Biostats - Regression and NonLinear MLE

Rob Leonard (robleonard@tamu.edu (mailto:robleonard@tamu.edu))

1) Linear Regression - Calcium Data

1a) Write the Design Matrix
X =

$X_{i,1}$	$X_{i,2}$
1	$Calcium_1$ = 1
...	...
1	$Calcium_{10}$ =1
1	$Placebo_1$ = 0
...	...
1	$Placebo_{11}$ = 0

1b) Fit regression model, report coefficient estimates and std. errors For the Placebo treatment: coeff is -0.27, std. error is 2.23 For the Calcium treatment: coeff is 5.27, std. error is 3.23

```
caldata = read.delim("calcium.txt", header=TRUE, sep="\t", stringsAsFactors=TRUE)
caldata$Treatment = ifelse(caldata$Treatment=="Placebo",0,1)
model1 = lm(Decrease ~ Treatment, data=caldata)
summary(model1)$coef
```

```
##               Estimate Std. Error    t value  Pr(>|t|)
## (Intercept) -0.2727273   2.226614 -0.1224852 0.9038010
## Treatment    5.2727273   3.226670  1.6341082 0.1186968
```

1c) What is the p-value for the null hypothesis of no treatment effect?
The p-value of no treatment effect is 0.119. This is for H_0: $Beta_1$ = 0.

1d) Repeat using a two-sample t-test, assuming equal variances.

```
table(caldata$Treatment) # get sample sizes for Calcium (10) and Placebo (11) groups
```

```
##
##  0  1
## 11 10
```

```
sig0 = sd(caldata$Decrease[11:21]) # std dev of Placebo group
sig1 = sd(caldata$Decrease[1:10]) # std dev of Calcium group
poolsig = sqrt(((10*(sig0)^2)+(9*(sig1)^2))/(11+10-2))# pooled std dev
ybar0 = mean(caldata$Decrease[11:21]) # mean of Placebo group
ybar1 = mean(caldata$Decrease[1:10]) # mean of Calcium group
TS = ((ybar1-ybar0)-0)/(poolsig*sqrt((1/10)+(1/11)))  # calculate test statistic
pval = 2*(1-pt(abs(TS),(11+10-2)))
pval
```

```
## [1] 0.1186968
```

The p-value of 0.119 is the same p-value obtained using regression. The results are the same.

1e) What is the estimated distribution of Decrease when Treatment = 1?

E[Y|Treatment=1] = E[$Beta_0$ + $Beta_1$*1 + e_i]= E[$Beta_0$] + E[$Beta_1$] + E[e_i]= $Beta_0$ + $Beta_1$

Var[Y|Treatment=1] = apply def of variance = E[$Beta_0 + Beta_1 * 1 + e_i]^2$-E[$(Beta_0 + Beta_1 * 1 + e_i)^2$]

=> E$(Beta_0^2 + 2Beta_0 Beta_1 + 2Beta_0 e_i + 2Beta_1 e_i + Beta_1^2 + e_i^2) - (Beta_0^2 + 2Beta_0 Beta_1 + Beta_1^2)$

$[E(Beta_0^2) - Beta_0^2] + [E(Beta_1^2) - Beta_1^2] + E(2Beta_0 e_i) + E(2Beta_1 e_i) + E(2Beta_0 Beta_1) - 2Beta_0 Beta_1 + [E(e_i^2) - E(e_i)^2]$

=> $var(Beta_0) + var(Beta_1) + var(e_i)$=$sigma^2$

So the distribution of Decrease when Treatment=1 is **Normal(mean=$Beta_0$ + $Beta_1$, var=$sigma^2$)**

