STAT-S631

Assignment 11

John Koo

```
dp <- loadNamespace('dplyr')
import::from(magrittr,
             `%>%`, `%<>%`)
library(ggplot2)
theme_set(theme_bw())
import::from(GGally,
             ggpairs)
import::from(car,
             Anova, boxCox, bcPower,
             powerTransform, invResPlot,
             invTranEstimate, invTranPlot)
import::from(effects,
             effect, Effect)
import::from(miscTools,
             rSquared)
```

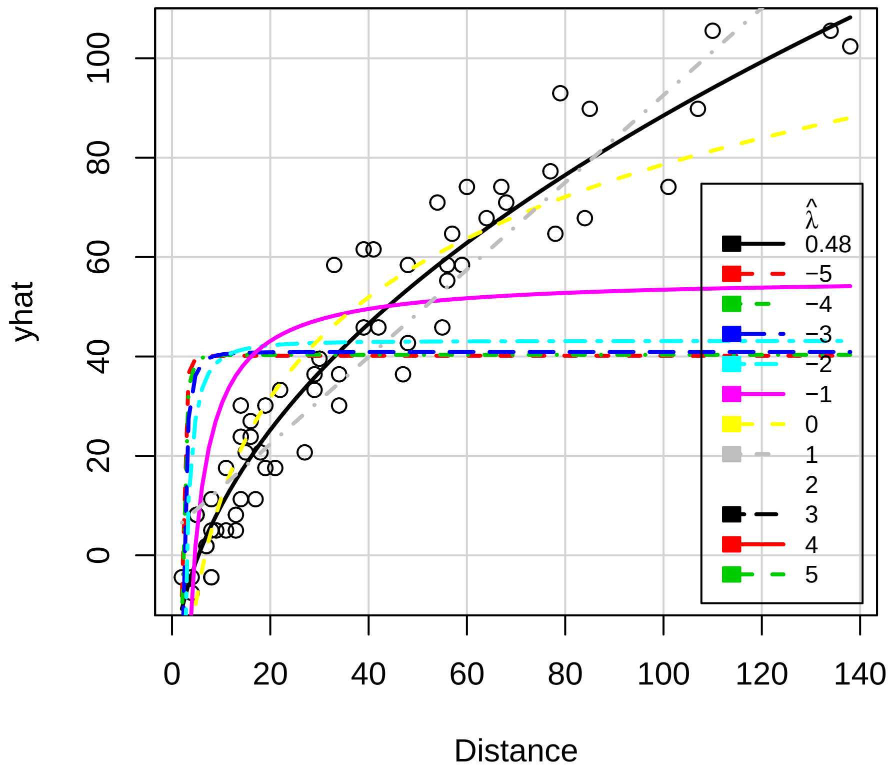
Problem 1

[From ALR 8.2]

```
stopping.df <- alr4::stopping
```

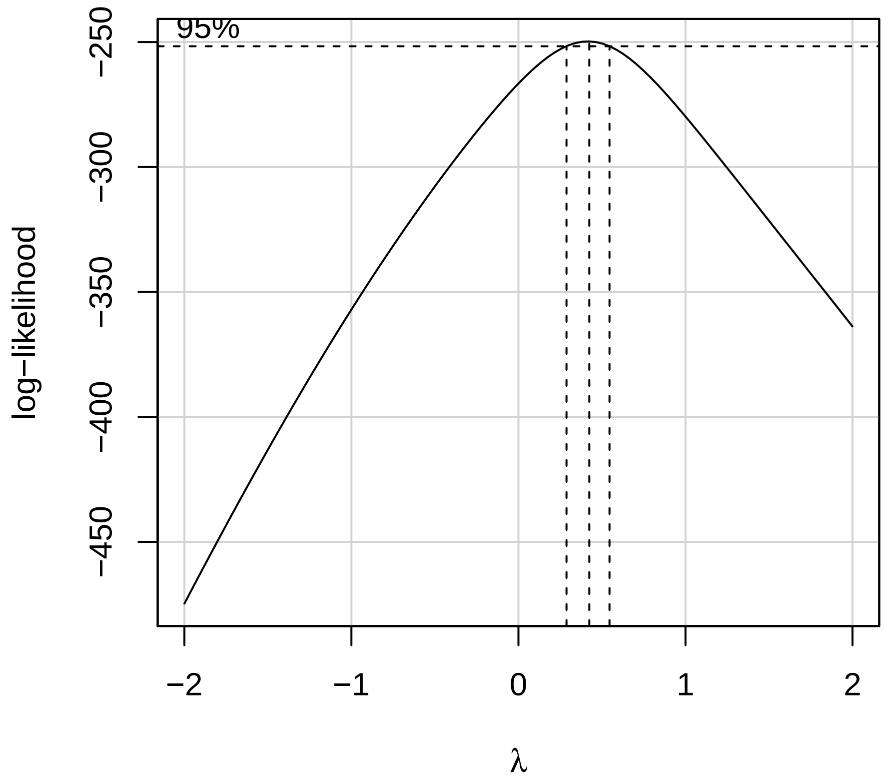
Part 1

```
lin.mod <- lm(Distance ~ Speed, data = stopping.df)

invResPlot(lin.mod, lambda = seq(-5, 5))
```




```
       lambda       RSS
1   0.4849737  4463.944
2  -5.0000000 57340.753
3  -4.0000000 56863.345
4  -3.0000000 55499.171
5  -2.0000000 50668.115
6  -1.0000000 33149.061
7   0.0000000  7890.434
8   1.0000000  7293.835
9   2.0000000 19819.302
10  3.0000000 30597.911
11  4.0000000 37471.316
12  5.0000000 41718.391
```

