Exploratory analysis of the ToothGrowth dataset

Daniel Fabian

June 17, 2015

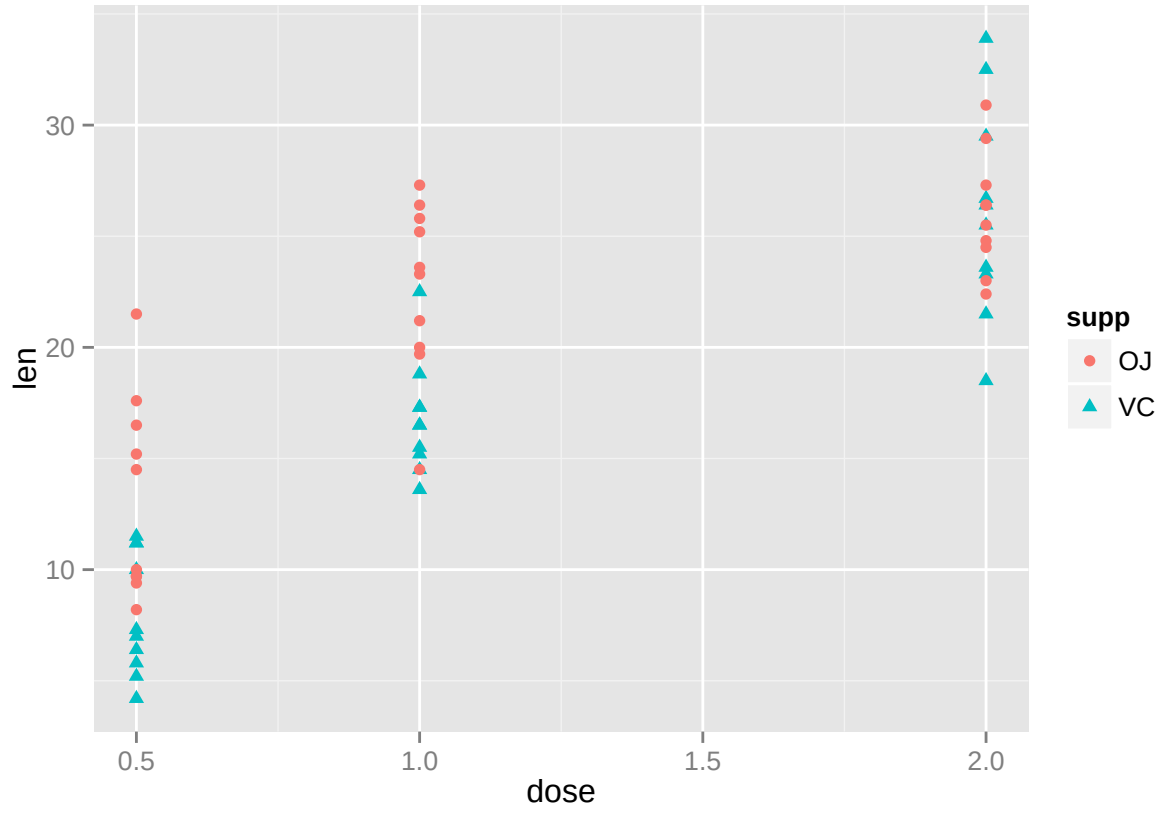
In the very beginning, we look at the summary

```
summary(ToothGrowth)
```

```
##       len        supp         dose
##  Min.   : 4.20   OJ:30   Min.   :0.500
##  1st Qu.:13.07   VC:30   1st Qu.:0.500
##  Median :19.25           Median :1.000
##  Mean   :18.81           Mean   :1.167
##  3rd Qu.:25.27           3rd Qu.:2.000
##  Max.   :33.90           Max.   :2.000
```

First, we look at the data, without caring much about the supplement and scales.

