Voltage Case Study

David Gerard

2018-12-07

Objectives

- Analyze Voltage vs Breakdown Time Case Study
- Lack of Fit F-test

Case Study: Voltage vs Breakdown Time, a Controlled Experiment

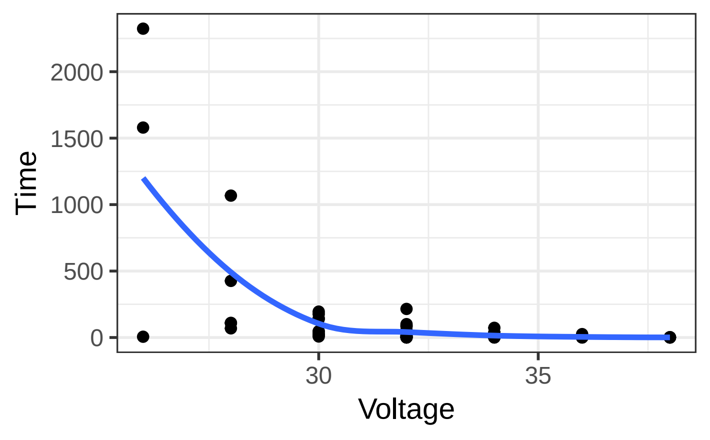
- Goal: study relationship between voltage and breakdown time of an electrical insulating fluid.
- The authors could control the voltage level of each trial.

```
library(Sleuth3)
data("case0802")
head(case0802)
```

```
##      Time Voltage  Group
## 1    5.79      26 Group1
## 2 1579.52      26 Group1
## 3 2323.70      26 Group1
## 4   68.85      28 Group2
## 5  108.29      28 Group2
## 6  110.29      28 Group2
```

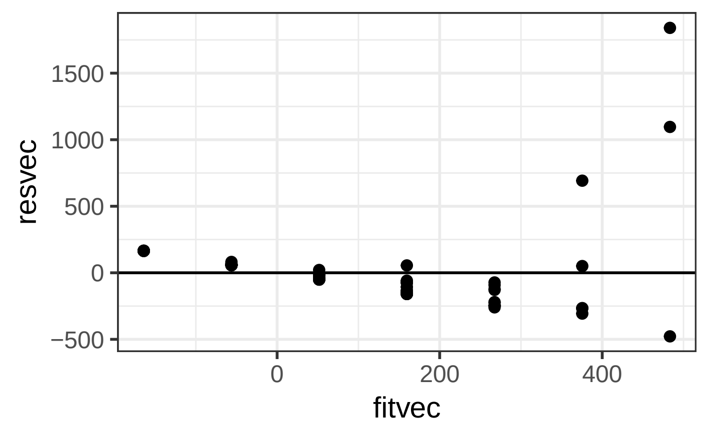
Step 1: Make a plot

```
library(ggplot2)
qplot(Voltage, Time, data = case0802) +
  geom_smooth(se = FALSE)
```



Try an initial fit with a residual plot

```
lmout <- lm(Time ~ Voltage, data = case0802)
resvec <- resid(lmout)
fitvec <- fitted(lmout)
qplot(fitvec, resvec) + geom_hline(yintercept = 0)
```

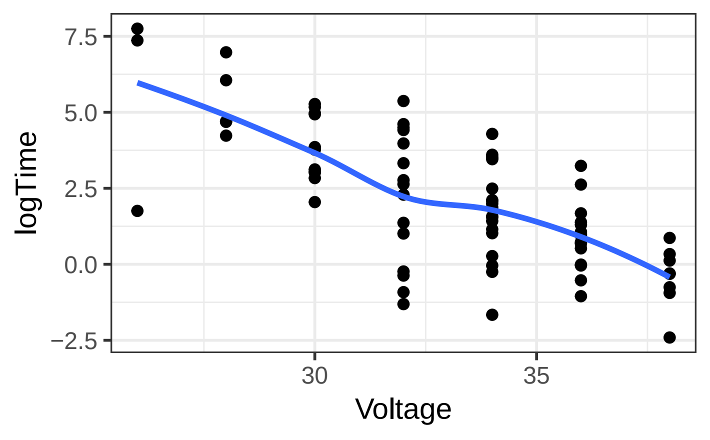


Conclusion?

