GRM HW2

Ericka Smith

10/6/2020

```r
corn <- read.table("~/generalized_regression_models/hw2/corn.txt", quote="\"", comment.char="")
names(corn) <- c("amt_fert", "block_num", "obs_yield")
corn$block_num %<>% factor()
```

Part A

```r
summary(aov(obs_yield ~ block_num + amt_fert, data=corn))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## block_num    2   6745    3373   8.765 0.000625 ***
## amt_fert     1   5819    5819  15.124 0.000336 ***
## Residuals   44  16930     385                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The sum of squares associated with variation: between blocks is 6745, with nitrogen is 5819, and with departures from additivity is 16930.

Part B

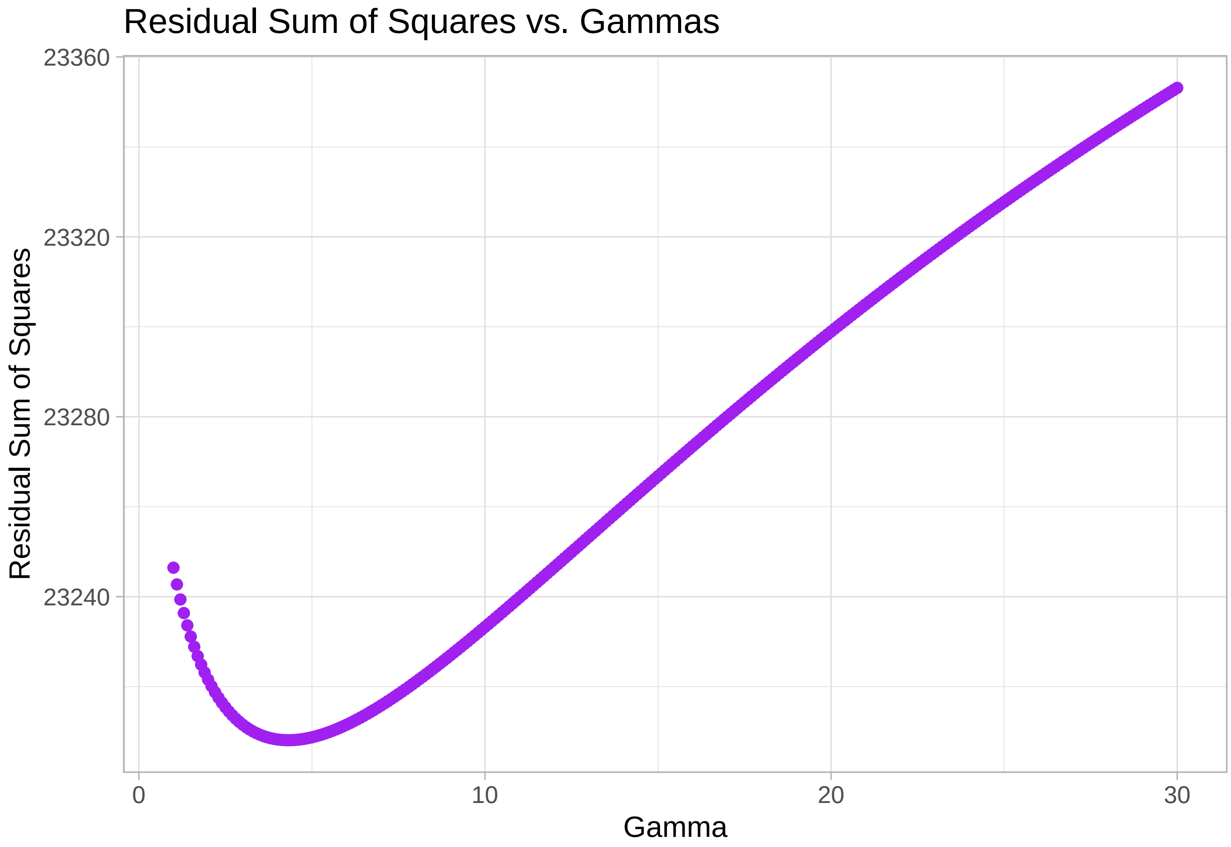
Since the block number factor is significant at level $\alpha = 0.05$ blocking does appear to have been effective in this experiment.$(p = 0.000625)$

Part C

