STAT645 - Homework 2

Salih Kilicli

9/3/2019

Problem 1: With the calcium data in "calcium.txt," consider the Decrease variable as your response and Treatment as your treatment. In what follows, I have recoded Treatment to equal 0 for placebo and 1 for calcium treatment.

(a) For the regression model $Decrease_i = \beta_0 + \beta_1 Treatment_i + \epsilon_i$, write down the model matrix.

The model can be defined as $Y = X\beta + \epsilon$ where the model matrix, X, is given by:

$$X = \begin{bmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{bmatrix}_{21x2}, \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ y_{21} \end{bmatrix}_{21x1}, \qquad \epsilon = \begin{bmatrix} e_1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ e_{21} \end{bmatrix}_{21x1}, \qquad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}_{2x1}$$

Notice, above X is a 21×2 matrix whose second column consists of 10 ones and 11 zeros representing calcium treatment and placebo treatment, respectively. In R;

```
setwd("~/Desktop/STAT645/Data")
calc <- read.table("calcium.txt", header=TRUE)
attach(calc)
Dummy=rep(0, 21)
Dummy[which(Treatment=="Calcium")] = 1
model0 = lm(Decrease ~ Dummy, data = calc)
model.matrix(model0)
```

```
##    (Intercept) Dummy
## 1            1     1
## 2            1     1
## 3            1     1
## 4            1     1
## 5            1     1
## 6            1     1
## 7            1     1
## 8            1     1
## 9            1     1
## 10           1     1
## 11           1     0
## 12           1     0
## 13           1     0
## 14           1     0
## 15           1     0
## 16           1     0
## 17           1     0
## 18           1     0
## 19           1     0
## 20           1     0
## 21           1     0
## attr(,"assign")
## [1] 0 1
```

```
detach(calc)
```

(b) Fit the above model, and report the coefficient estimates and standard errors.

```
setwd("~/Desktop/STAT645/Data")
calc <- read.table("calcium.txt", header=TRUE)
model1 = lm(Decrease ~ Treatment, data = calc)
summary(model1)
```

```
##
## Call:
## lm(formula = Decrease ~ Treatment, data = calc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.7273  -4.7273  -0.7273   5.0000  13.0000
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.000      2.335   2.141   0.0454 *
## TreatmentPlacebo   -5.273      3.227  -1.634   0.1187
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.385 on 19 degrees of freedom
## Multiple R-squared:  0.1232, Adjusted R-squared:  0.07708
## F-statistic:  2.67 on 1 and 19 DF,  p-value: 0.1187
```

(c) Based on the model, what is the p-value for the null hypothesis of no treatment effect? #meaning $H_0 : \beta_1 = 0$

```
p = summary(model1)[["coefficients"]][2,4]
cat("p-value for the null hypothesis of no treatment effect is p =", p)
```

```
## p-value for the null hypothesis of no treatment effect is p = 0.1186968
```

(d) Now analyze the same data using a two-sample t-test, assuming equal variances. How do the results compare to those you obtained using the regression model?

```
setwd("~/Desktop/STAT645/Data")
calc <- read.table("calcium.txt", header=TRUE)
attach(calc)
model1 = lm(Decrease ~ Treatment, data = calc)
t.test(Decrease ~ Treatment, alternative="two.sided", var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  Decrease by Treatment
## t = 1.6341, df = 19, p-value = 0.1187
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.48077 12.02622
## sample estimates:
## mean in group Calcium mean in group Placebo
##             5.0000000            -0.2727273
```

p-values in last 2 problem matches perfectly, so it is the same thing with testing the given Null hypothesis.

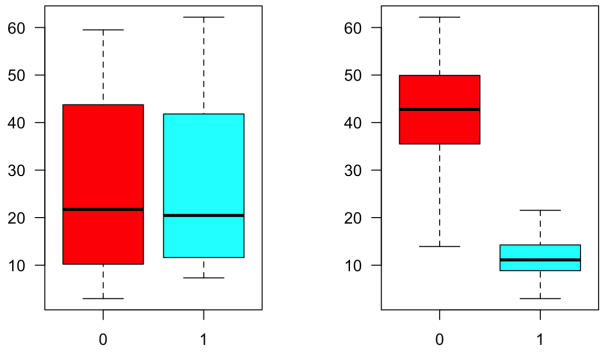
(e) Assuming that the ϵ_i are normally distributed, what is the estimated distribution of $Decrease$ when $Treatment = 1$?

