Lab 1 Solutions

January 22, 2017

Central Limit Theorem

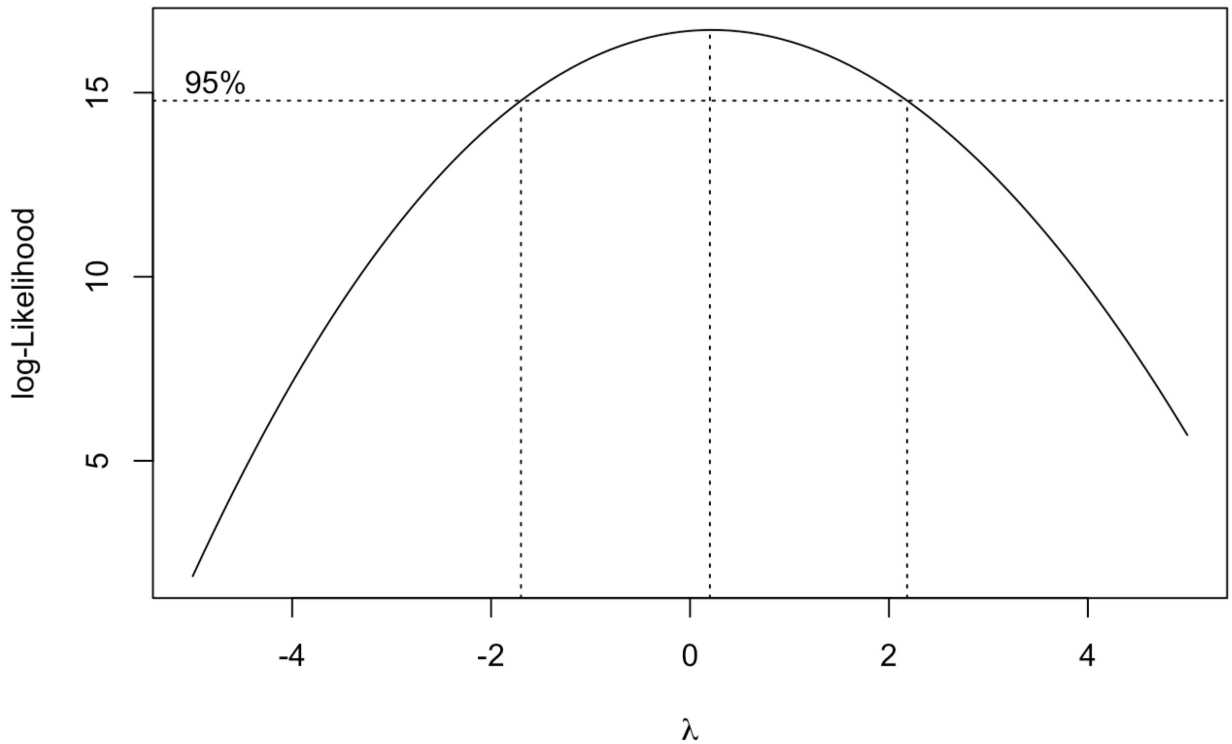
The *central limit theorem* is an **extremely** useful result in statistical analysis. The result says that the sample mean of an independent sample will be approximately normally distributed. In mathematical notation this can be written as $\bar{X}_n \sim N(\mu, \sigma^2/n)$.

Testing R Code

In the following code chunk we create some normal data, and then look at the parameter estimated for the box cox transformation.

```
library(MASS)

#create data
data = rnorm(100, mean = 10, sd = 1)
#do box-cox transformation
boxcox(data~1, lambda = seq(-5, 5, 1/10))
```



The plot we get shows a wide range of potential transformation values. If we do our analysis assuming this parameter is unknown, we could potentially lose a lot of power in our analysis due to this uncertainty. In practice and when possible, intuitable transformations can provide a much more powerful analysis.

Linear Regression Example

We will look at data that compares the chirps per second for the striped ground cricket, to the temperature in degrees farhenhit. Analyzing this relationship can give us an idea of what would be usual within our population. If we collect crickets from a separate area, quantifying the difference in the relationship of temperature and chriping time could help give us a deeper understanding of the differences in the areas.