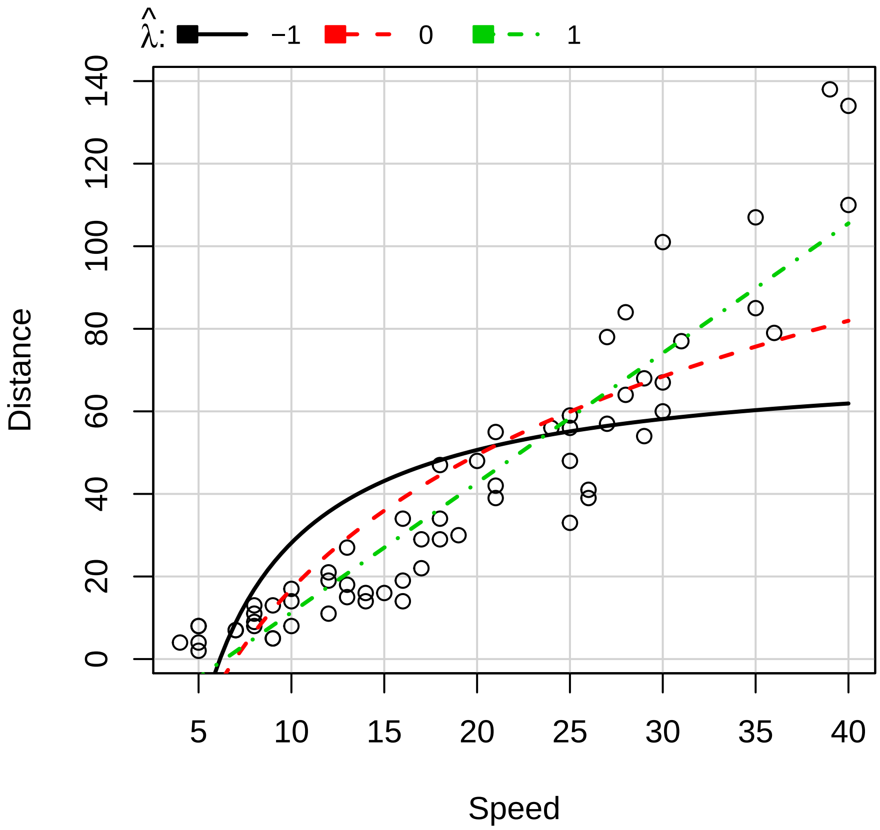
```
boxCox(lin.mod)
```



The optimal λ is 0.485. However, $\lambda = .5$ is in the 95% confidence interval, and it is the only integer value in the interval. So we will use $\lambda = .5$ for this problem.

Part 2

```
invTranPlot(Distance ~ Speed, data = stopping.df,
            lambda = seq(-1, 1), optimal = FALSE)
```

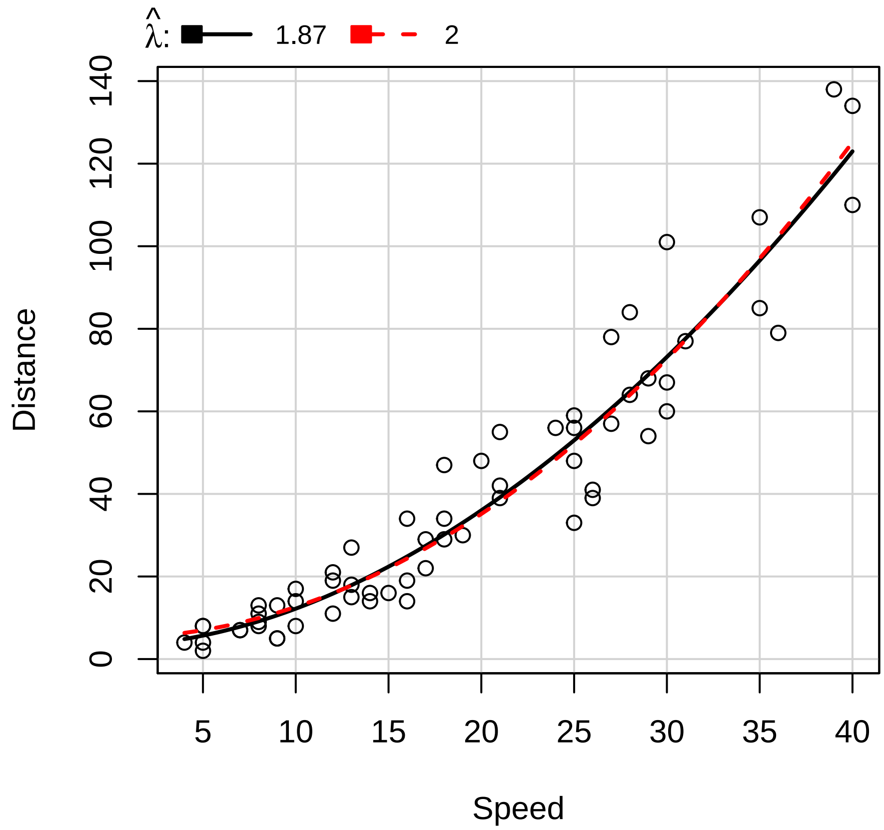



```
  lambda       RSS
1     -1 34951.108
2      0 18844.172
3      1  8310.166
```