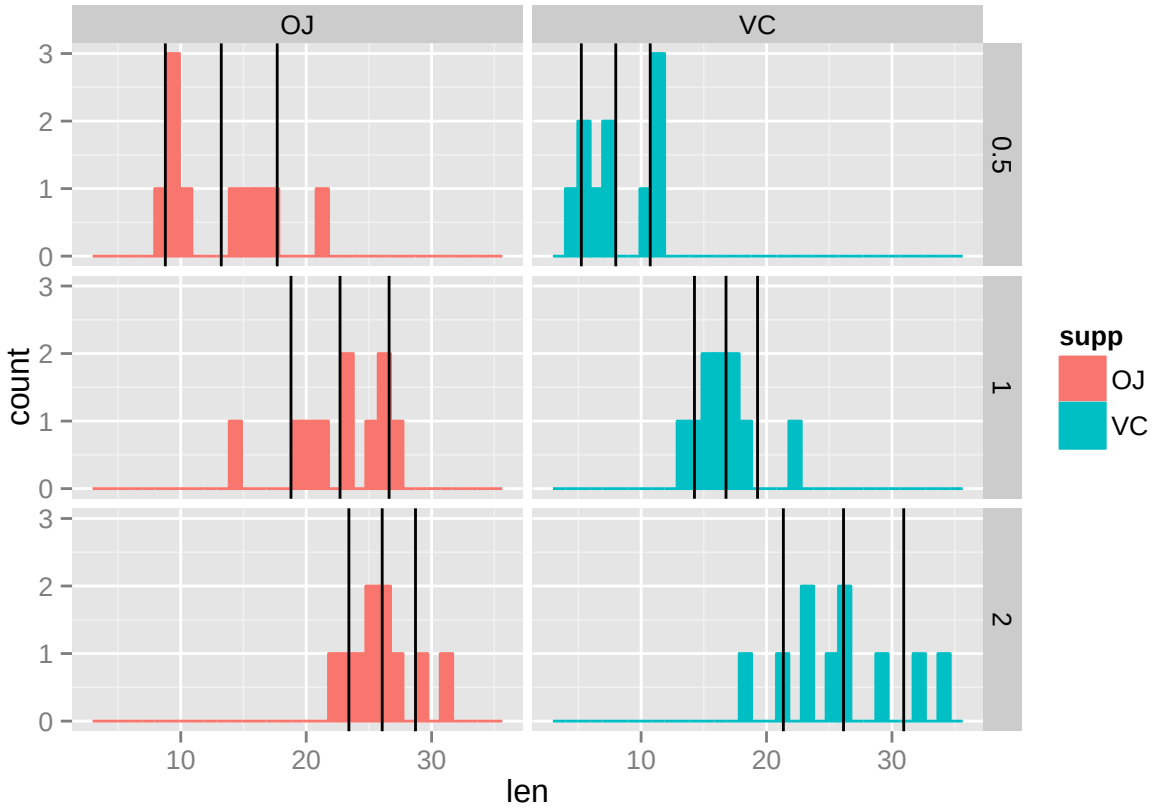
```
library(ggplot2)
g <- ggplot(ToothGrowth, aes(x = dose, y = len, colour = supp, shape = supp))
g <- g + geom_point()
g
```



it looks as though, the dose and the supplement seem to be relevant, so we look at a histogram with mean and standard deviation.

```
library(plyr)
means <- ddply(ToothGrowth, c("supp", "dose"), summarise, len.mean=mean(len), len.sd=sd(len))

g <- ggplot(ToothGrowth, aes(x = len, colour = supp, fill = supp))
g <- g + geom_histogram()
g <- g + geom_vline(data = means, aes(xintercept = len.mean))
g <- g + geom_vline(data = means, aes(xintercept = len.mean - len.sd))
g <- g + geom_vline(data = means, aes(xintercept = len.mean + len.sd))
g <- g + facet_grid(dose ~ supp)
g
```



Here, I assume that the two supplements are indepent and I treat the two supplements as two distinct cases and I split the data into two different datasets, on for `OJ` and one for `Vc`. Using the `lm`, I fit a linear model to the data and give the coefficients:

```
vcMod <- lm(len ~ dose, ToothGrowth[ToothGrowth$supp=="VC",])
summary(vcMod)$coefficients
```

```
##             Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  3.29500   1.427060  2.308943 2.854201e-02
## dose        11.71571   1.078756 10.860392 1.509369e-11
```

```
ojMod <- lm(len ~ dose, ToothGrowth[ToothGrowth$supp=="OJ",])
summary(ojMod)$coefficients
```

```
##             Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 11.550000   1.721951 6.707508 2.788784e-07
## dose         7.811429   1.301673 6.001070 1.824801e-06
```

the coefficient `dose` is the slope of the linear model's curve. When we look at the `Pr(>|t|)` value, this is the probability, that such a slope would have occured under the null hypothesis, that there is no `dose` parameter, i.e. that there is no linear trend in the data.