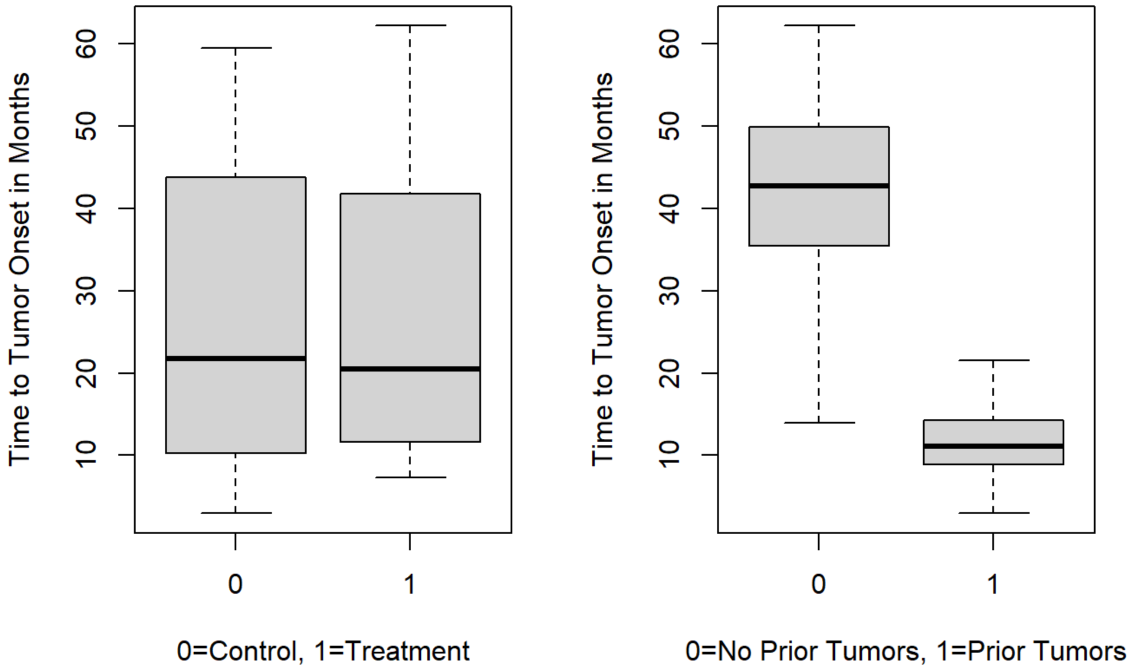
```
cat("Normal(",paste(summary(model1)$coef[1]+summary(model1)$coef[2]),",",paste(round(summary(model1)$sigma,3)),
"^2)")
```

```
## Normal( 5 , 7.385 ^2)
```

2) Tumor Onset

2a) Create Box Plots

```
onsetdata = read.delim("onset_data.csv", header=TRUE, sep=",")
par(mfrow=c(1,2))
boxplot(onset~tx, data=onsetdata, ylab="Time to Tumor Onset in Months", xlab="0=Control, 1=Treatment")
boxplot(onset~prior, data=onsetdata, ylab="Time to Tumor Onset in Months", xlab="0=No Prior Tumors, 1=Prior Tumo
rs")
```

