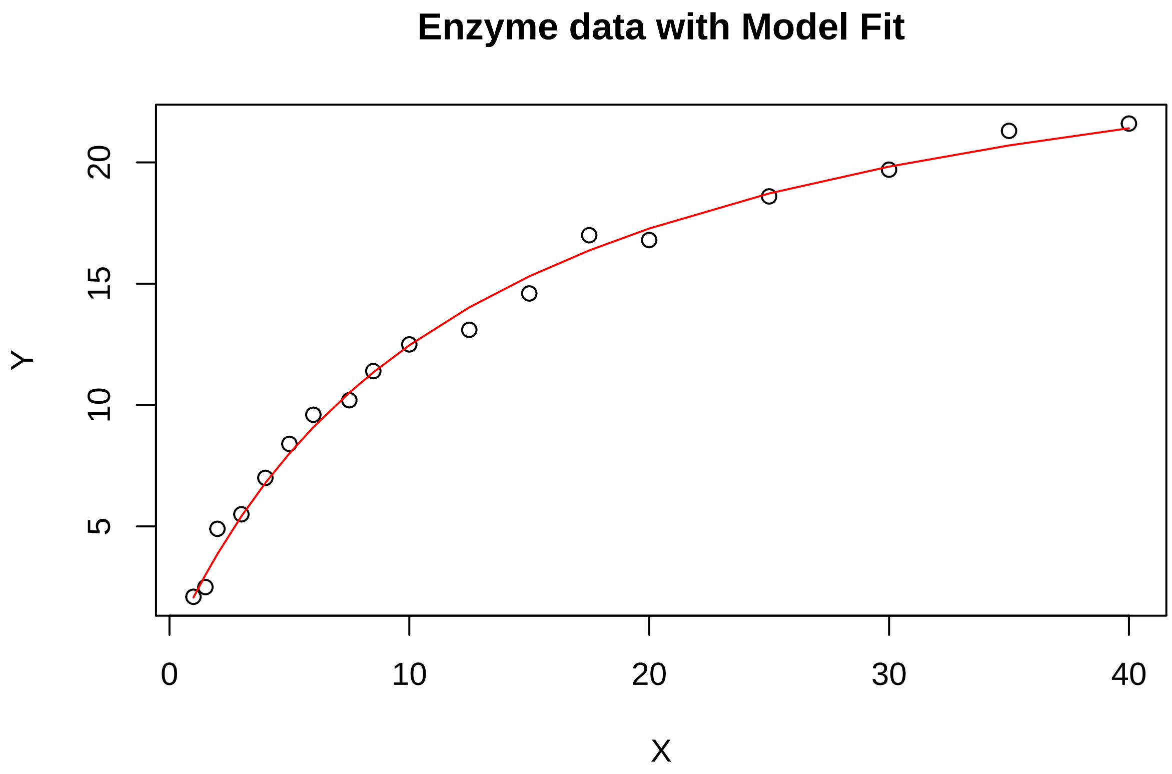
From the starting points and running the model we get gamma0-hat = 28.14 and gamma1-hat = 12.57.

Problem 2

Refer to the analysis of the enzyme kinetics in problem 1:

(a) Plot the estimated nonlinear regression function and data on the same graph. Does the fit appear to be adequate?

