Homework5

Dima Mikhaylov

10/5/2021

Question 1:

Why do we transform the response variable when the constant variance assumption is not met, instead of transforming the predictor variable?

Comment: Variance of response actually only depends on the variance of the error terms, because all other parameters are fixed in the estimation process. Therefore, transforming the response variable is used to mitigate non-constant variance problem, because it effectively changes how sensitive is the respose to the variability in the data. Generally, transforming response using a concave function such as `log()` or `sqrt()` should work well.

Question 2:

```
library(faraway)
data(cornnit)
head(cornnit)
```

```
##   yield nitrogen
## 1   115        0
## 2   128       75
## 3   136      150
## 4   135      300
## 5    97        0
## 6   150       75
```

a. What is the response variable and predictor for this study? Create a scatterplot of the data, and interpret the scatterplot

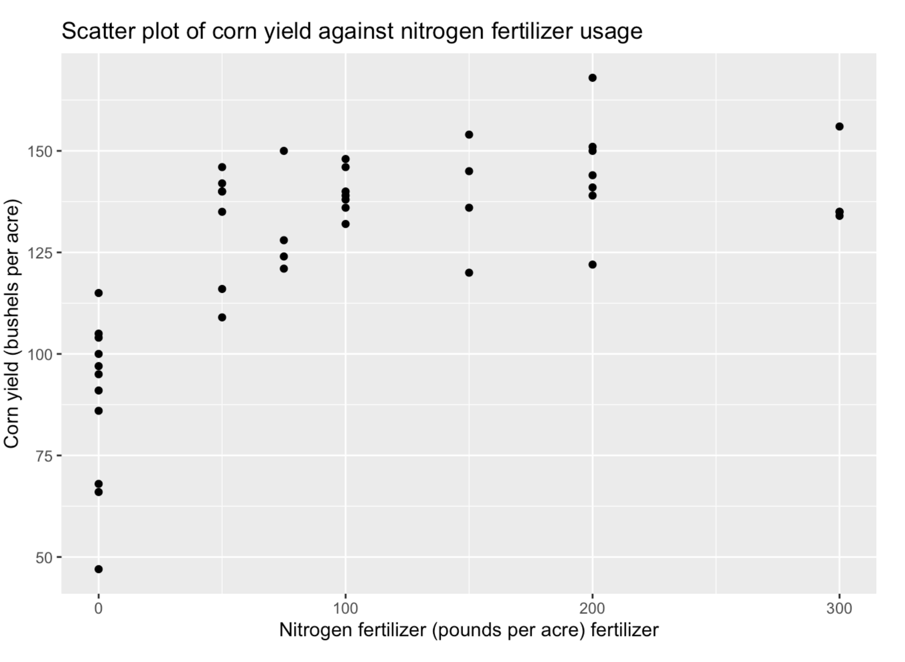
```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✓ ggplot2 3.3.5     ✓ purrr   0.3.4
## ✓ tibble  3.1.3     ✓ dplyr   1.0.7
## ✓ tidyr   1.1.3     ✓ stringr 1.4.0
## ✓ readr   2.0.1     ✓ forcats 0.5.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
ggplot(cornnit, aes(x=nitrogen, y=yield))+
  geom_point()+
  labs(x="Nitrogen fertilizer (pounds per acre) fertilizer",
       y="Corn yield (bushels per acre)",
       title="Scatter plot of corn yield against nitrogen fertilizer usage")
```

