Statistical Inference

Nguyen Toan

October 26th, 2014

Part 2: Basic inferential data analysis

1. Load the ToothGrowth data and perform some basic exploratory data analysis

```
# load the datas
library(datasets)
data(ToothGrowth)

# some basic exploratory data analyses
head(ToothGrowth)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

By using `?ToothGrowth`, we can get the explanations on the data.

```
A data frame with 60 observations on 3 variables.
  [,1]   len     numeric     Tooth length
  [,2]   supp    factor  Supplement type (VC or OJ).
  [,3]   dose    numeric     Dose in milligrams.
```

2. Provide a basic summary of the data

```
summary(ToothGrowth)
```

```
##       len        supp         dose      
##  Min.   : 4.2   OJ:30   Min.   :0.50  
##  1st Qu.:13.1   VC:30   1st Qu.:0.50  
##  Median :19.2           Median :1.00  
##  Mean   :18.8           Mean   :1.17  
##  3rd Qu.:25.3           3rd Qu.:2.00  
##  Max.   :33.9           Max.   :2.00  
```

3. Use confidence intervals and hypothesis tests to compare tooth growth by supp and dose.

We create a linear regression model with `len` explained by `dose` and `supp` and calculate the 95% confidence intervals for the coefficients.

```
fit <- lm(len ~ dose + supp, data=ToothGrowth)
confint(fit, level=0.95)
```

```
##              2.5 % 97.5 %
## (Intercept)  6.705  11.84
## dose         8.008  11.52
## suppVC      -5.890  -1.51
```

The result means that 95% of the time which we collect a different set of data and estimate parameters of the linear regression model, the coefficient estimations will vary in these confidence intervals.

```
summary(fit)
```

```
## 
## Call:
## lm(formula = len ~ dose + supp, data = ToothGrowth)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -6.600 -3.700  0.373  2.116  8.800 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    9.273      1.282    7.23  1.3e-09 ***
## dose           9.764      0.877   11.14  6.3e-16 ***
## suppVC        -3.700      1.094   -3.38   0.0013 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 4.24 on 57 degrees of freedom
## Multiple R-squared:  0.704,  Adjusted R-squared:  0.693
## F-statistic: 67.7 on 2 and 57 DF,  p-value: 8.72e-16
```

Here we consider the null hypothesis, which assumes that the coefficients in the linear regression model are zeros. From the summary of the model, we see that all *p*-values are less than 0.05, which means the null hypothesis is rejected with the 5% significance level. In other words, each variable significantly explains the variability in tooth length.

For example, the coefficient of `dose` is 9.7636, which means that increasing the dose 1 mg (while fixing `supp`) would increase the tooth length 9.7636 units.