From the scatterplot with fitted lines, we can see that none of these values of λ fit the data very well. $\lambda = 0$ or -1 do not lie near the points, and $\lambda = 1$ fails to capture the curvature, resulting in a pattern in the residuals.

Part 3

```r
invTranPlot(Distance ~ Speed, data = stopping.df, lambda = 2)
```




```
    lambda      RSS
1 1.868443 5823.372
2 2.000000 5869.232
```

```
invTranEstimate(stopping.df$Speed, stopping.df$Distance)
```

```
$lambda
[1] 1.868443

$lowerCI
[1] 1.617086

$upperCI
[1] 2.135815
```

The optimal λ and $\lambda = 2$ produce very similar results. In addition, 2 is in the 95% CI for the optimal λ (minimizing the log-likelihood).

Part 4

```
hald.mod <- lm(Distance ~ Speed + I(Speed ** 2),
               weights = I(Speed ** -2),
               data = stopping.df)
summary(hald.mod)
```

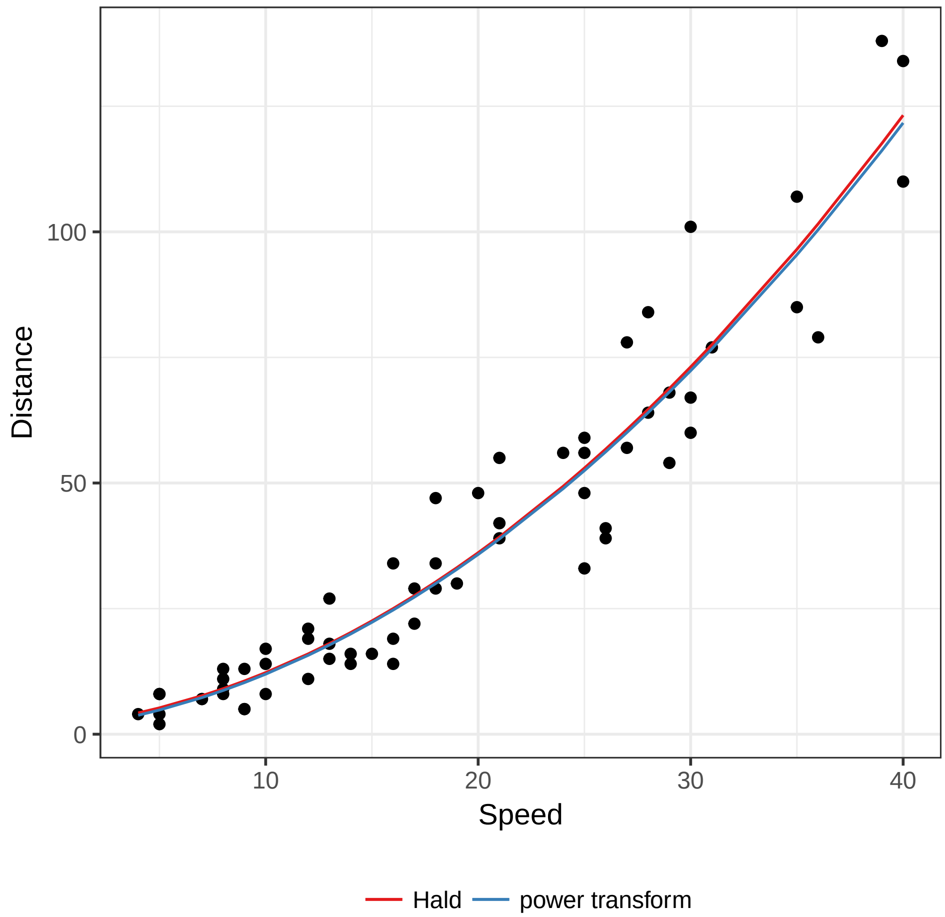
```
Call:
lm(formula = Distance ~ Speed + I(Speed^2), data = stopping.df,
```

```
    weights = I(Speed^-2))

Weighted Residuals:
     Min       1Q   Median       3Q      Max
-0.79915 -0.32983 -0.02599  0.27541  0.92972

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.50605    2.03544   0.740    0.462
Speed        0.41968    0.34326   1.223    0.226
I(Speed^2)   0.06557    0.01057   6.205  5.9e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4514 on 59 degrees of freedom
Multiple R-squared:  0.9131,    Adjusted R-squared:  0.9101
F-statistic: 309.8 on 2 and 59 DF,  p-value: < 2.2e-16
```

```r
power.mod <- lm(sqrt(Distance) ~ Speed,
                data = stopping.df)
summary(power.mod)
```

```
Call:
lm(formula = sqrt(Distance) ~ Speed, data = stopping.df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.49948 -0.54761  0.00469  0.53153  1.54350

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.932396   0.197909   4.711  1.5e-05 ***
Speed       0.252466   0.009274  27.223  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7209 on 60 degrees of freedom
Multiple R-squared:  0.9251,    Adjusted R-squared:  0.9239
F-statistic: 741.1 on 1 and 60 DF,  p-value: < 2.2e-16
```

```r
stopping.df %<>%
  dp$mutate(hald.pred = predict(hald.mod, newdata = stopping.df),
            power.pred = predict(power.mod, newdata = stopping.df) ** 2,
            hald.resid = Distance - hald.pred,
            power.resid = Distance - power.pred)
ggplot(stopping.df) +
  geom_point(aes(x = Speed, y = Distance))  +
  geom_line(aes(x = Speed, y = hald.pred, colour = 'Hald')) +
  geom_line(aes(x = Speed, y = power.pred, colour = 'power transform')) +
  theme(legend.position = 'bottom') +
  labs(colour = NULL) +
  scale_colour_brewer(palette = 'Set1')
```




```
mean(stopping.df$hald.resid ** 2)
```

```
[1] 93.77798
```

```
mean(stopping.df$power.resid ** 2)
```

```
[1] 94.10235
```

```
rSquared(stopping.df$Distance, stopping.df$hald.resid)
```

```
          [,1]
[1,] 0.9144326
```

```
rSquared(stopping.df$Distance, stopping.df$power.resid)
```

```
          [,1]
[1,] 0.9141366
```