$Decrease \sim N(\hat{\beta}_0 + \hat{\beta}_1, \hat{\sigma}^2) = N(5.00 - 5.273, (7.385)^2) = N(-0.273, 54.53822)$

Problem 2: With the onset data in "onset data.csv," conduct the following analysis.

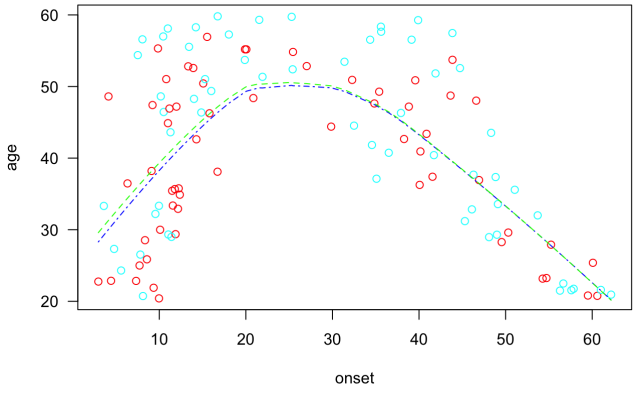
(a) Create side-by-side box plots comparing time to onset with (i) the tx variable and (ii) the prior variable. Comment.

```
setwd("~/Desktop/STAT645/Data")
ons = read.csv("onset_data.csv", header=TRUE)
par(mfrow = c(1,2))
boxplot(onset ~ tx, data = ons, col = rainbow(2), las=1)
boxplot(onset ~ prior, data = ons, col = rainbow(2), las=1)
```



(b) Create a scatterplot of onset vs. age. Color code the points by prior status. Also, fit and overlay separate lowess curves, one each for prior = 0 and prior = 1.

```
attach(ons)
plot(onset, age, col = rainbow(2), las=1)
lines(lowess(onset,age), col="blue", lty=4)
lines(lowess(onset,age+prior), col="green", lty=2)
```



(c) Fit the regression model

$$y_i = \beta_0 + \beta_1 tx_i + \beta_2 prior_i + \beta_3 age_i + \beta_4 (prior \times age)_i + \epsilon_i$$

Interpret all coefficients and report their estimates and standard errors.

```
model2 = lm(onset ~ tx + prior + age + I(prior*age))
summary(model2)
```

```
##
## Call:
## lm(formula = onset ~ tx + prior + age + I(prior * age))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.4079  -2.6405   0.4422   2.5607  12.0187
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     70.91644    2.29513  30.899   <2e-16 ***
## tx               2.11154    0.91918   2.297   0.0234 *
## prior          -69.24092    3.21787 -21.518   <2e-16 ***
## age             -0.71677    0.05334 -13.438   <2e-16 ***
## I(prior * age)   0.92996    0.07541  12.332   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.033 on 115 degrees of freedom
## Multiple R-squared:  0.9226, Adjusted R-squared:  0.9199
## F-statistic: 342.6 on 4 and 115 DF,  p-value: < 2.2e-16
```

All coefficients appear to be significant with $prior$ being the most effective one with negative (inversely proportional) effect on the response variable, and age being the least effective on $onset$ times. However, interaction between $prior$ and age variables are more effective than age variable itself.