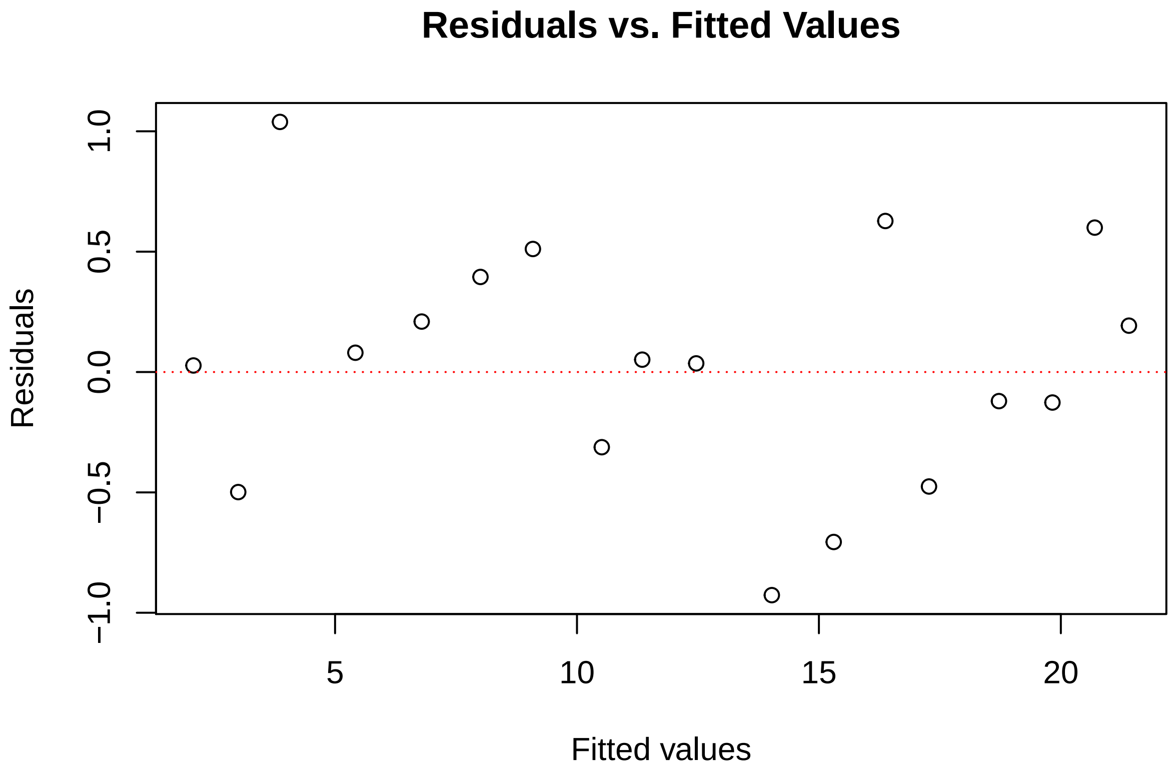
```r
plot(enzyme$X, enzyme$Y, xlab = "X", ylab = "Y", main = "Enzyme data with Model Fit")
lines(enzyme$X, model.start$m$fitted(), col = "red")
```