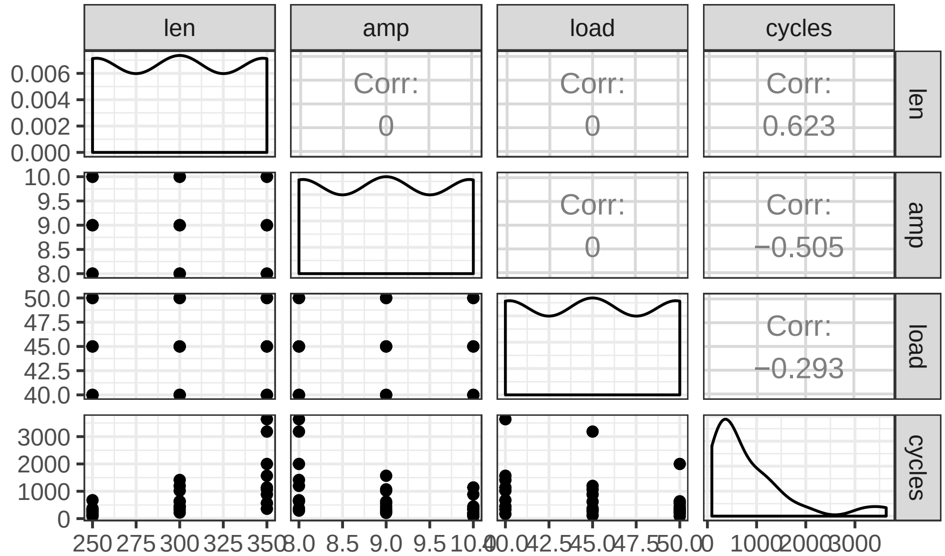
Problem 2

[From ALR 8.6]

```
wool.df <- car::Wool
```

Part 1

```r
ggpairs(wool.df)
```




```r
summary(wool.df)
```

```
      len           amp           load          cycles      
 Min.   :250   Min.   : 8   Min.   :40   Min.   :  90.0  
 1st Qu.:250   1st Qu.: 8   1st Qu.:40   1st Qu.: 312.0  
 Median :300   Median : 9   Median :45   Median : 566.0  
 Mean   :300   Mean   : 9   Mean   :45   Mean   : 861.4  
 3rd Qu.:350   3rd Qu.:10   3rd Qu.:50   3rd Qu.:1105.0  
 Max.   :350   Max.   :10   Max.   :50   Max.   :3636.0  
```

```r
wool.df %>%
  dp$select(len, amp, load) %>%
  table()
```

```
, , load = 40

     amp
len   8 9 10
  250 1 1  1
  300 1 1  1
  350 1 1  1

, , load = 45

     amp
len   8 9 10
  250 1 1  1
  300 1 1  1
  350 1 1  1

, , load = 50

     amp
```

```
len   8 9 10
  250 1 1  1
  300 1 1  1
  350 1 1  1
```

```r
dim(wool.df)
```

```
[1] 27  4
```

The values for `len`, `amp`, and `load` consist of just 3 values each. Each triple is unique, which matches the number of rows of the data frame (3^3). The values are evenly spaced out.