Below we load the data and use the `head` function to preview it.

```
#Loading in the data
library(readxl)
slr02 <- read_excel("~/Dropbox/BTRY6020/2017/lab1/slr02.xlsx")
head(slr02)
```

```
##      X    Y
## 1 18.0 88.6
## 2 16.0 71.6
## 3 17.8 93.3
## 4 16.4 84.3
## 5 17.1 80.6
## 6 15.5 75.2
```

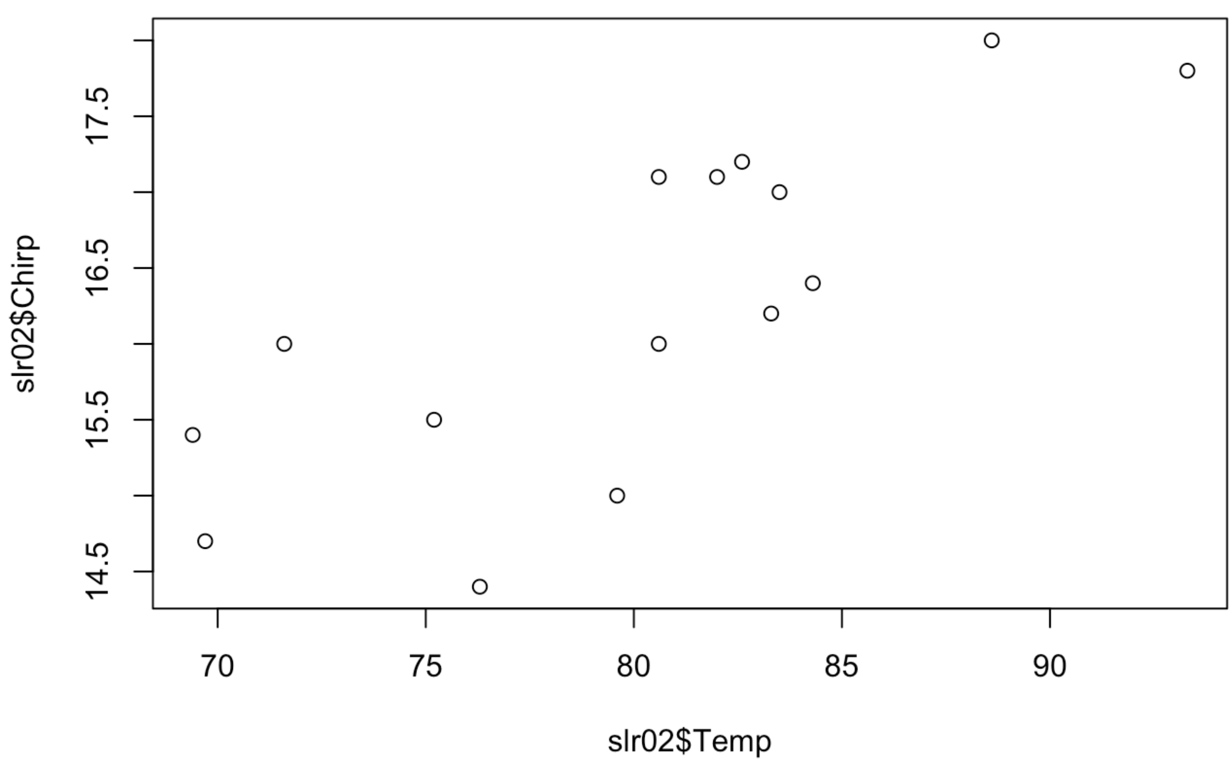
(Extra) We rename the first and second variable so the future analysis will be more clear.

```
names(slr02) = c("Chirp", "Temp")
```

Now that the data has ben loaded and we can begin to analyze it.

Let's plot the dataset.

```
plot(slr02$Temp, slr02$Chirp)
```



A linear regression seems appropriate for capturing the relationship between these variables. Since `Temp` is a variable we can control, we'll use it as the dependent variable. We can run and save the linear regression using the `lm` function.

```
cric.lm = lm(Chirp~Temp, data = slr02)
```

The `cric.lm` object is now loaded in our environment. We can see what information it has by using the `names` function.