- As the mean increases, the variability increases.
- We see a curved relationship between X and Y
- Clearly need a log-transformation

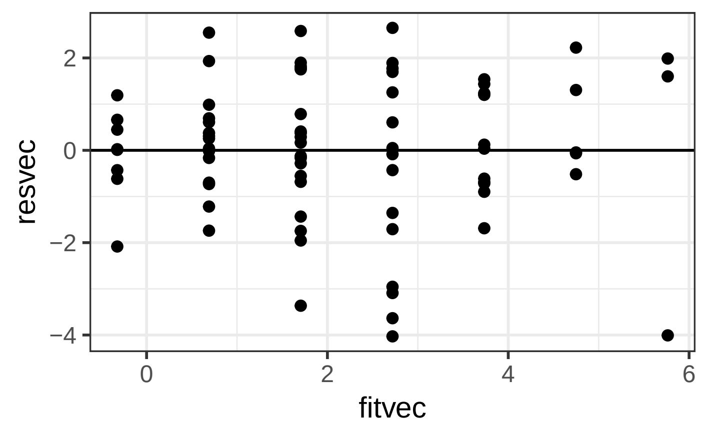
Log-transformation and Re-plot

```r
case0802$logTime <- log(case0802$Time)
qplot(Voltage, logTime, data = case0802) +
  geom_smooth(se = FALSE)
```



Log-transformation and Residual Plot

```
lmout <- lm(logTime ~ Voltage, data = case0802)
resvec <- resid(lmout)
fitvec <- fitted(lmout)
qplot(fitvec, resvec) + geom_hline(yintercept = 0)
```



Conclusion

- After the log-transformation, the data look pretty awesome.

Formal test if there is a relationship

- There is clearly a relationship here, but you need to report p-values to get published, so ...

```
sumout <- summary(lmout)
coef(sumout)
```

```
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  18.9555     1.9100   9.924 3.052e-15
## Voltage      -0.5074     0.0574  -8.840 3.340e-13
```

More interesting are coefficient estimates with confidence intervals

```
cbind(coef(lmout), confint(lmout))
```

```
##                        2.5 % 97.5 %
## (Intercept) 18.9555 15.1497 22.761
## Voltage     -0.5074 -0.6217 -0.393
```

Interpret on Original Scale

- A one kV increase results in a $\exp(-0.507) = 0.6$ multiplicative change in breakdown times.
- 95% confidence of

```r
exp(confint(lmout)[2, ])
```

```
##  2.5 % 97.5 %
##  0.537  0.675
```

- A one kV increase results in a 40% decrease in breakdown time, 95% confidence interval of

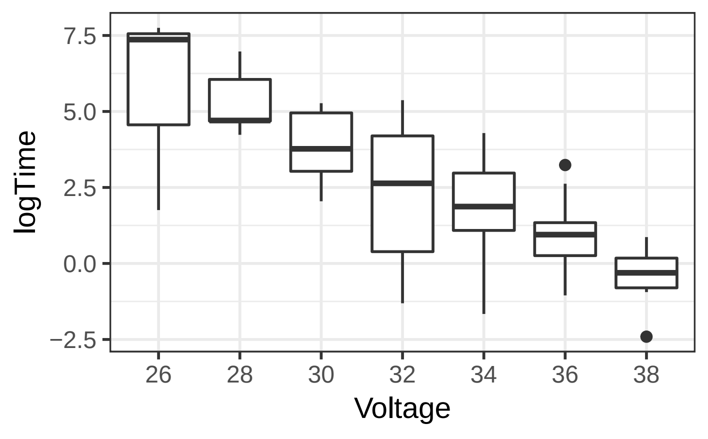
```r
(1 - exp(confint(lmout)[2, ])) * 100
```

```
##  2.5 % 97.5 %
##   46.3   32.5
```

Lack of Fit F-test

We could have viewed this as an ANOVA problem

```
case0802$VoltageFac <- as.factor(case0802$Voltage)
qplot(VoltageFac, logTime, data = case0802, geom = "boxplot
  xlab("Voltage")
```