Part 2

```r
wool.df %<>%
  dp$mutate(len = as.factor(len),
            amp = as.factor(amp),
            load = as.factor(load))

factor.2.mod <- lm(cycles ~ len * amp + len * load + amp * load,
                   data = wool.df)
summary(factor.2.mod)
```

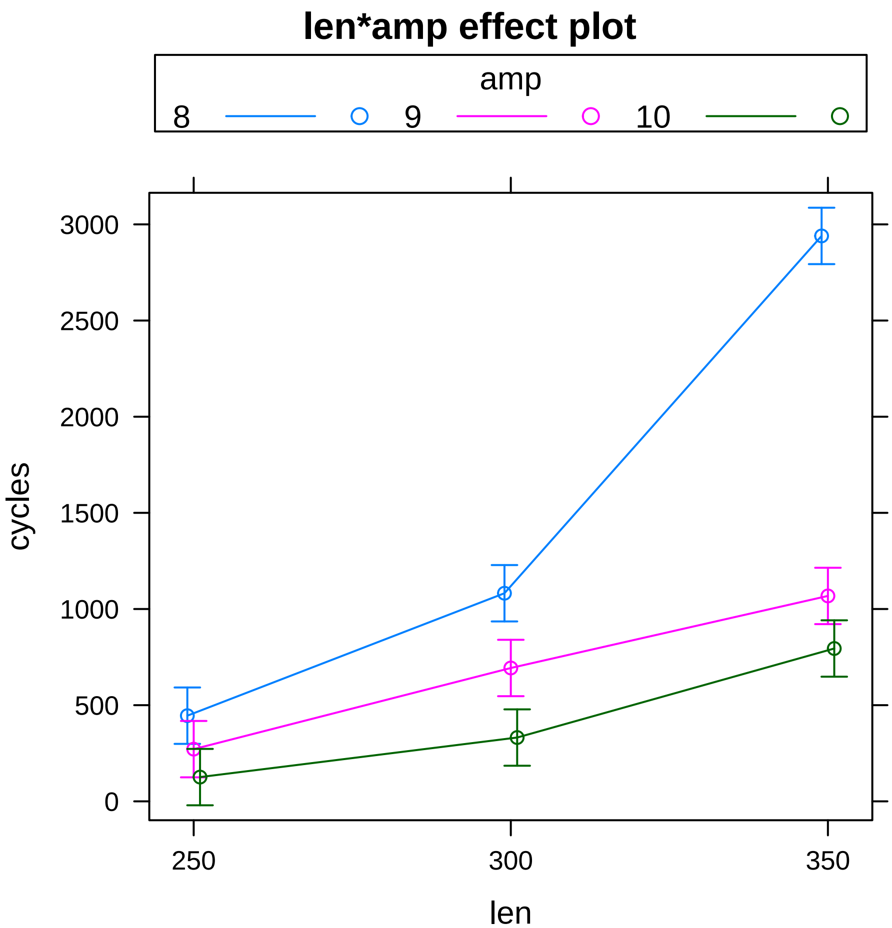
```
Call:
lm(formula = cycles ~ len * amp + len * load + amp * load, data = wool.df)

Residuals:
     Min       1Q   Median       3Q      Max 
-127.593  -39.148   -9.037   58.074  117.074 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    6.826e+02  9.237e+01   7.390 7.69e-05 ***
len300         7.809e+02  1.161e+02   6.728 0.000148 ***
len350         2.895e+03  1.161e+02  24.946 7.13e-09 ***
amp9          -2.944e+02  1.161e+02  -2.537 0.034879 *  
amp10         -5.713e+02  1.161e+02  -4.923 0.001160 ** 
load45        -2.041e+02  1.161e+02  -1.759 0.116697    
load50        -5.077e+02  1.161e+02  -4.374 0.002368 ** 
len300:amp9   -2.147e+02  1.271e+02  -1.688 0.129813    
len350:amp9   -1.698e+03  1.271e+02 -13.355 9.45e-07 ***
len300:amp10  -4.310e+02  1.271e+02  -3.390 0.009502 ** 
len350:amp10  -1.826e+03  1.271e+02 -14.362 5.40e-07 ***
len300:load45 -1.003e+02  1.271e+02  -0.789 0.452782    
len350:load45 -2.593e+02  1.271e+02  -2.040 0.075709 .  
len300:load50 -3.323e+02  1.271e+02  -2.614 0.030944 *  
len350:load50 -9.427e+02  1.271e+02  -7.414 7.52e-05 ***
amp9:load45    5.907e-13  1.271e+02   0.000 1.000000    
amp10:load45   1.843e+02  1.271e+02   1.450 0.185155    
amp9:load50    3.613e+02  1.271e+02   2.842 0.021747 *  
amp10:load50   5.717e+02  1.271e+02   4.496 0.002012 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 110.1 on 8 degrees of freedom
Multiple R-squared:  0.9952,    Adjusted R-squared:  0.9844
F-statistic: 92.25 on 18 and 8 DF,  p-value: 2.537e-07
```

```
Anova(factor.2.mod)
```

```
Anova Table (Type II tests)

Response: cycles
           Sum Sq Df  F value    Pr(>F)
len       8182253  2 337.4408 1.884e-08 ***
amp       5624249  2 231.9473 8.260e-08 ***
load      1753097  2  72.2987 7.554e-06 ***
len:amp   3555537  4  73.3162 2.433e-06 ***
len:load   732881  4  15.1122 0.0008457 ***
amp:load   283609  4   5.8481 0.0167886 *
Residuals   96992  8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(effect('len:amp', factor.2.mod), multiline = TRUE, ci.style = 'bars')
```