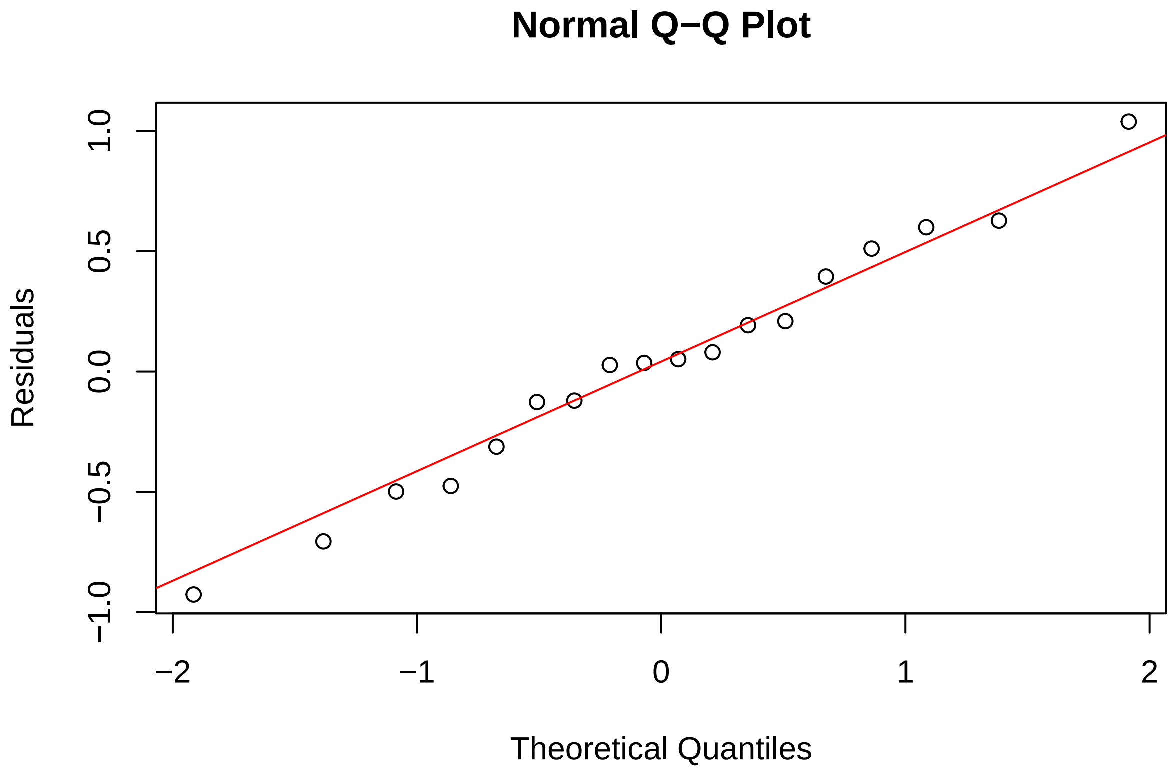
The fit using the starting points above appears to fit the data very well.

(b) Plot the residuals against the fitted values and obtain the normal qq-plot. Comment on the fit of the model.

```r
plot(model.start$m$fitted(), model.start$m$resid(), xlab = "Fitted values",
     ylab = "Residuals", main = "Residuals vs. Fitted Values")
abline(h = 0, col = "red", lty = 3)
```



```r
qqnorm(model.start$m$resid(), ylab = "Residuals")
qqline(model.start$m$resid(), col ="red")
```



The residuals against the fitted values show equal variance, and the normal qq-plot shows the data just about hugs the qq-line, which means our assumptions are satisfied and our model fit is appropriate.

(c) Assume that the fitted model is appropriate and that large sample inference can be employed. Report the test statistic and two-sided p-value of the test of H0: gamma1 = 20.

```
J = model.start$m$gradient()
sigma2 = sum(model.start$m$resid()^2)/(nrow(J) - ncol(J))
se.gamma1 = sqrt(sigma2)*sqrt( solve(t(J)%*%J)[2,2] )
n = 18

gamma1 = gamma.hat[2]
t = (gamma1 - 20) / se.gamma1
p.val = 2*pt(-abs(t), df = n-1)
t; p.val
```

```
##    gamma1
## -9.731382
```

```
##       gamma1
## 2.304276e-08
```

With a test statistic of -9.73 and a p-value of 2.3e-08, we can reject the H0 and conclude that gamma1 is not equal to 20.

##Problem 3

Refer to the analysis of the enzyme kinetics in problems 1 and 2. Perform a bootstrap with 1000 samples, and compute 95% percentile confidence intervals for gamma1. Is it close to the confidence interval based on the large sample theory?

```
gamma_function = function(data, i){
 d2 = data[i,]
 model = nls(Y ~ gamma0*X / (gamma1 + X), data = d2, start = list(gamma0 = 21, gamma1 = 0))
 gammahat = model$m$getAllPars()
 return(gammahat[2])
}

bootstrap_gamma1 = boot(enzyme, gamma_function, R = 1000)
boot.ci(boot.out = bootstrap_gamma1, conf = .95, type = "norm")
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = bootstrap_gamma1, conf = 0.95, type = "norm")
##
## Intervals :
## Level      Normal
## 95%   (11.13, 14.11 )
## Calculations and Intervals on Original Scale
```

```
# large sample theory
CI.gamma1 = gamma.hat[2] + c(-1,1)*se.gamma1*qt(p = 0.975, df = nrow(J)-ncol(J))
CI.gamma1
```

```
## [1] 10.95684 14.19204
```

After 1000 bootstrap samples, the bootstrap CI is (11.24, 14.07). Compared to the CI using the design matrix above, our bootstrap CI is narrower (i.e., better).