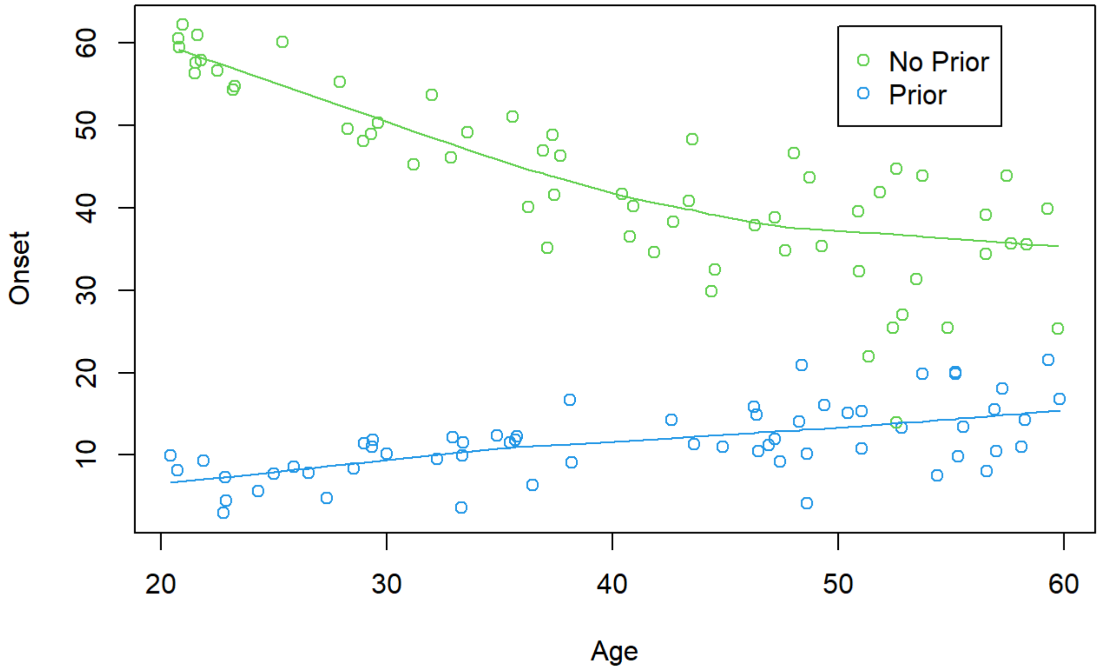


The distributions of the time to onset of a tumor (in months) is very similar for both the control and treatment groups. The treatment group has a similar median to the control group, but does have slightly more right skew, but doesn’t look significant.

The distribution of the time to onset for those without a prior tumor compared to those with a prior tumor is very different. The median time to onset for those without a prior tumor is much larger, over 40 months, while the time for those with a prior tumor is much shorter, with a median of just over 10. The spread of the distribution for those without a prior is also much wider than those with a prior tumor.

2b) Create a scatter plot of onset vs. age.

```
plot(onsetdata$age,onsetdata$onset, col=as.integer(onsetdata$prior+3), xlab="Age", ylab="Onset")
legend(x=50,y=62,c("No Prior","Prior"),col=c(3,4),pch=c(1,1))
lowessP = lowess(onsetdata$age[onsetdata$prior==1],onsetdata$onset[onsetdata$prior==1])
lines(lowessP,col=4)
lowessNP = lowess(onsetdata$age[onsetdata$prior==0],onsetdata$onset[onsetdata$prior==0])
lines(lowessNP,col=3)
```



2c) Fit regression model and interpret coefficients, report estimates and std. errors

```
model2 = lm(onset~tx+prior+age+prior:age, data=onsetdata)
summary(model2)$coef
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  70.9164387 2.29513166  30.898636 1.591689e-57
## tx            2.1115406 0.91917896   2.297203 2.341646e-02
## prior       -69.2409193 3.21786773 -21.517640 3.951810e-42
## age          -0.7167661 0.05333879 -13.437988 2.521552e-25
## prior:age     0.9299603 0.07541030  12.332007 8.883129e-23
```

All explanatory variables are highly statistically significant (very low p-values).
The coefficient for tx indicates that for those getting treatment, we expect the mean time to tumor onset to be 2 months longer than for those not getting treatment, controlling for the other explanatory variables.

For those without a prior tumor, we expect the mean tumor onset time to increase by 69.24 months compared to those who had a prior tumor, controlling for the other factors.

Older patients are expected to have a shorter mean tumor onset time of 0.72 months for each additional year of age as compared to younger patients, controlling for the other variables.

For those who had a prior tumor, age also plays a role in expected mean tumor onset time. For each additional year of age, tumor onset is expected to occur 0.93 months sooner than patients who are 1 year younger.

2d) Use matrix manipulation using a design matrix to verify the estimates and standard errors from above. First verify the coefficients:

```
desmat = data.matrix(model.matrix(model2)) # get design matrix from the model
ymod2 = onsetdata$onset  # get y repsonse vector
betahat = solve(t(desmat)%*%desmat)%*%t(desmat)%*%ymod2
betahat # Coefficient Estimates
```

```
##                    [,1]
## (Intercept)  70.9164387
## tx            2.1115406
## prior       -69.2409193
## age          -0.7167661
## prior:age     0.9299603
```

Then verify the standard errors:

```
residm2 = residuals(model2)
mseM2 = (sum(residm2^2)/(model2$df))

varM2 = mseM2*solve(t(desmat)%*%desmat) #
diagsM2 = diag(varM2)
stderrm2 = diagsM2^.5
stderrm2 # these match the std. error of lm model2 output
```

```
## (Intercept)          tx       prior         age   prior:age
##  2.29513166  0.91917896  3.21786773  0.05333879  0.07541030
```

2e) 95% confidence interval for the mean difference in onset times between the treatment and control groups

```
(lowbound=betahat[2]-qt(0.975,model2$df-1)*stderrm2[2])
```

```
##        tx
## 0.2906542
```

```
(upbound=betahat[2]+qt(0.975,model2$df-1)*stderrm2[2])
```

```
##       tx
## 3.932427
```

The confidence interval is **(0.291,3.93)**.

