

## STAT 457 , Fall 2010

### Lab # 6

#### More Basic Linear Regression Practice

##### Topics Covered:

- Correction from last week's lab
- Linear Regression from "start to end", of data in which a linear model is appropriate
- Linear Regression of a model in which we have to do a log transform in order to get a linear fit

=====

Last week we had script `firstset.R`, as reproduced below.

```
# generate descriptive stats and graphs on weight, height

# get basic histogram of weight
hist(vecheight, main="UM Football WEIGHT Distribution",
     xlab="pounds", ylab="frequency")

# generate relevant summary statistics
summary(vecweight)
cat("std dev=",sd(vecweight),"\n")

# get weight as relative frequency on histogram
wt <- vecweight/length(vecweight)
hist(wt, main="UM Football WEIGHT Distribution",
     xlab="pounds", ylab="rel frequency")
lines(density(wt)) # this gets a profile line on histogram

# generate another displays of WEIGHT
cat("TEAM WEIGHTS", stem(vecweight), "\n")

# generate another display
boxplot(vecweight, main="Team WEIGHT", ylab="inches")

# generate another
f1 <- fivenum(wt)
cat("five number summary =", f1, "\n")

# compare WEIGHT and HEIGHT
# first produce a scatterplot
plot(vecheight, vecweight, main="Scatterplot of WEIGHT vs HEIGHT")

# now do a linear regression model with the data and store in variable
line1
# also print out r
line1 <- lm(vecweight ~ vecheight)
line1
cat("r=", cor(vecweight,vecheight), "\n")
```

```

# plot line of best fit on scatterplot
plot(vecheight, vecweight, main="Scatterplot of WEIGHT vs HEIGHT",
      xlab="inches", ylab="pounds")
abline(line1)

# do scatterplot labeled by position, using first letter
plot(vecweight, vecheight, pch=as.character(vecposition),
main="scatterplot of
      WEIGHT on HEIGHT, Coded with letters")

# do scatterplot labeled by position, using symbol
plot(vecweight, vecheight, pch=as.numeric(vecposition),
main="Scatterplot of
      WEIGHT on HEIGHT, Coded with Symbols", xlab=")

# do residual plot
res1 <- resid(line1)
plot(vecheight, res1, main="Residual Plot-WEIGHT vs HEIGHT",
      xlab="pounds", ylab="residuals")
abline(0,0)      # plot a line with intercept=0 followed by slope=0

```

I call your attention to the purple colored lines involving the histogram of the weight distribution shown in the script above. It needs to be changed. We need to get rid of the `wt` assignment line and forget about getting a density shape line. I can't get the shape feature to work for me, so for now let's just get the shape by looking at the bars.

```

First, delete the line wt <- vecweight/length(vecweight)
which is the line above the hist(wt , etc. line).
Also, delete lines(density(wt)) # this gets a profile line on
      histogram
      which is the line below the hist(wt , etc. line).
Then change the following lines removing the original words and replacing them
by the purple words
hist(vecweight, main="UM Football WEIGHT Distribution",
      xlab="pounds", ylab="frequency")

```

The histogram of weight given last week had nonsense numbers for pounds, and the density line was a little low in profile. I missed seeing the problem, but a student caught it for me. So, by making those changes above and rerunning `firstset.R`, you should end up with a WEIGHT histogram which looks like

=====

I want to look at the relationship between two quantitative sets of data, CHILD vs ADULT, which is a listing of professional baseball ticket prices for children and adults at games in various ball parks. There is also other price data from the ball park, including prices for parking and for soda pop (called PARKING and SODA, respectively) in this data set, called `fandata.txt`. I wrote a script called `readfandata.R`, which is listed below and which you can download from the

web site. We do a linear regression on the variables CHILD (which is the response variable, or y axis variable) and ADULT, and do a scatterplot with line of fit placed on the plot, as well as produce a residual plot with x axis on the plot.

```
# open fan price data
# assuming you have already made Desktop the root directory for R
gameprice <- read.table("fandata.txt")
gameprice
colnames(gameprice) <- c("team", "Adult", "Child", "Parking", "Soda")
gameprice
child <- gameprice[,3]
adult <- gameprice[,2]
child
adult

# create a linear model using the lm() command
# then run a scatterplot with model
# line on the plot
# make child the y or response variable
# and call the model pricemodell
pricemodell <- lm(child~adult)
plot(adult,child)
abline(pricemodell)

# save residuals and make residual plot
# plot x axis on plot
residual1 <- resid(pricemodell)
plot(adult,residual1)
abline(0,0)

# make summary of model
summary(pricemodell)
```

Run this script in R, A LINE AT A TIME, rather than all at once, so you can see what was done and how it was accomplished. Ask me questions during lab if you are confused about any of it.