STAT 2132
orly olbum

HW2 Probs 4,5

4. $Y \sim$ Bernoulli

$$\text{logit}\{P(Y=1|X)\} = \beta_0 + \beta_1 X \quad \rightarrow \quad P(Y=1|X) = \frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}$$

$Z=1$ if included in enriched study, $=0$ if not

$P(Z=1|Y=1)=\gamma_1$

$P(Z=1|Y=0)=\gamma_0$

$\gamma_1 > \gamma_0 > 0$, individuals selected based on Y (NOT X)

ⓐ Show that

$$\text{logit}\{P(Y=1|X,Z=1)\} = \beta_0^* + \beta_1 X$$

where $\beta_0^* = \beta_0 + \log(\gamma_1/\gamma_0)$

From Bayes:

$$\Pr(B_i|A) = \frac{\Pr(B_i)\Pr(A|B_i)}{\sum_j \Pr(B_j)\Pr(A|B_j)}$$

$B_i \Rightarrow Y$, $A \Rightarrow Z$ (since selection is based on Y, not worried about X)

$$\Pr(Y=1|Z=1) = \frac{\Pr(Y=1)\Pr(Z=1|Y=1)}{\Pr(Y=1)\Pr(Z=1|Y=1) + \Pr(Y=0)\Pr(Z=1|Y=0)}$$

$$\Pr(Y=1) = \frac{\exp(\beta_0+\beta_1 X)}{1+\exp(\beta_0+\beta_1 X)} \qquad \Pr(Y=0) = \frac{1}{1+\exp(\beta_0+\beta_1 x)}$$

$$\Pr(Z=1|Y=1) = \gamma_1 \qquad \Pr(Z=1|Y=0) = \gamma_0$$

$$\Pr(Y=1|Z=1) = \frac{\left[\frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}\right]\gamma_1}{\left[\frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}\right]\gamma_1 + \left[\frac{1}{1+\exp(\beta_0+\beta_1 x)}\right]\gamma_0}$$

$$= \frac{\gamma_1 \exp(\beta_0+\beta_1 X) / [1+\exp(\beta_0+\beta_1 X)]}{\{\gamma_1 \exp(\beta_0+\beta_1 X)/[1+\exp(\beta_0+\beta_1)]\} + \{\gamma_0/[1+\exp(\beta_0+\beta_1 x)]\}}$$

$$= \frac{\gamma_1 \exp(\beta_0+\beta_1 X) / [1+\exp(\beta_0+\beta_1 X)]}{[\gamma_1 \exp(\beta_0+\beta_1 x) + \gamma_0]/[1+\exp(\beta_0+\beta_1 x)]}$$

$$= \frac{\gamma_1 \exp(\beta_0+\beta_1 X)}{\gamma_1 \exp(\beta_0+\beta_1 x) + \gamma_0} = \frac{\exp(\beta_0+\beta_1 X)}{\gamma_0/\gamma_1 + \exp(\beta_0+\beta_1 X)}$$