2f) 95% confidence interval for the mean response of a treated individual, age 35, with no prior tumor incidence?

```
data2f = data.frame(tx=1, prior=0, age=35)
predict(model2, data2f, interval="confidence", level=0.95)
```

```
##        fit      lwr      upr
## 1 47.94117 46.25858 49.62375
```

The 95% confidence interval is **(46.26, 49.62)**.

3) Derive MLE for sigma

$$N(0,\sigma^2)\,pdf = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y-o)^2}{2\sigma^2}}$$

$$L(\sigma^2|y_1,\dots,y_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y_i)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi\sigma}}\right)^n \prod_{i=1}^{n} e^{-\frac{(y_i)^2}{2\sigma^2}} = (2\pi)^{-n/2} * (\sigma)^{-\frac{n}{2}} * e^{-\frac{\left(\sum_{i=1}^{n} y_i^2\right)}{2\sigma^2}}$$

take ln of both sides

$$\ln\left(L(\sigma^2|y_1,\dots,y_n)\right) = -\frac{\mathrm{n}}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma) - \left(\frac{\left(\sum_{i=1}^{n} y_i^2\right)}{2\sigma^2}\right)$$

$$\frac{\partial}{\partial\sigma^2}\ln\left(L(\sigma^2|y_1,\dots,y_n)\right) = 0 - \frac{n}{2\sigma^2} + \left(\frac{\left(\sum_{i=1}^{n} y_i^2\right)}{2\sigma^4}\right) = set\ to\ 0$$

$$\frac{\left(\sum_{i=1}^{n} y_i^2\right)}{2\sigma^4} = \frac{n}{2\sigma^2} \Rightarrow \boldsymbol{\hat{\sigma}^2} = \frac{\left(\sum_{i=1}^{\boldsymbol{n}} \boldsymbol{y_i^2}\right)}{\boldsymbol{n}}$$

4) Gamma Distribution - Time to Infection

$$f(x_1, \ldots, x_n) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x_1^{\alpha-1} e^{-\frac{x_1}{\beta}} * \cdots * \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x_n^{\alpha-1} e^{-\frac{x_n}{\beta}}$$

$$= \left(\frac{1}{\Gamma(\alpha)\beta^{\alpha}}\right)^{n} \left(\prod_{i=1}^{n} x_i^{\alpha-1}\right) e^{-\frac{1}{\beta}\sum_{i=1}^{n} x_i}$$

$$\ln[f(x_1, \ldots, x_n)] = \ln\left(\prod_{i=1}^{n} x_i^{\alpha-1}\right) - \frac{1}{\beta}\sum_{i=1}^{n} x_i - \ n * \ln(\Gamma(\alpha)\beta^{\alpha})$$

from Method of Moments an STAT 641

$$\hat{\alpha} = \frac{\bar{Y}^2}{E[Y^2]-\bar{Y}^2} \quad and\ \hat{\beta} = (E[Y^2] - \bar{Y}^2)/\bar{Y}$$

Read in datafile

```
data4 = read.delim("gamma.csv", header=TRUE, sep=",")

mean4 = mean(data4$x)  # calc mean
meansq4 = (mean4)^2    # calc mean squared
var4 = var(data4$x)    # calc variance
ahat = meansq4/var4    # find an intial alpha value
bhat = var4/mean4      # find an initial beta value

funct4 = function(abvec) {a=abvec[1];b=abvec[2]; lglk=log(prod((x)^(a-1)))-(1/b)*sum(x)-n*log(gamma(a)*b^a);
return(-lglk)}     # set up function to minimize

n = length(data4$x)
x = data4$x

out=nlm(funct4,c(ahat,bhat))
```

```
## Warning in nlm(funct4, c(ahat, bhat)): NA/Inf replaced by maximum positive value
```

```
out
```

```
## $minimum
## [1] 222.6099
##
## $estimate
## [1] 2.232943 1.772003
##
## $gradient
## [1] -1.781971e-06 -3.272019e-06
##
## $code
## [1] 1
##
## $iterations
## [1] 10
```

$\hat{alpha}$ = 2.233 $\hat{beta}$ = 1.772