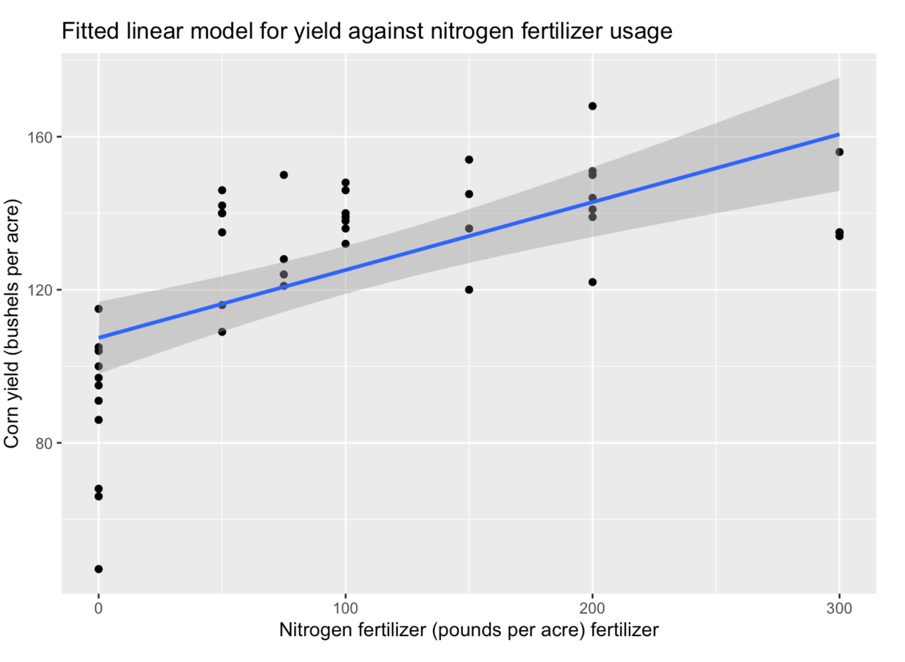


Comments: corn yeild is the response (y), and fertilizer usage is the predictor (x). Overall, there seems to be some sort of reltionship between the variables, however it is probably not a linear relationship.

b. Fit a linear regression without any transformations. Create the corresponding residual plot. Based only on the residual plot, what transformation will you consider first? Be sure to explain your reason.

```
ggplot(cornnit, aes(x=nitrogen, y=yield))+
  geom_point()+
  geom_smooth(method = "lm", se = TRUE)+
  labs(x="Nitrogen fertilizer (pounds per acre) fertilizer",
       y="Corn yield (bushels per acre)",
       title="Fitted linear model for yield against nitrogen fertilizer usage")
```

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



It seems that variance is affecting confidence bounds for the smaller x values, as well as for the larger x valuess. For the former observations the reason could be related to several outliers, extremely low yields recorded at zero nitrogen usage. For the later, very high nitrogen usage, variance of yields may be increasing simply due to a larger value of yield.

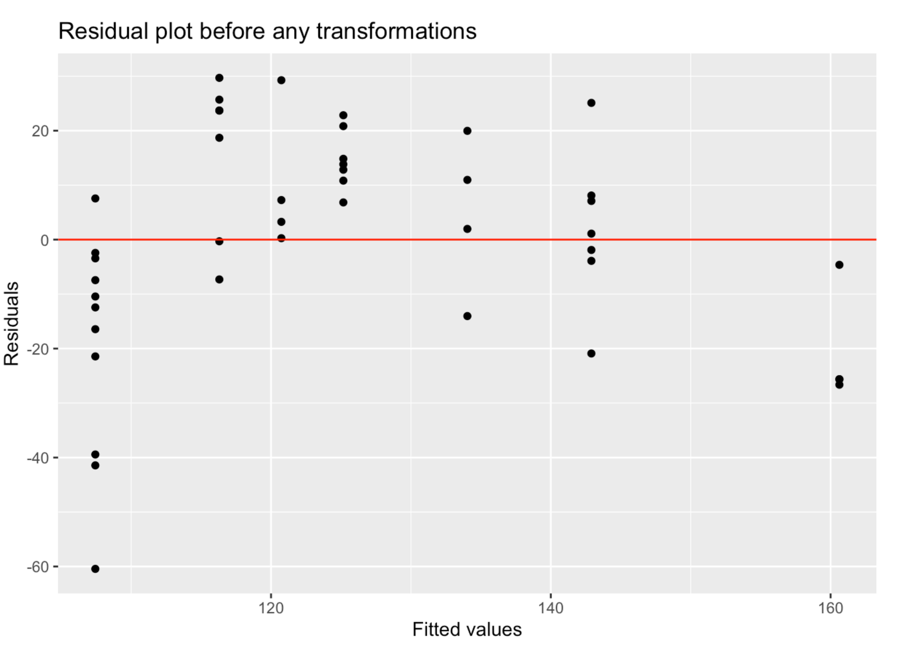
```
base_model = lm(yield ~ nitrogen, data=cornnit)
summary(base_model)
```

```
##
## Call:
## lm(formula = yield ~ nitrogen, data = cornnit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.439 -10.939   1.534  14.082  29.697
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 107.43864    4.66622   23.02  < 2e-16 ***
## nitrogen      0.17730    0.03377    5.25 4.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.53 on 42 degrees of freedom
## Multiple R-squared:  0.3962, Adjusted R-squared:  0.3818
## F-statistic: 27.56 on 1 and 42 DF,  p-value: 4.713e-06
```

The summary output confirms some relationship with significant F-statistic, but poor R-squared. Additional analysis of residuals would be needed to confirm linearity of this relationship, i.e. confirm the true predictive power of the model.

Create the corresponding residual plot.