Solve $\rightarrow$ $\beta_0^* = \beta_0 + \log(\gamma_1/\gamma_0)$

$\beta_0 = \beta_0^* - \log(\gamma_1/\gamma_0)$ $\rightarrow$ plug in:

$$= \frac{\exp(\beta_0^* - \log(\gamma_1/\gamma_0) + \beta_1 x)}{\gamma_0/\gamma_1 + \exp(\beta_0^* - \log(\gamma_1/\gamma_0) + \beta_1 x)} = \frac{\frac{\exp(\beta_0^*)\exp(\beta_1 x)}{\gamma_1/\cancel{\gamma_0}}}{\cancel{\gamma_0/\gamma_1} + \frac{\exp(\beta_0^*)\exp(\beta_1 x)}{\gamma_1/\cancel{\gamma_0}}}$$

$$= \frac{\exp(\beta_0^* + \beta_1 x)}{1+\exp(\beta_0^* + \beta_1 x)} = \text{logit}\{P(Y=1|X,Z=1)\}$$

$$= \beta_0^* + \beta_1 X$$

4, continued

ⓑ can estimated effect of X from an enriched study be used to infer effect in the general population?

In 4ⓐ we showed that β_1 remains the same in the enriched study, which means **we can infer the effect** of X from the enriched study in the entire population

ⓒ can estimated prob of $Y=1|X=x_0$ from an enriched study be used to infer probability in the general population?

Since the $P(Y=1/X=x_0)$ changes when Z is considered, (β_0 changes, as we saw in ⓐ), we cannot use the estimated $Pr(Y=1|X=x_0)$ in the general population, and we have not addressed $Pr(Y=0|X=x_0)$

we would however **be** able to infer odds ratio, which only relies on β_1

5. Y is rv with mgf $M(\theta) = E\{e^{\theta y}\}$

Assume $M(\theta) < \infty$ for all $\theta \in (-\epsilon, +\epsilon)$, $\epsilon > 0$

(ie, all moments of Y exist, $E(Y^k) = M^{(k)}(0)$

ⓐ $K(\theta) = \log\{M(\theta)\}$ → cumulant generating function

Show that $K'(0) = E(Y)$ and $K''(0) = var(Y)$

$$K'(\theta) = \frac{1}{M(\theta)} M'(\theta) = \frac{M'(\theta)}{M(\theta)}$$

$$K'(0) = \frac{1}{M(0)} M'(0) = \frac{1}{E(e^0)} E(X) = E(X) = \mu$$

from mgf properties, $M'(0) = E(Y)$

$$K''(\theta) = \frac{M(\theta) \cdot M''(\theta) - [M'(\theta)]^2}{[M(\theta)]^2}$$

$$K''(0) = \frac{M(0) \cdot M''(0) - [M'(0)]^2}{[M(0)]^2}$$

$$= M''(0) - [M'(0)]^2$$

$$= E(Y^2) - [E(Y)]^2 = \sigma^2 = var(Y)$$

by properties of mgf

ⓑ $f_0(y)$ is density wrt Lebesgue measure

$M(\theta) = \int e^{y\theta} f_0(y)dy$ is mgf

$K(\theta) = \log[M(\theta)]$ is cgf

Assume 0 lies in interior of $R = \{\theta : M(\theta) < \infty\}$

Define family of densities

$$f(y;\theta) \propto e^{\theta y} f_0(y), \quad \theta \in R$$

what is the normalizing constant for $f(y;\theta)$ in terms of θ?

Since we know $\int$pdf $=1$, we have some constant c so that

$\int c f(y;\theta)dy = 1$

$c \int e^{\theta y} f_0(y)dy = 1$

$c = 1/M(\theta)$, which is our normalizing constant

5, continued

ⓒ Show that

$$\ell(y;\theta) = \log\{f(y;\theta)\} = h(y) + \theta y - K(\theta), \theta \in R$$

for some function h that only depends on y.