```r
# make sequence of gammas and resy df to hold data
gamma_nums <- seq(1,30,by=0.1)
resy <- data.frame(
  gammas =gamma_nums,
  rss = rep(0, length(gamma_nums)))
# function that gives the residual sums of squares
get_rss <- function(the_gamma){
  anova(lm(obs_yield ~ log(the_gamma+amt_fert), data=corn))["Residuals", "Sum Sq"]
}
# use map to combine feed gamma values into rss function
resy$rss <- map_dbl(resy$gammas, ~get_rss(.x))
# find least squares estimate
(lse <- resy$gammas[which(resy$rss==min(resy$rss))])
```

```
## [1] 4.3
```

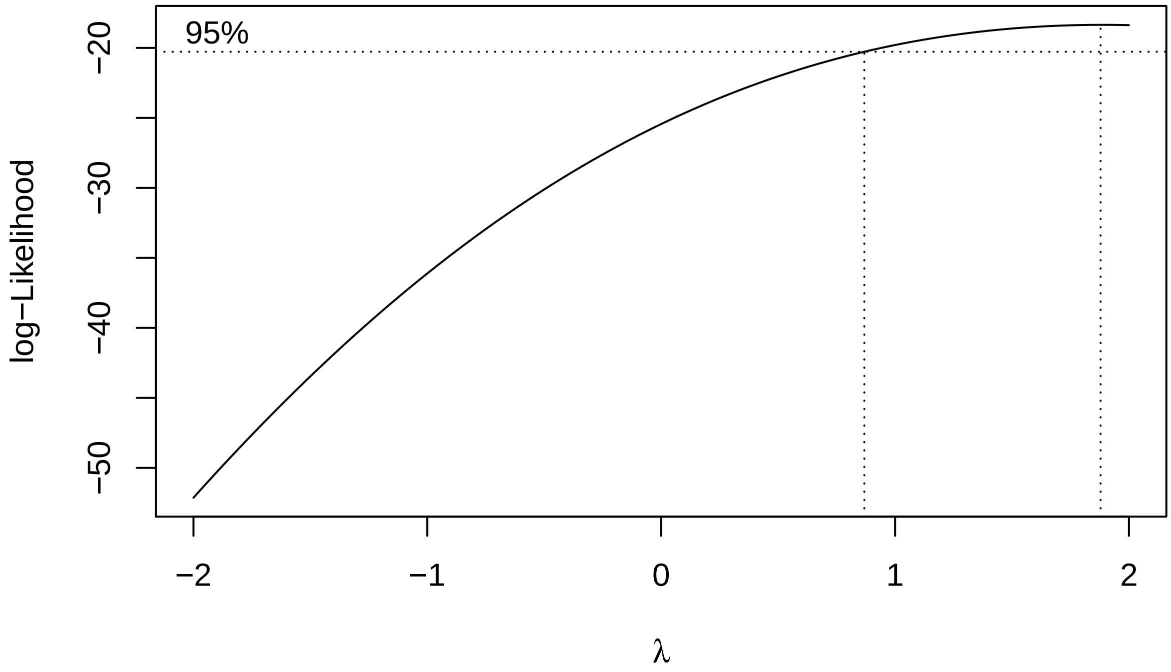
```
# make plot
ggplot(resy)+
  geom_point(data=resy,aes(x=gammas, y=rss), color="purple")+
  theme_light()+
  ggtitle("Residual Sum of Squares vs. Gammas")+
  xlab("Gamma")+
  ylab("Residual Sum of Squares")
```



The least squares estimate is 4.3.

Part D

```
mod <- lm(obs_yield ~ log(lse+ amt_fert), data=corn)
the_bc <- boxcox(mod)#, lambda = seq(0.5,2.5, by=0.1))
```



