Chapter 8.1, 8.5: ANOVA as an Assessment of Fit

In these notes we are going to compare three models that we have studied this semester:

- Equal mean model (i.e., a 1-sample approach)
- A SLR
- ANOVA (i.e., separate means model)

We are going to go over the mechanics for performing an **extra sum of squares test** to assist us in determining which model is "best" for describing the data.

Some housekeeping:

```
library(Sleuth3)
source("http://www.math.montana.edu/parker/courses/STAT411/diagANOVA.r")
```

8.1 Case study of breakdown times under different voltages

Here's the data for the case study in section 8.1.2 where 'breakdown time' of an insulating fluid was studied in an experiment under uniform conditions. Lets fit the SLR

$$\mu\{\text{Time}|\text{Voltage}\} = \beta_0 + \beta_1 \text{Voltage}$$

```
# Get data and plot it

summary(case0802)
```

```
##       Time             Voltage         Group
##  Min.   :   0.090   Min.   :26.00   Group1: 3
##  1st Qu.:   1.617   1st Qu.:31.50   Group2: 5
##  Median :   6.925   Median :34.00   Group3:11
##  Mean   :  98.558   Mean   :33.13   Group4:15
##  3rd Qu.:  38.383   3rd Qu.:36.00   Group5:19
##  Max.   :2323.700   Max.   :38.00   Group6:15
##                                     Group7: 8
```

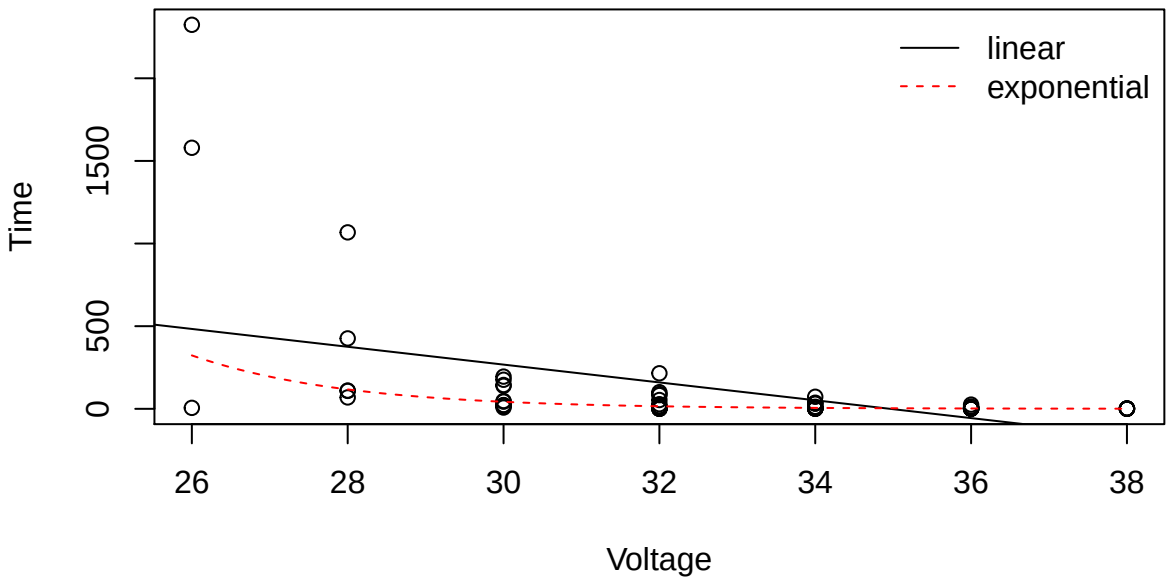
```
dim(case0802)   # This tells us that n=76
```

```
## [1] 76  3
```