For a significance level of $\alpha = 0.05$, we reject the null hypothesis that the coefficients for the `len` and `amp` interaction terms is 0.

Part 3

```r
factor.1.mod <- lm(cycles ~ len + amp + load, data = wool.df)
summary(factor.1.mod)
```

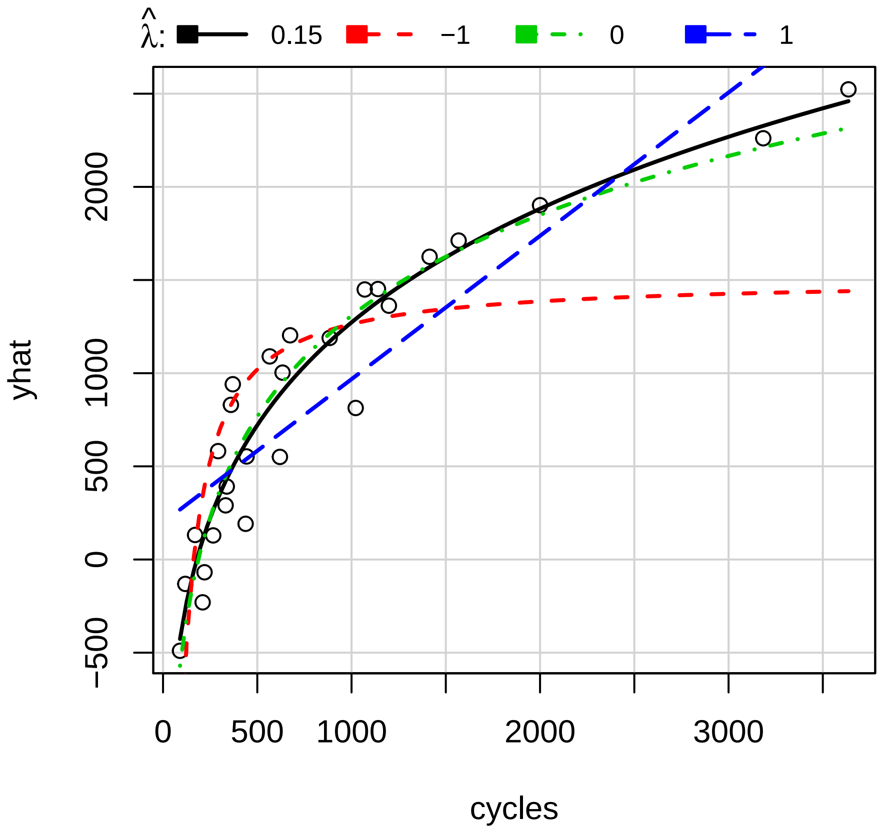
```

Call:
lm(formula = cycles ~ len + amp + load, data = wool.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-570.81 -308.43  -53.81  227.57 1112.63 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   1203.4      246.0   4.891 8.83e-05 ***
len300         421.4      227.8   1.850 0.079096 .  
len350        1320.0      227.8   5.795 1.14e-05 ***
amp9          -811.6      227.8  -3.563 0.001948 ** 
amp10        -1071.7      227.8  -4.705 0.000136 ***
load45        -262.6      227.8  -1.153 0.262611    
load50        -621.7      227.8  -2.729 0.012918 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 483.2 on 20 degrees of freedom
Multiple R-squared:  0.7692,	Adjusted R-squared:  0.6999 
F-statistic: 11.11 on 6 and 20 DF,  p-value: 1.769e-05
```

```r
anova(factor.2.mod, factor.1.mod)
```

```
Analysis of Variance Table

Model 1: cycles ~ len * amp + len * load + amp * load
Model 2: cycles ~ len + amp + load
  Res.Df     RSS  Df Sum of Sq      F    Pr(>F)    
1      8   96992                                   
2     20 4669020 -12  -4572028 31.425 2.158e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

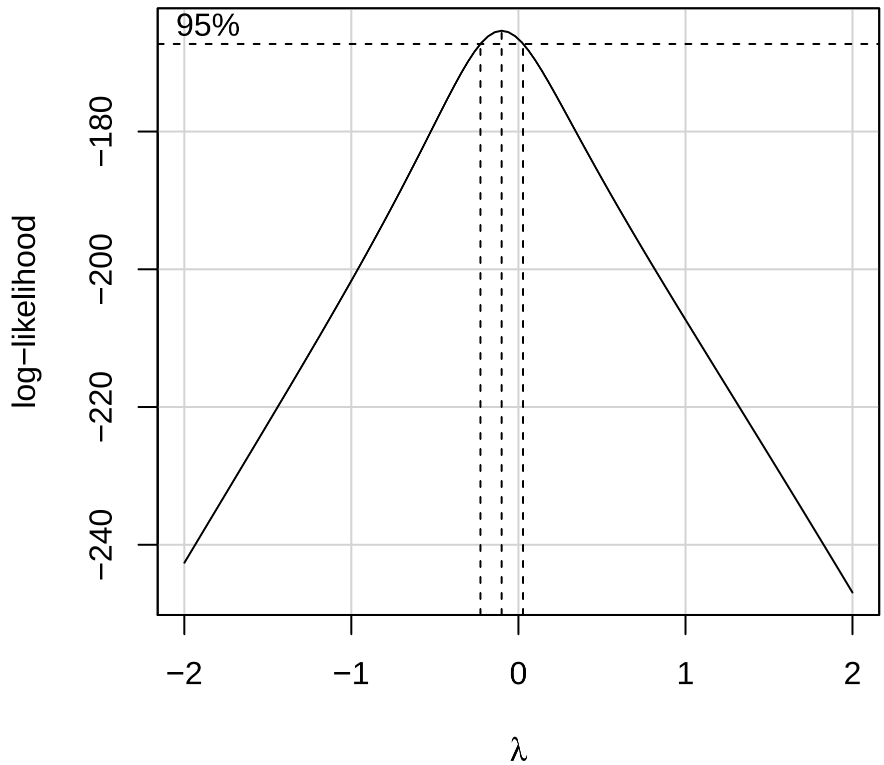
The ANOVA test confirms the text's assertion.