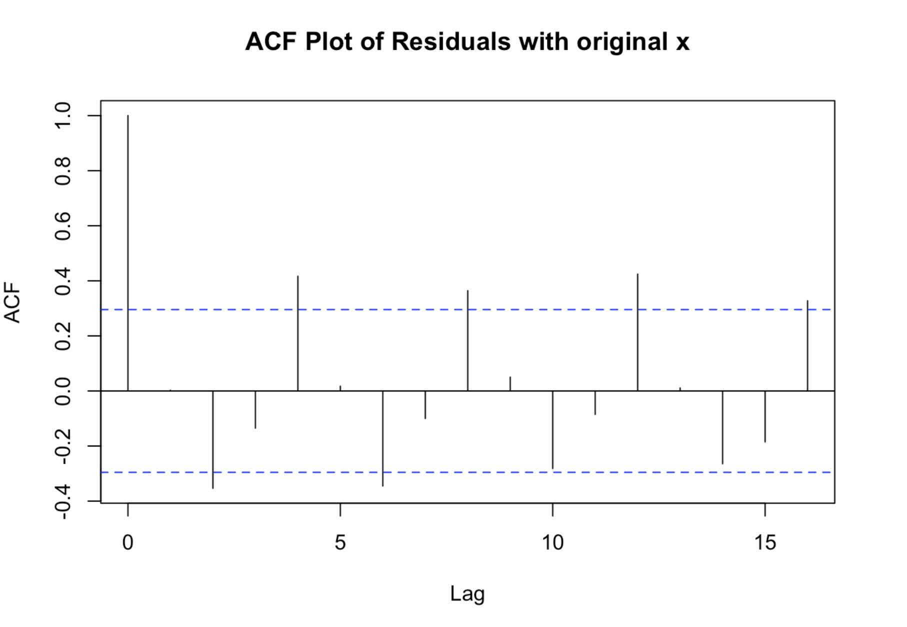
```
y_fitted <- base_model$fitted.values
residuals <- base_model$residuals
cornnit <- data.frame(cornnit, y_fitted, residuals)

ggplot(cornnit, aes(x=y_fitted, y=residuals))+
  geom_point()+
  geom_hline(yintercept = 0, color='red')+
  labs(x="Fitted values", y="Residuals", title= "Residual plot before any transformations")
```



Based only on the residual plot, what transformation will you consider first? Be sure to explain your reason.

Comment: based on the Residual plot, it seems errors may be following a certain pattern. Errors need to be i.i.d and on the Residual plot this should result in points being randomly scattered, i.e. 1) not displaying any pattern (mean 0), and 2) spread of residuals for each fitted value of x should be constant (constant variance). In this case, it is hard to tell about non-constant variance, but non-zero mean is obviously present, as some observations are clustered below the zero line and some are clearly grouped above the zero line. Additional checks for auto-correlation may be warranted.

```
acf(residuals, main="ACF Plot of Residuals with original x")
```



Comment: from the ACF plot (above) it also seems that residuals are correlated in lags 2, 4, 6, 8, (almost 10), 12, (almost 14), 16 and this is why the estimated standard errors from the regression model may underestimate the true standard errors. Without correcting for non-linearity in response-predictor relationship, one may erroniously conclude that regression parameters are significant.

c. Create a Box Cox plot for the profile log-Likelihoods. How does this plot aid in your data transformation?

```
boxcox(base_model)
```

```
## Error in boxcox(base_model): could not find function "boxcox"
```

Comment: Box Cox provides an analytical way to eveluate constant variance and normality assumptions with estimating significance of lambda. From the plot above, y could be raised to the power close to 1.4.

```
summary(lm(yield**1.4 ~ nitrogen, data = cornnit))
```

```
##
## Call:
## lm(formula = yield^1.4 ~ nitrogen, data = cornnit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -491.35 -110.33   12.33  127.92  279.10
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 710.5988    42.4344  16.746  < 2e-16 ***
## nitrogen      1.6452     0.3071   5.357 3.32e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 186.7 on 42 degrees of freedom
## Multiple R-squared:  0.4059, Adjusted R-squared:  0.3918
## F-statistic:  28.7 on 1 and 42 DF,  p-value: 3.32e-06
```

Comment: this transformation does not seem to improve R-squared.

d. Perform the necessary transformation to the data. Re fit the regression with the transformed variable(s) and assess the regression assumptions. You may have to apply transformations a number of times. Be sure to explain the reason behind each of your transformations. Perform the needed transformations until the regression assumptions are met. What is the regression equation that you will use?