**1** Copy and paste relevant graphs and output information from your efforts to the WORD document for this week. Comment on the appropriateness of a linear fit for this data and the quality of fit we got with the linear model.

**2** Now, run a linear regression model on the relationship, if any, between the other 2 variables in this data set, namely, between PARKING and SODA. Make PARKING the response variable (y axis variable). I included a partially written script (where I just provided the comments, without the code!!), called yourturn.R

```
# Now it is your turn
# make a model of parking prices vs soda prices
# first call the 4th column a vector park
# and last column vector sodacost
```

```
# now run the linear model with the lm() command
# and call the result model2
```

```
# now find summary statistics on model2
# and plot the scatterplot
# with line of best fit on it
```

```
# now find the residuals and plot them
# on the y axis, with sodacost on the
# x axis for the residual plot
# and be sure to include x axis line
```

You will have to fill in the code required to do the tasks for a linear model analysis on this part. Copy and paste relevant output and graphs/tables to your WORD document.

=====

Now, download the data file `TVlife06.txt`, which contains various life expectancies of different countries (in years) and the number of TV sets per 1000 people in the countries. Column 2 is the life expectancy, and last column is the number of TV's per 1000 people in each country.

We want to do a linear regression analysis to see if there is any relationship between the number of TV's (as explanatory variable, on the x axis) and life expectancy as response variable. We will also do linear regressions of the following transformed variables:

log(life) vs TV

log(life) vs log(TV)

life vs log(TV)

Note that log is the natural log. So, for example, in R, `logTV <- log(TV)`.

Download the script `logscripts.R` and follow the instructions to do the code for accomplishing these 4 regression studies. This script is shown below.

```
# RELATING THE NUMBER OF TV'S TO LIFE EXPECTANCY
# =====
# FIRST, make sure Desktop is your default directory

# NEXT, read in datafile TVlife06.txt using read.table() command
# and call the table tvlife

# NEXT, give column names to replace V1, V2, and V3 by using colnames()

# NEXT, make variable tvlife[,2] be life
# and variable tvlife[,3] be TV
# and make variable called logTV be log(TV)
# and variable called loglife be log(life)
# remember that log() is the natural log
```

```

# NEXT, run linear model of life on TV
# and call it model10
# also make summary of model10
# and scatterplot with best fit line on it
# and make residual plot with x axis on it

# NEXT, run linear model of loglife on TV
# (which is loglife ~ TV)
# and call it model20
# get scatterplot and summary stats on model20
# which has line of best fit on it.
# Also get residual plot of model2 with x axis on it

# NEXT, run linear model of loglife on logTV
# and call it model30
# Include summary and scatter plot,
# with line on plot
# and residual plot with 0 line on plot

# NEXT, run linear model of life on logTV
# include scatterplot with line on it
# and summary and residual plot with x axis on it.
# Call the model model40

```

I will keep a copy of the code I used to do this log problem at lab, for your inspection, if and as needed, as well as I will also post my script on our web page after Tuesday. I encourage you to generate your own version of code, however. Consult my script if you find yourself running out of time.

**3** Copy/paste relevant results and graphs/tables of this log study into WORD. Comment on which of the 4 regressions:

life vs TV  
 loglife vs TV  
 loglife vs logTV  
 life vs log TV

Seems to be the best fit, or if linear regression provides an adequate fit at all, for any of them.

=====  
**EXTRA CREDIT:**

Put “nice” titles and x and y axes labels on the linear regressions. Include any other graphs (histograms, stemplots, etc.) or statistical summaries you think are useful in your regression studies, for any or all of the problems above. Make sure that these other graphs have labels, appropriately.