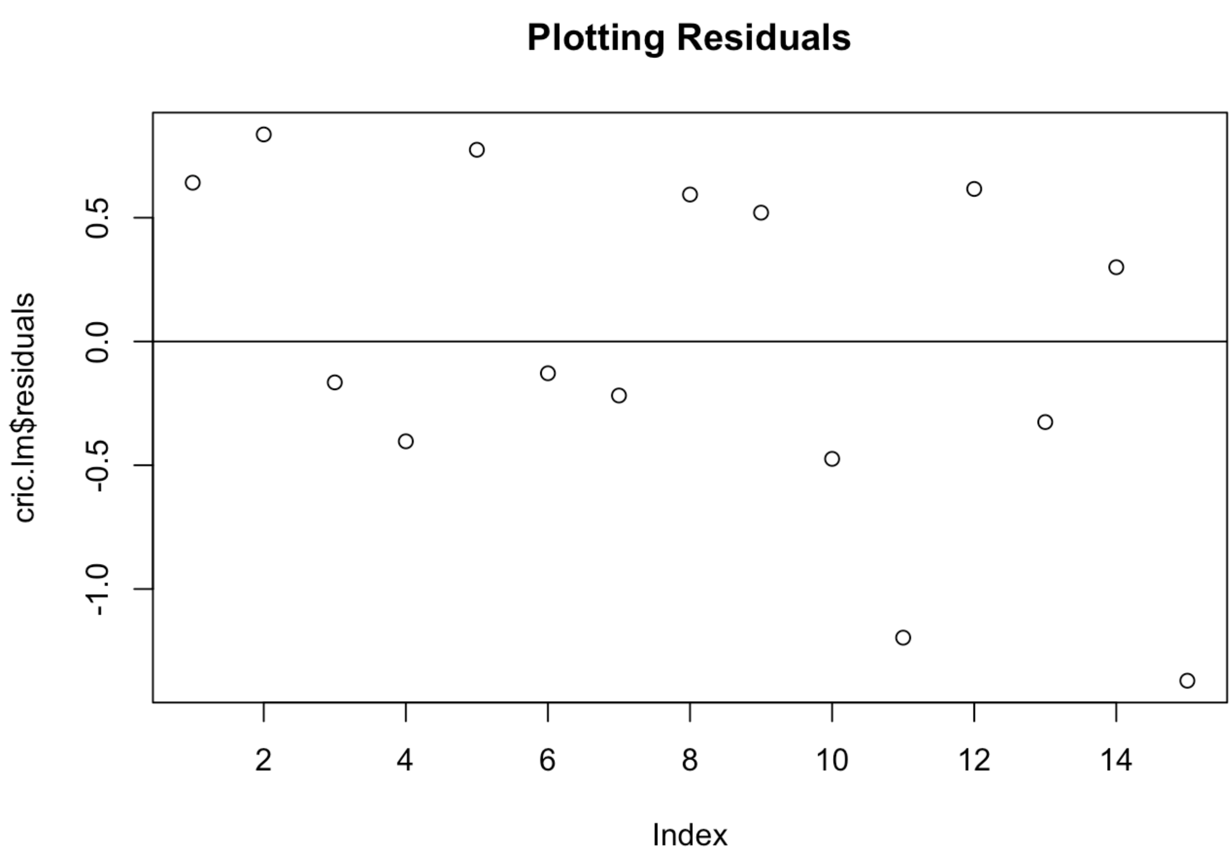
```
names(cric.lm)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

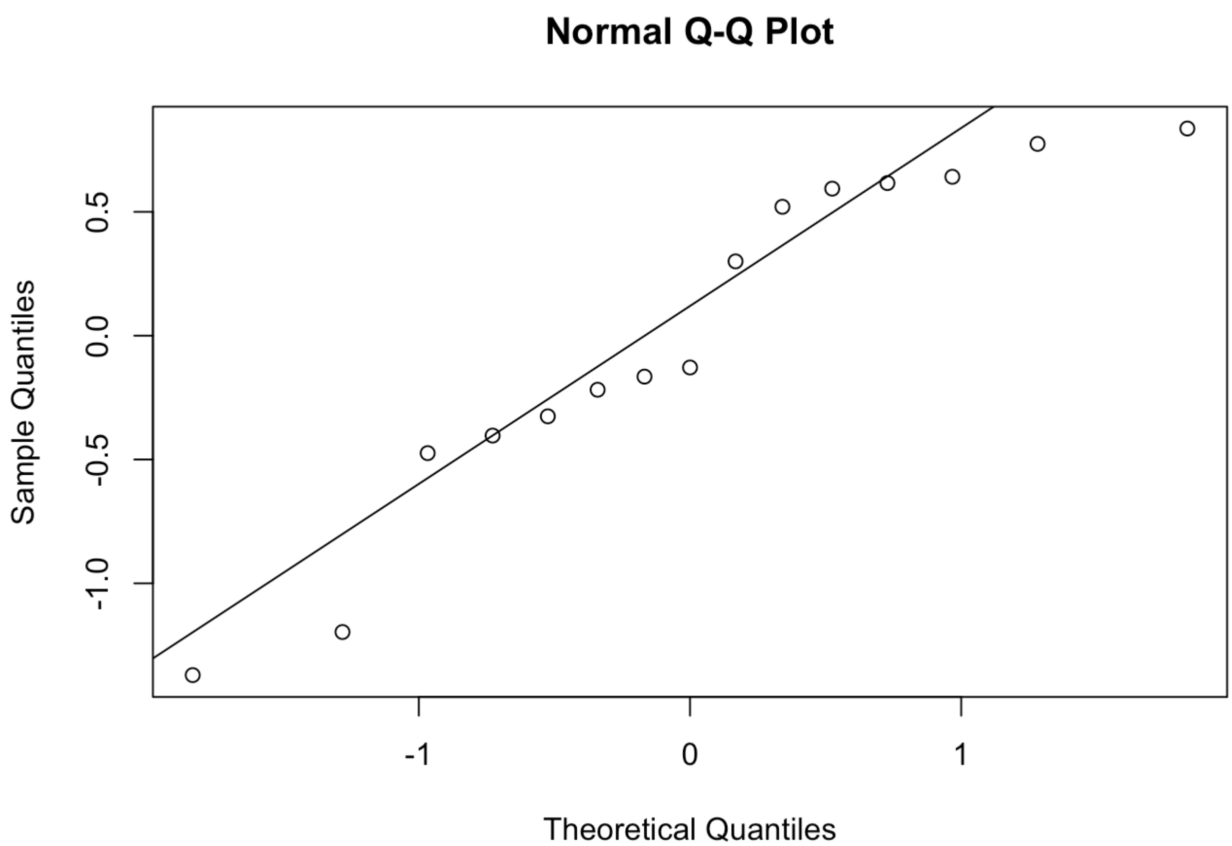
For example, since `residuals` show up, that means I can run the line `cric.lm$residuals` and obtain the fitted residuals for my linear regression.

In the next few lines of code we analyze the residuals for heteroskedasticity and normality.