Note: in part 2d, there are a number of solutions that will work. You must clearly document your reasons for each of your transformations

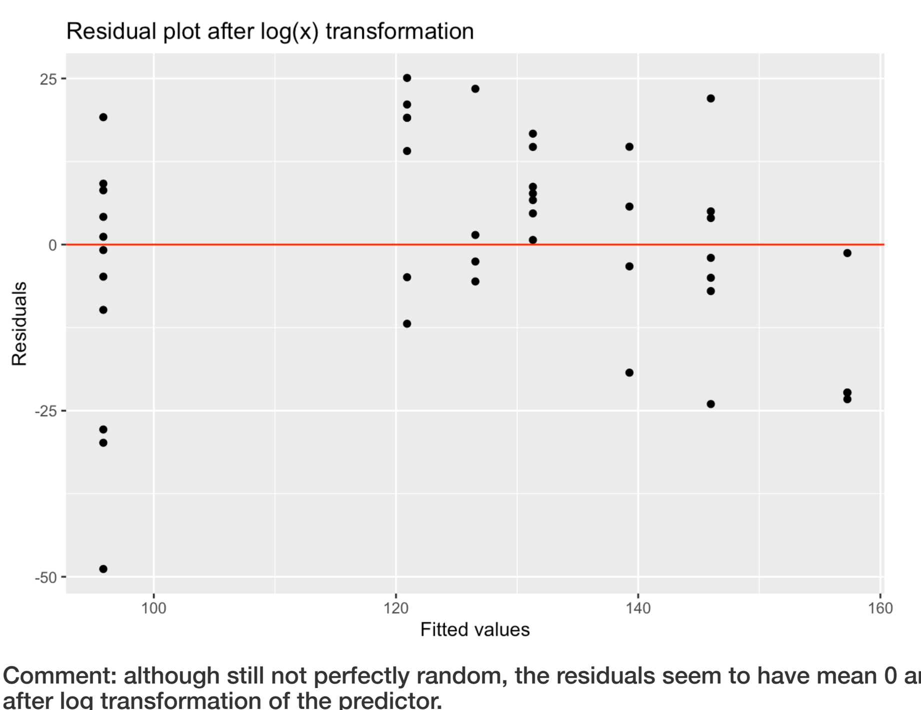
Comment: applying sqrt transformation to predictor to correct for non-zero means of residuals and also cope with zero values of individual observations:

```
x_star<-sqrt(cornnit$nitrogen)
cornnit<-data.frame(cornnit,x_star)

x_star_model <- lm(yield ~ x_star, data=cornnit)
summary(x_star_model)
```

```
##
## Call:
## lm(formula = yield ~ x_star, data = cornnit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.827  -5.912   1.311  10.401  25.087
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   95.827      4.474  21.416  < 2e-16 ***
## x_star         3.548      0.440   8.063 4.58e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.55 on 42 degrees of freedom
## Multiple R-squared:  0.6075, Adjusted R-squared:  0.5981
## F-statistic:    65 on 1 and 42 DF,  p-value: 4.577e-10
```

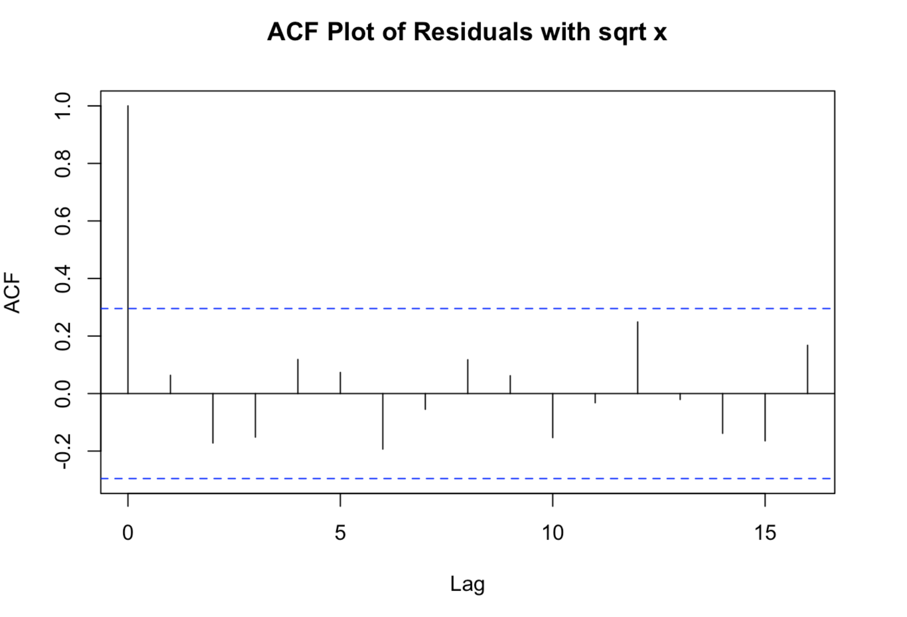
Comment: `sqrt()` transformation to the predictor, produces significant F-statistics and much better R-squared.

```
y_fitted_x_star <- x_star_model$fitted.values
residuals_x_star <- x_star_model$residuals
cornnit <- data.frame(cornnit, y_fitted_x_star, residuals_x_star)

ggplot(cornnit, aes(x=y_fitted_x_star, y=residuals_x_star))+
  geom_point()+
  geom_hline(yintercept = 0, color='red')+
  labs(x="Fitted values", y="Residuals", title= "Residual plot after log(x) transformation")
```