```
(mle_lambda <- the_bc$x[which(the_bc$y==max(the_bc$y))])
```

```
## [1] 1.878788
```

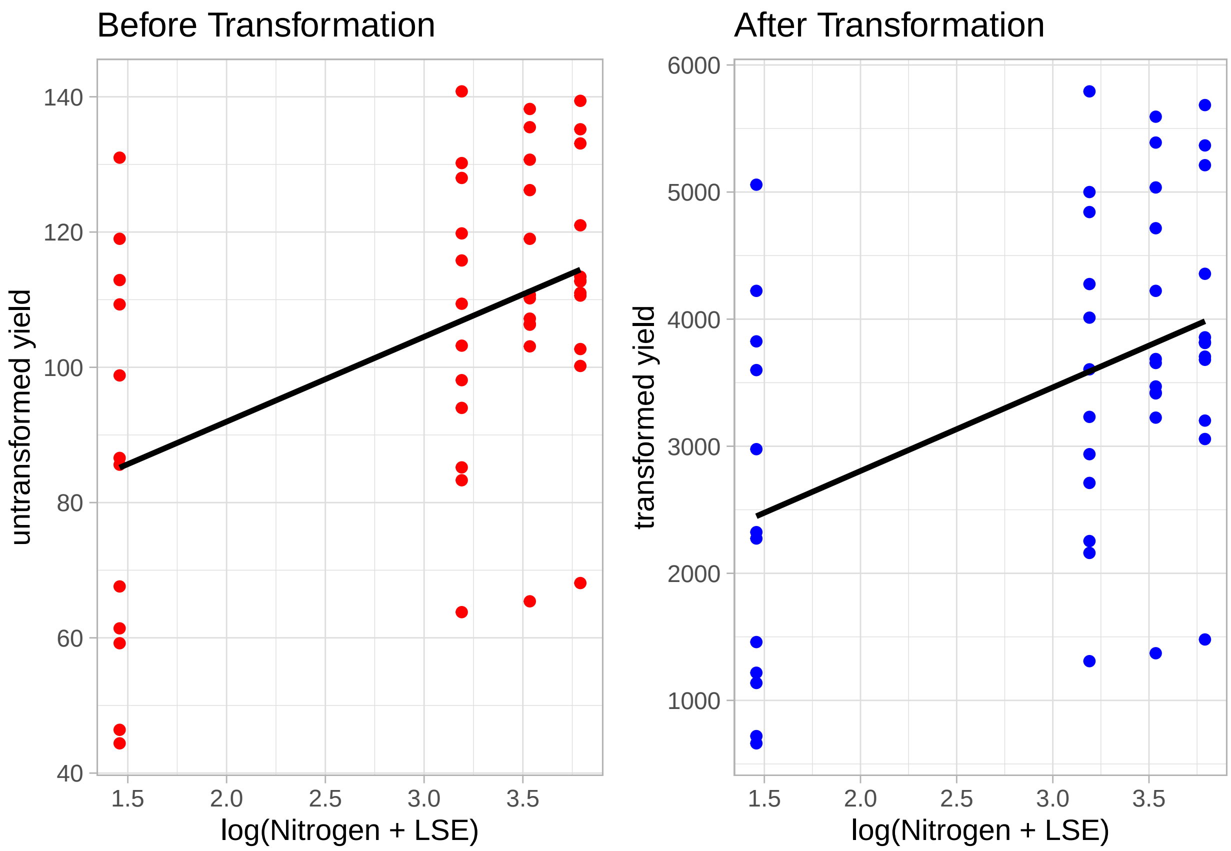
The model in C cannot be substantially improved by response transformation using the Box-Cox method. Though $\hat{\lambda}_{MLE} = 1.878788$, 1 is within our 95% confidence interval.

Since $\lambda = 1$ is equivalent to using the original data this tells us that the transformation is unnecessary.

(Part D Continued)

```
# transform data
transformed <- ((corn$obs_yield^mle_lambda)-1)/mle_lambda
# original plot
p_before <- ggplot()+
  geom_point(aes(x=log(lse+corn$amt_fert), y=corn$obs_yield), color="red")+
    stat_smooth(aes(x=log(lse+corn$amt_fert), y=corn$obs_yield), color="black", method = "lm", se=FALSE
  theme_light()+
  ggtitle("Before Transformation")+
  ylab("untransformed yield")+
  xlab("log(Nitrogen + LSE)")
# plot with transformed data
p_after <- ggplot()+
  geom_point(aes(x=log(lse+corn$amt_fert), y=transformed), color="blue")+
  stat_smooth(aes(x=log(lse+corn$amt_fert), y=transformed),color="black", method = "lm", se=FALSE)+
  theme_light()+
  ggtitle("After Transformation")+
  xlab("log(Nitrogen + LSE)")+
  ylab("transformed yield")
# arrange
ggarrange(p_before, p_after, ncol=2, nrow = 1)
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



Just to be thorough I have created plots with transformed and untransformed data. As you can see, the shape of the data is approximately the same. This further provides evidence against using a Box-Cox transformation.