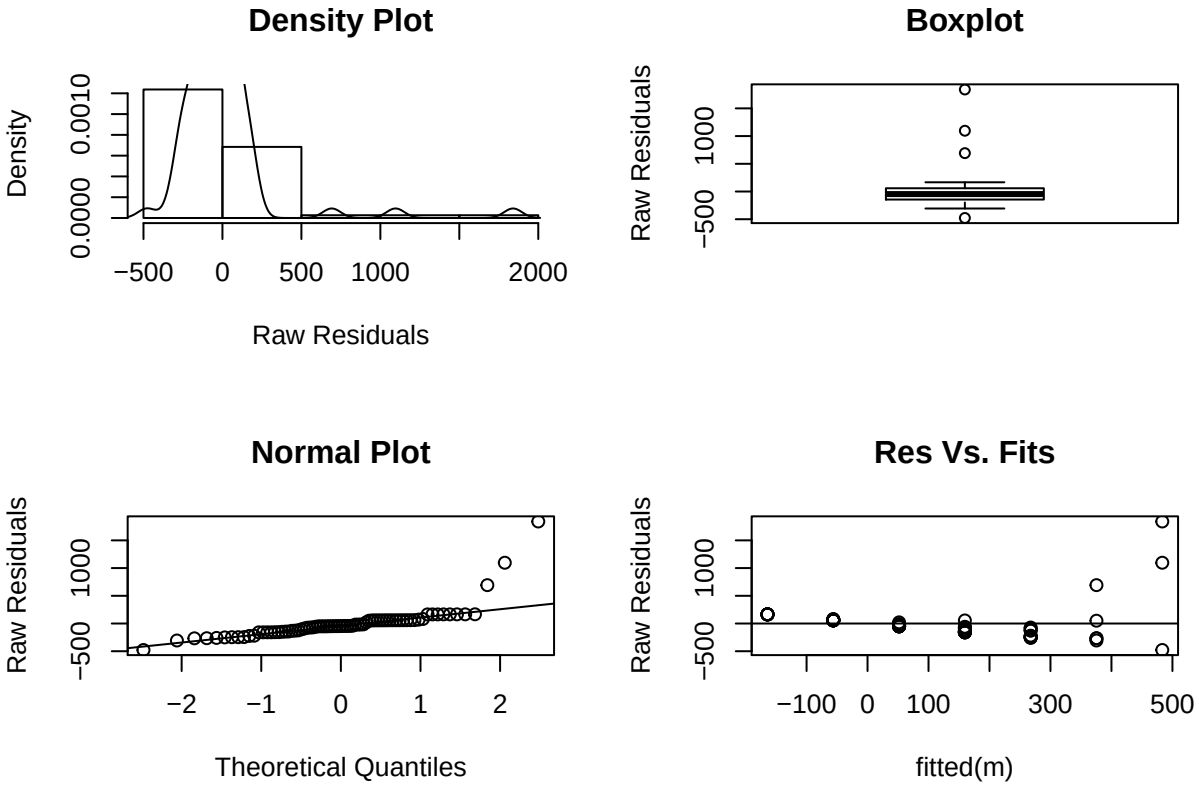
```
m.BAD=lm(Time ~ Voltage,data=case0802)

plot(Time ~ Voltage,data=case0802)
abline(coef(m.BAD))   # Show the line

volt=seq(26,38,length=100)
lines(volt,exp(18.96 -0.507*volt),col="red",lty=2)  # We'll see below where this comes from
legend("topright",legend = c("linear","exponential"),bty = "n",col = c("black","red"), lty =c(1,2))
```




```
diagANOVA(m.BAD)
```



The scatterplot and residual vs. fits plot show that Time and Voltage do not have a linear relationship. Your book suggests what to do: log-transform the response! The new SLR model is

$$\mu\{\log \text{Time}|\text{Voltage}\} = \beta_0 + \beta_1 \text{Voltage}$$

If the residuals are normal (i.e., symmetric) then

$$\mu\{\log \text{Time}|\text{Voltage}\} = \text{Median}\{\log \text{Time}|\text{Voltage}\} = \log \text{Median}\{\text{Time}|\text{Voltage}\}.$$

Exponentiating both sides gives median Time as negative exponential function of Voltage:

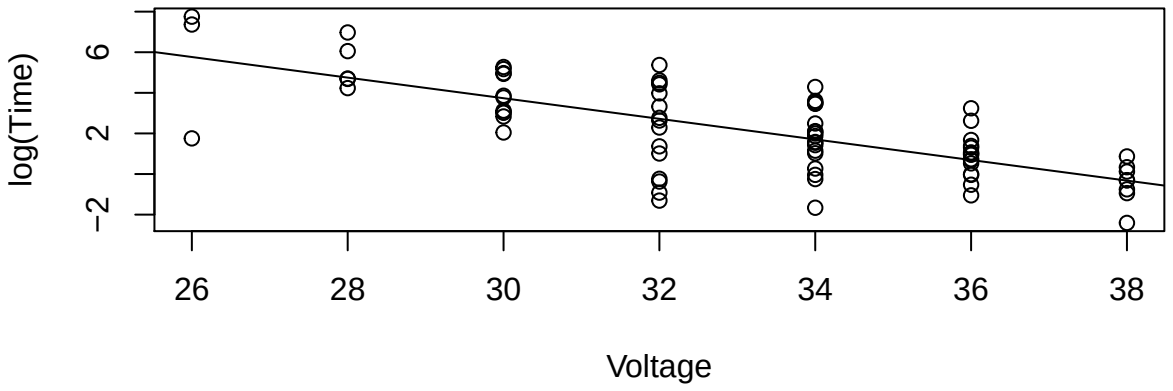
$$\text{Median}\{\text{Time}|\text{Voltage}\} = e^{\beta_0} e^{\beta_1 \text{Voltage}}$$

```
m.SLR=lm(log(Time) ~ Voltage,data=case0802)
summary(m.SLR)
```

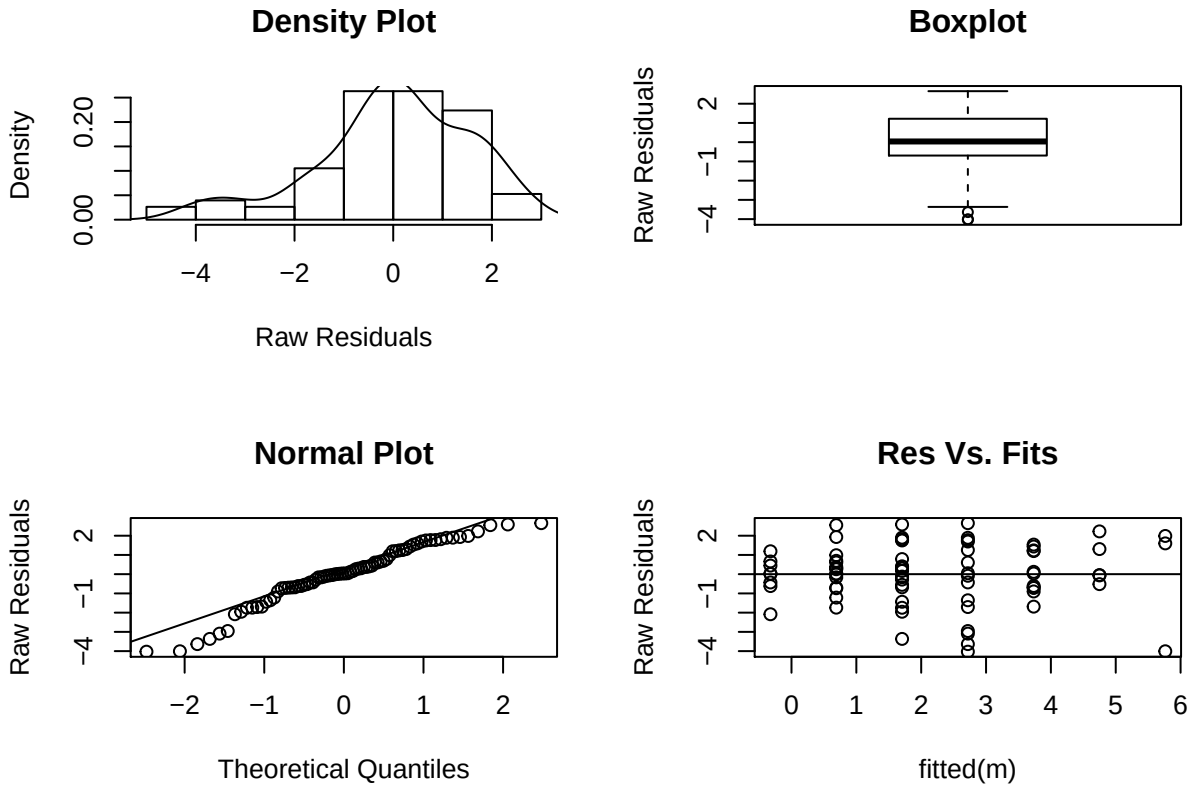
```
##
## Call:
## lm(formula = log(Time) ~ Voltage, data = case0802)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0291 -0.6919  0.0366  1.2094  2.6513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.9555     1.9100   9.924 3.05e-15 ***
## Voltage      -0.5074     0.0574  -8.840 3.34e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.56 on 74 degrees of freedom
## Multiple R-squared:  0.5136, Adjusted R-squared:  0.507
## F-statistic: 78.14 on 1 and 74 DF,  p-value: 3.34e-13
```