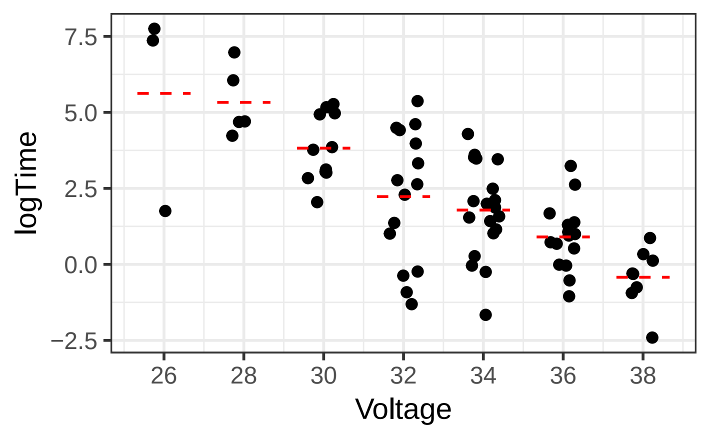
Which is better?

- If the linear model appears to fit fine, it is **always preferred**.
- You can interpolate with the linear model (not ANOVA).
- The linear model has easier interpretations.
- The linear model has fewer parameters
- We can formally test if the linear model does not fit using the F-testing strategy if we have replicates at given values of X

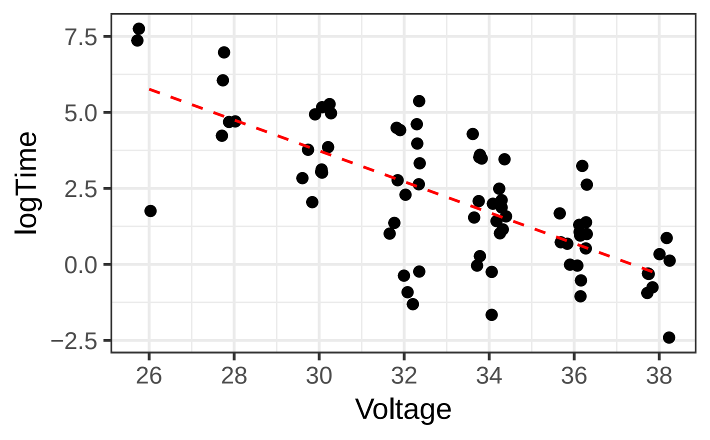
Lack-of-fit F-test

- $H_0 : E[Y_i] = \beta_0 + \beta_1 X_i$ (mean is based on line)
- $H_A : E[Y_i] = \mu_j$ where $X_i = j$ (mean is based on group)