```r
invResPlot(factor.1.mod)
```




```
      lambda     RSS
1  0.1452334 1340826
2 -1.0000000 5544947
3  0.0000000 1429311
4  1.0000000 3591351
```

```
boxCox(factor.1.mod)
```




```r
summary(powerTransform(factor.1.mod))
```

```
bcPower Transformation to Normality
   Est Power Rounded Pwr Wald Lwr bnd Wald Upr Bnd
Y1   -0.1005           0      -0.2249       0.0239

Likelihood ratio tests about transformation parameters
                           LRT df      pval
LR test, lambda = (0)  2.38372  1 0.1226053
LR test, lambda = (1) 83.89818  1 0.0000000
```

The best value of λ (the one that maximizes the log-likelihood) is -0.1005. However, 0 is within the 95% confidence interval, so we cannot say that -0.1005 is better than 0 (for $\alpha = 0.05$). So we will select $\lambda = 0$.

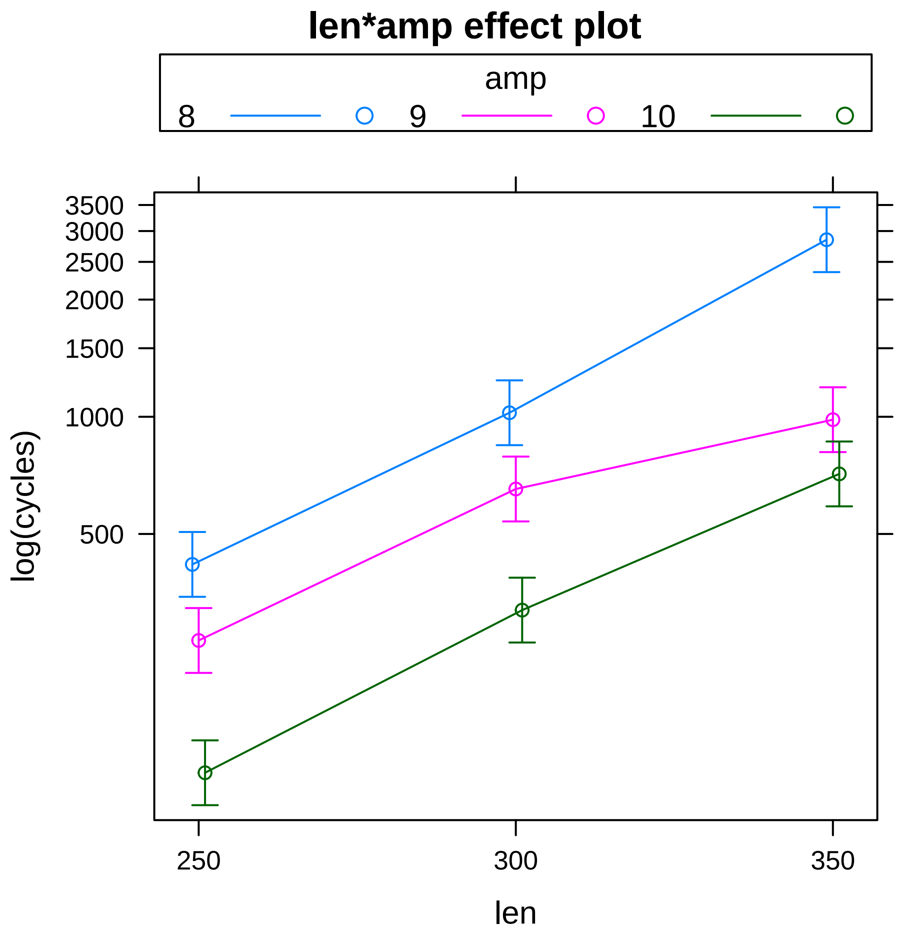
Part 4