Let's plot these data and the residuals on the log-scale to assess the fit:

```
plot(log(Time) ~ Voltage,data=case0802)
abline(coef(m.SLR))
```



```
diagANOVA(m.SLR)
```



The scatterplot and residual vs. fits plot for an SLR of log(Time) vs Voltage indicate much better fit to the data.

We have not looked at an ANOVA table for SLR before. Compare with Display 8.8.

```
anova(m.SLR)  # Compare with Display 8.8.
```

```
## Analysis of Variance Table
##
## Response: log(Time)
##           Df Sum Sq Mean Sq F value   Pr(>F)
## Voltage    1 190.15 190.151  78.141 3.34e-13 ***
## Residuals 74 180.07   2.433
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA table allows us to use sums of squares to confirm that $R^2 = 51\%$ (cf. `summary(m.SLR)` output above). The formula $R^2 = 1 - \frac{SSE}{SST}$ was provided in the Chapter 5 notes:

```
1-180.07/(180.07 + 190.15)
```

```
## [1] 0.5136135
```

Because only 7 voltages were tested in this experiment, we can also consider a ANOVA fit to these data.

```
m.ANOVA = lm(log(Time) ~ as.factor(Voltage),data=case0802)
anova(m.ANOVA)   # Compare with Display 8.8.
```

```
## Analysis of Variance Table
##
## Response: log(Time)
##                    Df Sum Sq Mean Sq F value    Pr(>F)
## as.factor(Voltage)  6 196.48  32.746  13.004 8.871e-10 ***
## Residuals          69 173.75   2.518
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
1 - 173.75/(173.75 + 196.48)  # can also check via summary(m.ANOVA)
```

```
## [1] 0.5306971
```

The last model we will need is the "equal means" or "single mean" model. We can fit it like this:

```
m.null = lm(log(Time) ~ 1,data=case0802)
anova(m.null)
```

```
## Analysis of Variance Table
##
## Response: log(Time)
##           Df Sum Sq Mean Sq F value Pr(>F)
## Residuals 75 370.23  4.9364
```

QUESTIONS:

You might be tempted to compare the R^2 values. Go ahead, which R^2 is larger?

But R^2 always increases as the number of parameters increases. How many parameters are in each of the models `m.SLR`, `m.ANOVA` and `m.null` above? So which R^2 is expected to be larger?

8.5 Extra sum of squares test

The extra sum of squares test can help us to decide whether a "reduced model" sufficiently describes the data; or whether a more complicated "full model" better descrbes it. The test is implemented by:

1. H_0: the reduced model is the true model vs. H_a: the full model is the true model
2. Check the assumptions for the reduced and full models
3. Test statistic is
$$F_{stat} = \frac{\frac{SSE_R - SSE_F}{DFE_R - DFE_F}}{SSE_F/DFE_F}$$
where SSE_R and SSE_F are the sums of squares error for the reduced and full models respectively, and DFE_R and DFE_F are the associated degrees of freedom error. The *extra* sum of squares, $ESS = SSE_R - SSE_F$, is interpreted as the extra variability in the response that is explained by the more complicated full model; or as the drop in the unexplained variability of the response when using the full model instead of the simpler reduced model. The test statistic F_{stat} compares this drop in the SSE to the unexplained variability in the full model, $\hat{\sigma}^2_F = SSE_F/DFE_F$. This drop must be large to conclude that the additional parameters in the full model substantially improve model fit.
4. p-value is $P(F > F_{stat})$ where $F \sim F(DFE_R, DFE_F)$

5. Make a decision
6. State a conclusion - a small p-value leads you to conclude that the model under H_a is "better"

It is easy to implement the extra sum of squares test in `R` using `anova()`. Let's look at 3 examples:

EXAMPLE 1: To compare the "equal means" (reduced) model to the SLR (full) model, the hypotheses are:

H_0: the single mean model is the true model vs. H_a: SLR is the true model

Let R test these hypotheses:

```
anova(m.null,m.SLR)
```

```
## Analysis of Variance Table
##
## Model 1: log(Time) ~ 1
## Model 2: log(Time) ~ Voltage
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     75 370.23
## 2     74 180.07  1    190.15 78.141 3.34e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This is the same output as we got for `anova(m.SLR)` on p.4!

QUESTION: Sketch out the 6 steps of the hypothesis test including a conclusion. So is increasing model complexity from 1 to 2 parameters appropriate?

EXAMPLE 2: To compare the "equal means" (reduced) model to the ANOVA (full) model, the hypotheses are:

H_0: the equal mean model is the true model vs. H_a: ANOVA is the true model

Let R test these hypotheses:

```
anova(m.null,m.ANOVA)
```

```
## Analysis of Variance Table
##
## Model 1: log(Time) ~ 1
## Model 2: log(Time) ~ as.factor(Voltage)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     75 370.23
## 2     69 173.75  6    196.48 13.004 8.871e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This is the same output as we got for `anova(m.ANOVA)` on p.5! This is the same output as we got for `anova(m.SLR)` on p.4!

QUESTION: Sketch out the 6 steps of the hypothesis test including a conclusion. So is increasing model complexity from 1 to 7 parameters appropriate?

EXAMPLE 3: When using an extra sum of squares test to compare a regression (reduced) model, here SLR, to ANOVA (full) model, the test is sometimes called a **lack of fit test**. It is only possible to compare a regression to ANOVA if you have replicate values of the response at some or all of the explanatory variable values. In this case the hypotheses to test are:

$$H_0\text{: SLR is the true model vs. } H_a\text{: ANOVA is the true model}$$

and the test statistic is (see p. 220 of the Sleuth):

$$F_{stat} = \frac{\frac{SSE_{SLR} - SSE_{ANOVA}}{DFE_{SLR} - DFE_{ANOVA}}}{SSE_{ANOVA}/DFE_{ANOVA}}$$

To implement a lack of fit test in R:

```
anova(m.SLR,m.ANOVA)  # Compare with Display 8.10
```

```
## Analysis of Variance Table
##
## Model 1: log(Time) ~ Voltage
## Model 2: log(Time) ~ as.factor(Voltage)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     74 180.07
## 2     69 173.75  5    6.3259 0.5024 0.7734
```

```
# Calculate the test stat for comparing SLR to ANOVA by hand
(6.326/5)/( 173.75/69)
```

```
## [1] 0.5024391
```

```
# Calculate the p-value for comparing SLR to ANOVA by hand
1-pf(0.5024,5,69)
```

```
## [1] 0.7734214
```

The SSE and DFE for each model are from the ANOVA tables provided earlier for each model.

QUESTION: Sketch out the 6 steps of the hypothesis test including a conclusion. So is increasing model complexity from 2 to 7 parameters appropriate?