$$f(y;\theta) \propto e^{\theta y} f_0(y) \quad , \theta \in R$$

$$\log\{f(y;\theta)\} = \log\{c e^{\theta y} f_0(y)\}$$

$$= \log(c) + \log(e^{\theta y}) + \log(f_0(y))$$

$$= \log(\tfrac{1}{M(\theta)}) + \theta y + \log(f_0(y))$$

$$= -\log(M(\theta)) + \theta y + \log(f_0(y))$$

$$= \underset{\underset{-K(\theta)}{\downarrow}}{-K(\theta)} + \underset{\underset{\theta y}{\downarrow}}{\theta y} + \underset{\underset{h(y)}{\downarrow}}{\log[f_0(y)]}$$

ⓓ Y has density $f(y;\theta)$

Show that cgf $K_\theta(t) = K(\theta+t) - K(\theta)$

Use to show $E(Y) = K'(\theta)$, $var(Y) = K''(\theta)$

$$K_\theta(t) = \log\{M_\theta(t)\} = \log\{E(e^{tY})\}$$

$$= \log\{E(e^{(\theta+t-\theta)Y})\}$$

$$= \log\{E(e^{(\theta+t)Y})\} + \log\{E(e^{-\theta y})\}$$

$$= K(\theta+t) - K(\theta)$$

$K(\theta) =$

$$K_\theta(t) = \log M_\theta(t)$$

$$= k_1 t + k_2 \frac{t^2}{2!} + k_3 \frac{t^3}{3!} + \dots$$

$$K'(\theta) = K_\theta'(t) = t + k_2 t + k_3 \cdot \frac{t^2}{2}$$

$$K'(\theta) = K_\theta'(0) = \frac{M_\theta'(\theta)}{M_\theta(\theta)} = E(Y) \quad (\text{from ⓐ})$$

$$K''(\theta) = K_\theta''(0) = \frac{M_\theta(0) \cdot M_\theta''(0) - M_\theta'(0)^2}{M_\theta(0)^2} = \sigma^2 = var(Y) \quad (\text{from ⓐ})$$

5, continued

ⓔ $Y_i \sim f(y; x_i^T\beta)$, $i=1,\dots,n$, $x_i, \beta \in \mathbb{R}^p$, Y_i's independent

$g(\beta) = \sum_{i=1}^{n} \ell(Y_i; x_i^T\beta)$ is log-likelihood

ⓘ Show that $g(\beta)$ is concave

From ⓒ, $\ell(Y_i; x_i^T\beta) = h(y) + \theta y - K(\theta)$

$= h(y_i) + x_i^T\beta y_i - K(\theta)$

$\Rightarrow \sum_{i=1}^{n} \{h(y_i) + x_i^T\beta y_i - K(\theta)\}$

We derive wrt $\beta^{(twice)}$, and are left with the $[x_i^T\beta y_i]$ term which, because it's positive, makes $g(\beta)$ negative semidefinite (and $K''(\theta) = var(Y) \geq 0$)

so $g(\beta)$ is concave wrt β

ⓘⓘ Show that MLE $\hat{\beta}$ satisfies $X^T\{Y - E_{\hat{\beta}}(Y)\} = 0$, where $X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$, $Y = (Y_1, \dots, Y_n)^T$, $E_{\hat{\beta}}(Y)$ is expectation of

Y under $Y_i \sim f(y; x_i^T\hat{\beta})$ for $\forall i = 1, \dots, n$

Derive $g(\beta)$ wrt $\beta = \hat{\beta}$, set to 0

$$\left.\frac{\partial g(\beta)}{\partial \beta}\right|_{\beta=\hat{\beta}} = 0$$

$$\sum_{i=1}^{n} X_i^T Y_i - \sum_{i=1}^{n} X_i^T K'(X_i^T\beta) = 0$$

$$= (X_1, \dots, X_n)^T \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} - \sum_{i=1}^{n} X_i^T E_{\hat{\beta}}(Y_i) = 0$$

$$\Rightarrow (X_1, \dots, X_n)^T \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} - [X_1, \dots, X_n]^T E_{\hat{\beta}}(Y_i) = 0$$

$$\Rightarrow X^T\{Y - E_{\hat{\beta}}(Y)\} = 0$$