```r
factor.1.log.mod <- lm(log(cycles) ~ len + amp + load, data = wool.df)
factor.2.log.mod <- lm(log(cycles) ~ len * amp + len * load + amp * load,
                       data = wool.df)
anova(factor.2.log.mod, factor.1.log.mod)
```

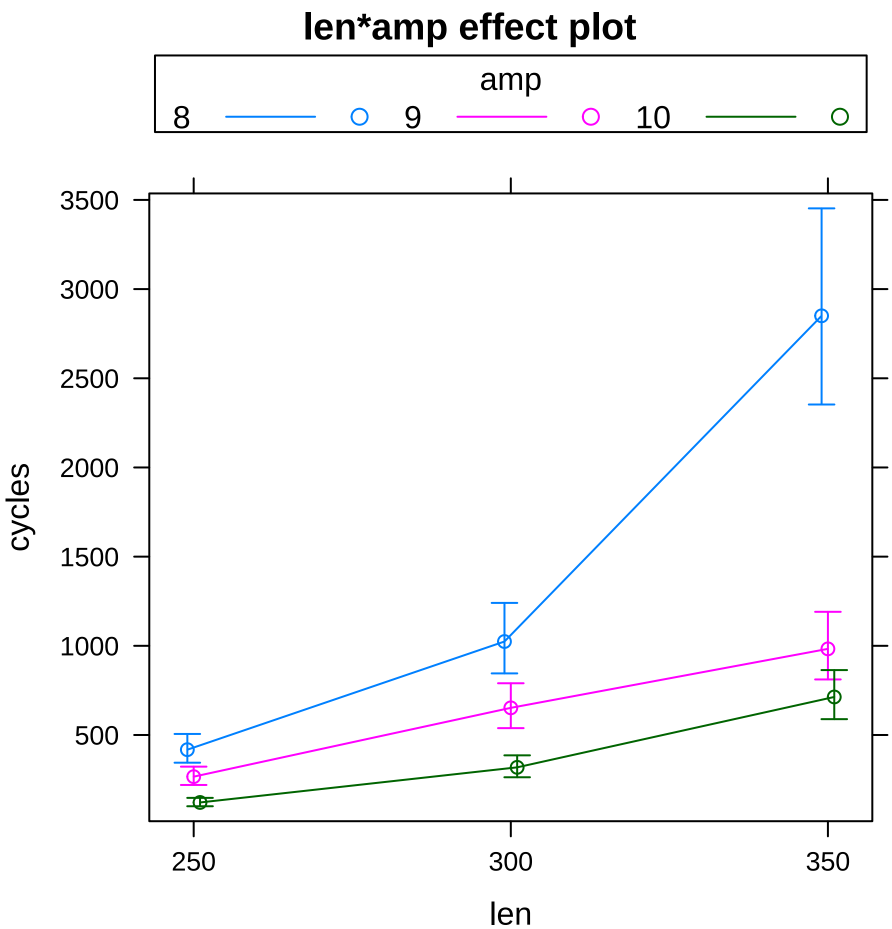
```
Analysis of Variance Table

Model 1: log(cycles) ~ len * amp + len * load + amp * load
Model 2: log(cycles) ~ len + amp + load
  Res.Df     RSS  Df Sum of Sq     F Pr(>F)
1      8 0.16591
2     20 0.71742 -12  -0.55151 2.216 0.1325
```

From the ANOVA test, we fail to reject the null hypothesis that all of the coefficients for the interaction terms is 0.

```
plot(Effect(c('len', 'amp'), factor.2.log.mod,
            transformation = list(link = log, inverse = exp)),
     multiline = TRUE,
     ci.style = 'bars')
```




```
plot(Effect(c('len', 'amp'), factor.2.log.mod,
            transformation = list(link = log, inverse = exp)),
     multiline = TRUE,
     axes = list(y = list(type = 'response', lab = 'cycles')),
     ci.style = 'bars')
```



The confidence interval increases with `cycles`.