We can also get the confidence intervals for the coefficients.

```
confint(vcMod)
```

```
##                 2.5 %   97.5 %
## (Intercept) 0.3717998  6.21820
## dose        9.5059827 13.92545
```

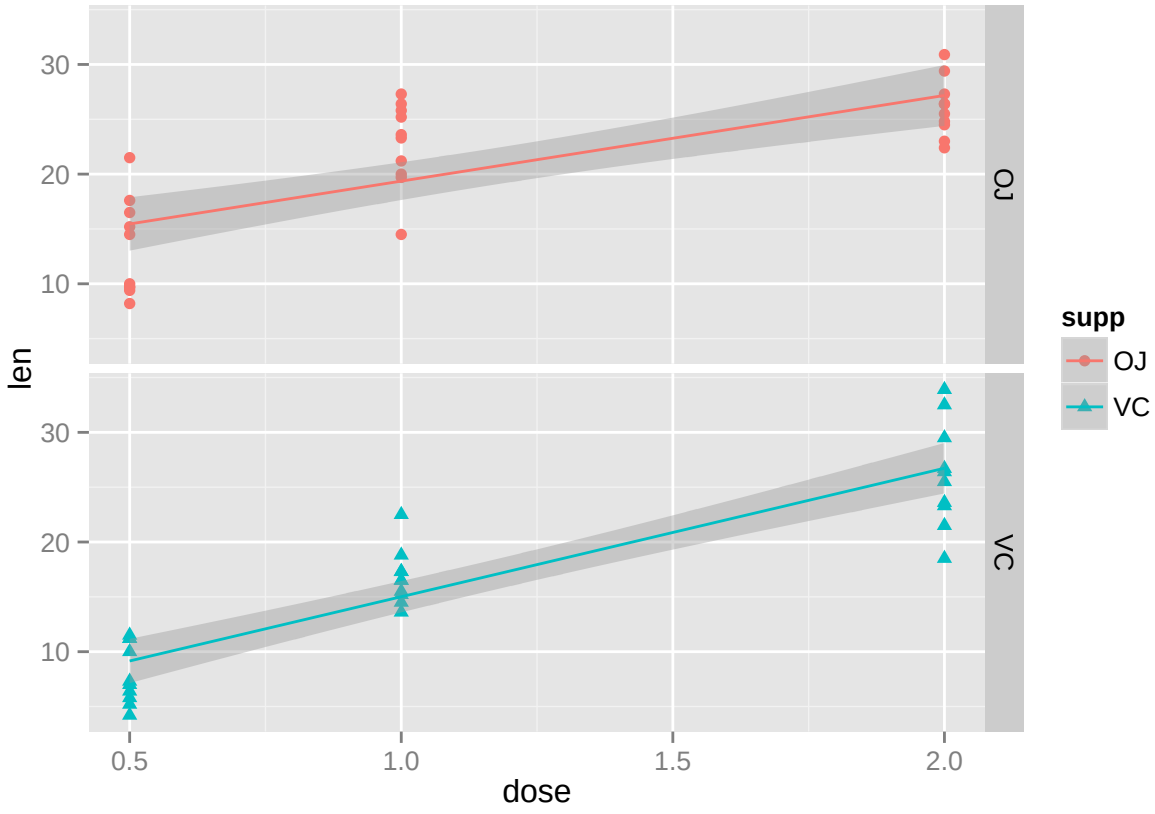
```
confint(ojMod)
```

```
##                2.5 %   97.5 %
## (Intercept) 8.022743 15.07726
## dose        5.145073 10.47778
```

So with the p-value, we got a very high confidence, that there is indeed a trend in the data, that cannot be explained with the null hypothesis. And with the confidence intervals we also give a 95% confidence interval for the slope and intercept terms.

So with the fitted model, the data looks thus:

```
g <- ggplot(ToothGrowth, aes(x = dose, y = len, colour = supp, shape = supp))
g <- g + geom_point()
g <- g + geom_smooth(aes(group = supp), method="lm")
g <- g + facet_grid(supp ~ .)
g
```



We see, that the linear model fits the data reasonably well, but maybe a further transformation would be useful. Notice, that we are given `dose` values only for 0.5, 1 and 2. So there is a always a factor of 2 between them. So maybe looking at the data under a *log* transformation could be a further idea for some exploratory analysis; but it is beyond the scope of the current project.