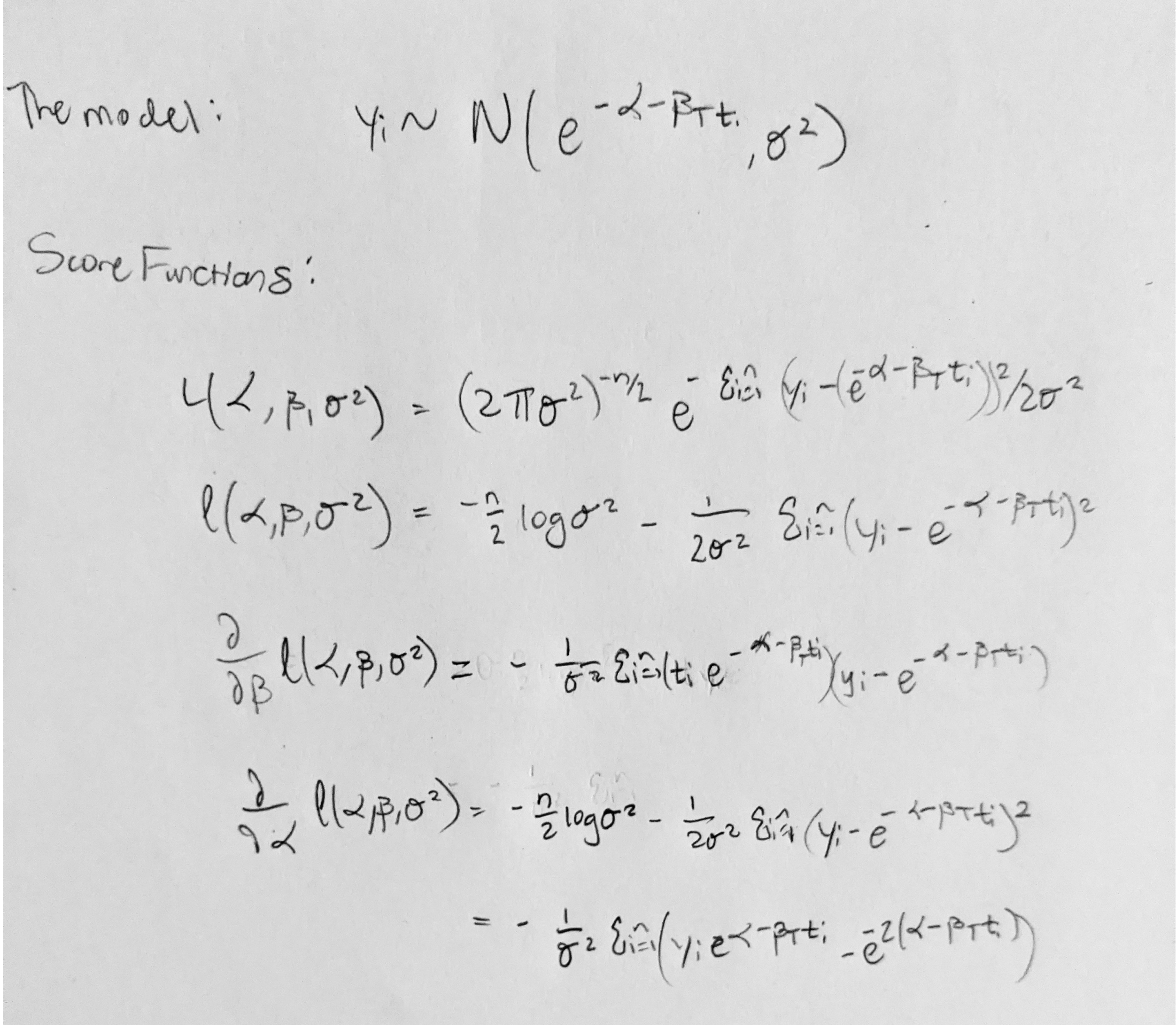
Part E

Based on my analysis higher amounts of Nitrogen fertilizer appear to be associated with greater corn yield, after accounting for differences between blocks.