```
#plotting residuals
plot(cric.lm$residuals, main = "Plotting Residuals")
#add a line through 0
abline(h = 0)
```

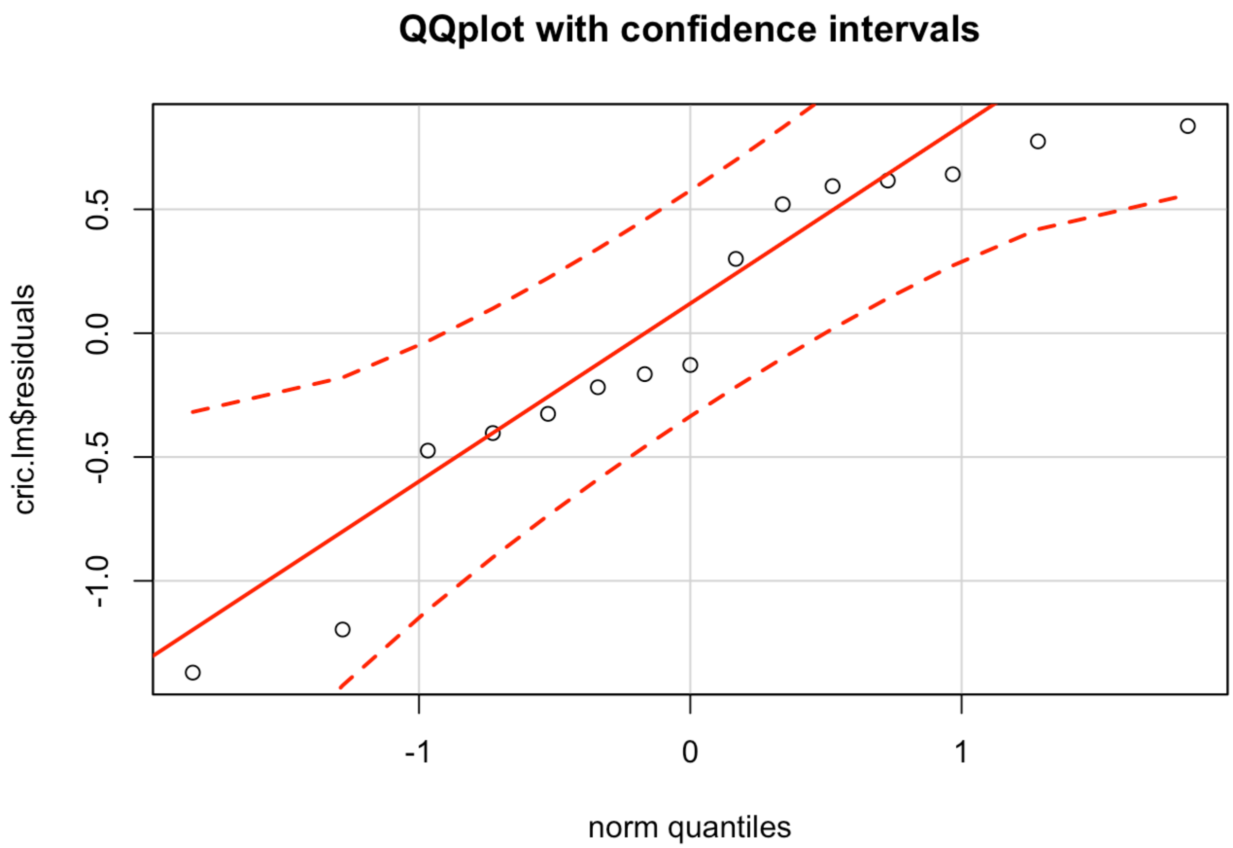