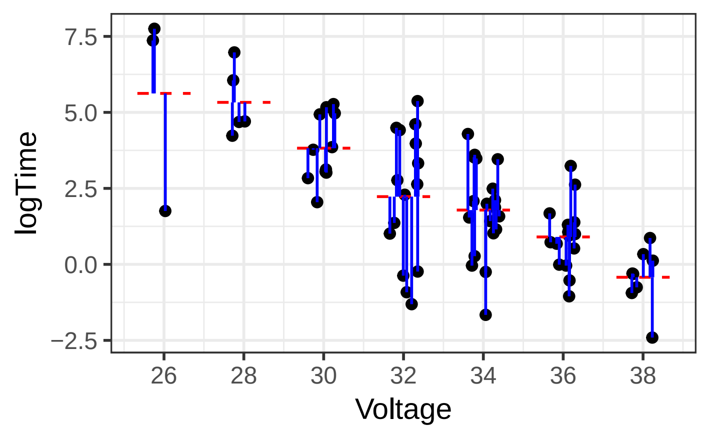
Full Model Estimates



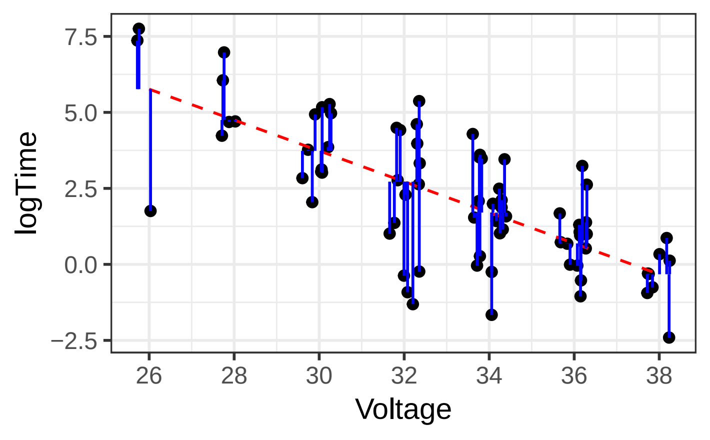
Reduced Model Estimates



Residuals Fnder Full



Residuals Fnder Reduced



Lack-of-fit F-test

- $RSS_{full} = 173.7489$
- $df_{full} = n - I =$ 76 - 7 $= 69$

Lack-of-fit F-test

- $RSS_{full} = 173.7489$
- $df_{full} = n - I =$ 76 - 7 $= 69$
- $RSS_{reduced} = 180.0748$
- $df_{reduced} = n - 2 =$ 76 - 2 $= 74$

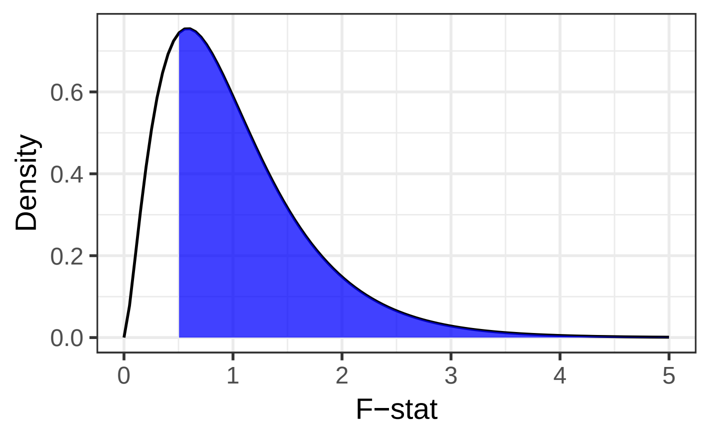
Lack-of-fit F-test

- $RSS_{full} = 173.7489$
- $df_{full} = n - I =$ 76 - 7 $= 69$
- $RSS_{reduced} = 180.0748$
- $df_{reduced} = n - 2 =$ 76 - 2 $= 74$
- $ESS = RSS_{reduced} - RSS_{full} = 6.3259$
- $df_{extra} = df_{reduced} - df_{full} =$ 74 - 69 - 5

Lack-of-fit F-test

- $RSS_{full} = 173.7489$
- $df_{full} = n - I =$ 76 - 7 $=$ 69
- $RSS_{reduced} = 180.0748$
- $df_{reduced} = n - 2 =$ 76 - 2 $=$ 74
- $ESS = RSS_{reduced} - RSS_{full} = 6.3259$
- $df_{extra} = df_{reduced} - df_{full} =$ 74 - 69 - 5
- F-statistic $= \frac{ESS/df_{extra}}{RSS_{full}/df_{full}} = 0.5024$

Compare to an $F_{5,69}$



Calculate p-value

```r
pf(0.5024, df1 = 5, df2 = 69, lower.tail = FALSE)
```

```
## [1] 0.7734
```

Lack of Fit in R

- Create a factor variable

```
case0802$VoltageFac <- as.factor(case0802$Voltage)
```

- Fit both the ANOVA and regression models

```
aout  <- aov(logTime ~ VoltageFac, data = case0802)
lmout <- lm(logTime ~ Voltage, data = case0802)
```

Lack of Fit in R

- Use `anova()` to get ANOVA table

```
anova(lmout, aout)
```

```
## Analysis of Variance Table
##
## Model 1: logTime ~ Voltage
## Model 2: logTime ~ VoltageFac
##   Res.Df RSS Df Sum of Sq   F Pr(>F)
## 1     74 180
## 2     69 174  5      6.33 0.5   0.77
```