Problem 2

```
beans <- data.frame(
  temps = c(0, 10, 20, 0, 10, 20, 0, 10, 20, 0, 10, 20),
  weeks = c(2,2,2,4,4,4,6,6,6,8,8,8),
  aa_loss = c(45,45,34,47,43,28,46,41,21,46,37,16)
)
beans$temps %<>% factor()
```

Part A



The model: $y_i \sim N(e^{-\alpha-\beta_T t_i}, \sigma^2)$

Score Functions:

$$L(\alpha, \beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} e^{-\sum_{i=1}^n (y_i - (e^{-\alpha-\beta_T t_i}))^2/2\sigma^2}$$

$$\ell(\alpha, \beta, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - e^{-\alpha-\beta_T t_i})^2$$

$$\frac{\partial}{\partial\beta}\ell(\alpha, \beta, \sigma^2) = -\frac{1}{\sigma^2}\sum_{i=1}^n (t_i e^{-\alpha-\beta_T t_i})(y_i - e^{-\alpha-\beta_T t_i})$$

$$\frac{\partial}{\partial\alpha}\ell(\alpha, \beta, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - e^{-\alpha-\beta_T t_i})^2$$

$$= -\frac{1}{\sigma^2}\sum_{i=1}^n (y_i e^{\alpha-\beta_T t_i} - e^{2(\alpha-\beta_T t_i)})$$

Figure 1: image

```
mod2 <- glm(aa_loss ~ temps:weeks, family=gaussian(link = "log"), data = beans)
summary(mod2)
```

```
##
## Call:
## glm(formula = aa_loss ~ temps:weeks, family = gaussian(link = "log"),
##     data = beans)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.49820  -0.38546   0.08059   0.80866   0.86028
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.8354389  0.0186464 205.694 3.49e-16 ***
## temps0:weeks  -0.0007716  0.0037540  -0.206 0.842286
## temps10:weeks -0.0236121  0.0040351  -5.852 0.000382 ***
## temps20:weeks -0.1329785  0.0061597 -21.588 2.23e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1.124727)
##
##     Null deviance: 1226.9167  on 11  degrees of freedom
## Residual deviance:    8.9978  on  8  degrees of freedom
## AIC: 40.599
##
## Number of Fisher Scoring iterations: 3
```

Printed above is the summary from this model. Each temperature factor grouping has it's own row along with the intercept. For the 0 degrees Fahrenheit group the coefficient was not significant ($p = 0.842286$) but each of the other coefficients were significant, including the intercept.

Part B

```
betas <- c(0.0007716,0.0236121,0.1329785)
log(2)/betas
```

```
## [1] 898.324495  29.355592   5.212476
```

For the 0 degrees Fahrenheit group we can argue that the ascorbic acid concentration remains the same indefinitely due to zero being within the confidence interval for the estimate of the beta coefficient. However I'll move forward under the assumption that we can use the given estimates.

Since $E(Y) = e^{-\alpha-\beta_T t_i}$ and $e^{-\alpha}$ is our initial concentration we want to find where $e^{-\beta_T t_i} = 2$

By rearranging we can see that this occurs at $t_i = \frac{log(2)}{\beta_T}$

Solving for t we find that the ascorbic acid concentration is expected to halve for

- 0 degrees Fahrenheit after 898.34 weeks
- 10 degrees Fahrenheit after 29.36 weeks
- 20 degrees Fahrenheit after 5.21 weeks

I considered that using the Delta method or Fieller's may be best for finding the confidence intervals. However, I don't know that I believe that level of complexity is warranted here.

Problem 3

Based off of this model we can say that there may be some correlation between education and extroversion based on the change in p-value and estimate for x1 (education) when it is regressed with x2 (extroversion) vs. when it is not included. However the difference is quite small and in both models education is not a significant predictor of sales performance, while extroversion is a significant predictor. This makes it difficult to get any good information about x2 (education) because it does not have evidence of a relationship with our y variable (sales performance).