```
#QQ plot, add line
qqnorm(cric.lm$residuals)
qqline(cric.lm$residuals)
```



Here we see that heterskedasticity appears to be satisfied. We could potentially be worried about the QQplot. To be certain, we can use the `car` package to include confidence intervals on our qqplot.

```
##QQ plot with confidence intervals
library(car)
qqPlot(cric.lm$residuals, main = "QQplot with confidence intervals")
```



Here we that points fall within the confidence bands. From here we should be satisfied with the robustness of the results of our analysis.

Conclusions

Our analysis yields the following fit.

```
summary(cric.lm)
```

```
##
## Call:
## lm(formula = Chirp ~ Temp, data = slr02)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3705 -0.3645 -0.1284  0.6049  0.8364
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.9194     2.2968   2.577 0.022977 *
## Temp          0.1291     0.0286   4.514 0.000582 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7178 on 13 degrees of freedom
## Multiple R-squared:  0.6105, Adjusted R-squared:  0.5805
## F-statistic: 20.38 on 1 and 13 DF,  p-value: 0.0005822
```

Therefore, we predict that for every 1 degree increase in temperature the chirps per second will increase by 0.1291 (within the range of temperatures we recorded).

This captures approximately 58% of the variability in our dataset. Therefore this is potentially room for other variables to account for phenotypic differences in crickets.