(d) Use matrix manipulation using a design matrix to verify the estimates and standard errors from above.

```
X = model.matrix(onset ~ tx + prior + age + I(prior*age), data = ons)
Y = matrix(onset, byrow=TRUE)
B = solve(t(X)%*%X,t(X)%*%Y)

sigma2.hat.1= sum((Y-X%*%B)^2)/nrow(X)
sigma2.hat.2= sum((Y-X%*%B)^2)/(nrow(X)-ncol(X))
mysel =sqrt(sigma2.hat.1) *sqrt(diag(solve(t(X)%*%X)))  #not sure how to find
mysel
```

```
##    (Intercept)             tx          prior            age I(prior * age)
##     2.24680769     0.89982565     3.15011556     0.05221575     0.07382253
```

```
mysel2 =sqrt(sigma2.hat.2) *sqrt(diag(solve(t(X)%*%X)))  #not sure how to find
mysel2
```

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

(e) What is a 95% confidence interval for the mean difference in onset times between the treatment and control groups, holding prior status and age constant?

It is simply confidence interval for the coefficient of tx_i variable, β_1, since it is the mean difference betweentreatment and control groups:

$$\mu_{treatment} - \mu_{control} = E[Y_i \mid tx_i = 1, age_i, prior_i] - E[Y_i \mid tx_i = 0, age_i, prior_i] = \beta_1$$

.

```
model2 = lm(onset ~ tx + prior + age + I(prior*age))
confint(model2, level=0.95)[2, 1:2]
```

```
##    2.5 %   97.5 %
## 0.290824 3.932257
```

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

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

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

Problem 3: Suppose that $y_1, y_2, \ldots, y_n$ are i.i.d. realizations from the $N(0, \sigma^2)$ distribution. Derive the maximum likelihood estimator of σ^2.

Probability distribution function for each y_i is given by $f(y_i \mid 0, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} e^{-\frac{(y_i - 0)^2}{2\sigma^2}}$. Since y_i's are i.i.d, the likelihood function is:

$$\mathcal{L}(\sigma^2) = f(y_1, y_2, \cdots, y_n \mid \sigma^2) = \prod_{i=1}^{n} f(y_i \mid \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\left(\frac{\sum_{i=1}^{n} y_i^2}{2\sigma^2}\right)}$$

Then, taking log of $\mathcal{L}(\sigma)^2$, we get the log-likelihood function;

$$log(\mathcal{L}(\sigma^2)) = -\frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} {y_i}^2$$

Taking derivative of log-likelihood function with respect to the parameter σ and setting it equal to zero yields the maximum likelihood estimator since log is monotonic function and its maximum is the same with the likelihood function.

$$0 = \frac{\partial}{\partial\sigma} log(\mathcal{L}(\sigma^2)) = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} {y_i}^2$$

which yields the maximum likelihood estimator as $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} {y_i}^2$.

Problem 4: Suppose the times to infection following exposure to a particular bacteria follow the gamma distribution with shape parameter α, scale parameter β, and pdf

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta}$$

Use the *nlm* function in R to compute the maximum likelihood estimates for the data in "gamma.csv."

Log-likelihood function for gamma distribution is given by;

$$log(\mathcal{L}(\alpha,\beta)) = log(f(x_1,x_2,\cdots,x_n \mid \alpha,\beta)) = -nlog(\Gamma(\alpha)\beta^{\alpha}) + (\alpha-1)\sum_{i=1}^{n} log(x_i) - \frac{1}{\beta}\sum_{i=1}^{n} x_i$$

Maximizing log-likelihood function is the same with minimizing -[log-likelihood] function. Therefore,

```
setwd("~/Desktop/STAT645/Data")
gdata = read.csv("gamma.csv", header=TRUE)

# Gamma minus log likelihood = gml, alpha=a, beta=b, lgamma(x)=log(gamma(x))

gml <- function(theta,dat)
 {
 a = theta[1]; b = theta[2]; n = length(dat); sumx = sum(dat); sumlogx = sum(log(dat));
 gml = n*a*log(b) + n*lgamma(a) + sumx/b - (a-1)*sumlogx
 return(gml)
}

# End function gml

mle = nlm(gml,c(1,1),dat=gdata)
mle
```

```
## $minimum
## [1] -703828.2
##
## $estimate
## [1] 7173.618 1814.771
##
## $gradient
## [1] -97.128853   3.952784
##
## $code
## [1] 5
##
## $iterations
## [1] 7
```