Comment: although still not perfectly random, the residuals seem to have mean 0 and constant variance, after log transformation of the predictor.

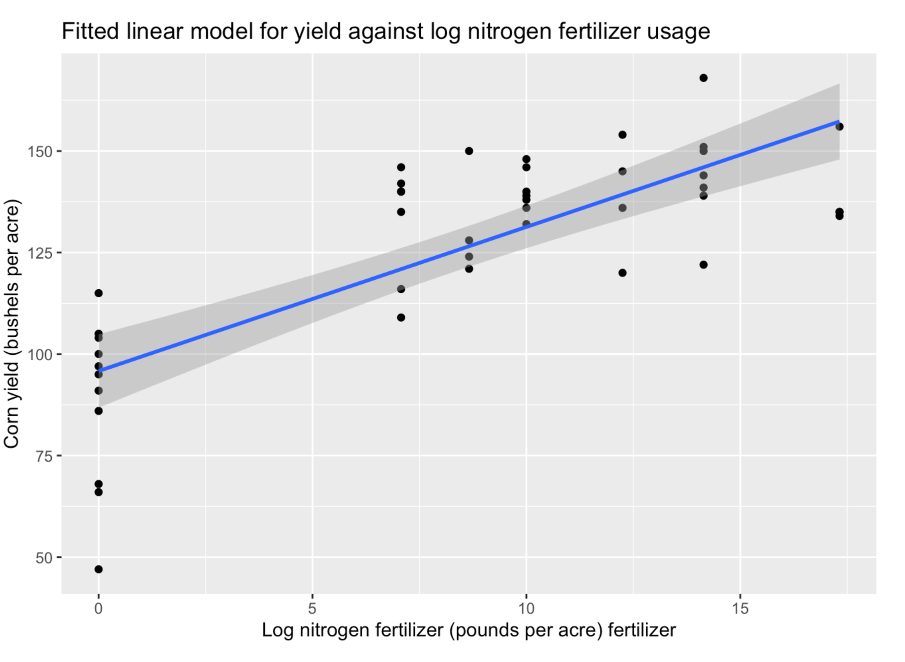
```
acf(residuals_x_star, main="ACF Plot of Residuals with sqrt x")
```



Comment: numerous auto-correlations that were due to uneven sampling of nitrogen usage also was mitigated by this transformation.

```
ggplot(cornnit, aes(x=x_star, y=yield))+
  geom_point()+
  geom_smooth(method = "lm", se = TRUE)+
  labs(x="Log nitrogen fertilizer (pounds per acre) fertilizer",
       y="Corn yield (bushels per acre)",
       title="Fitted linear model for yield against log nitrogen fertilizer usage")
```

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



Question 3

a. Based only on Figure 1, would you recommend transforming the predictor, x, or the response, y, first? Briefly explain your choice.

Comment: it seems there are both problems present - non-zero mean and non-constant variance. I would recommend transforming the response first as it should help with non-constant mean and also could improve non-zero mean. If mean is not zero after this, the predictor can be transformed next.

b. The profile log-likelihoods for the parameter, λ, of the Box-Cox power transformation, is shown in Figure 2. Your classmate says that you should apply a log transformation to the response variable first. Do you agree with your classmate? Be sure to justify your answer.

Comment: yes, I agree. This should work well because 0 is in the range of appropriate values given 95% confidence requirement.

c. Your classmate is adament on applying the log transformation to the response variable, and fits the regression model. The R output is shown in Figure 3. Write down the estimated regression equation for this model. How do we interpret the regression coefficients β^1 and β^0 in context?

Model: log(conc) = 1.5 - 0.45 * (time)

Comment: since only response is log-transformed, the interpretation should state that the predicted value is multiplied by a factor of exp(B1_hat) when predictor increases by one unit of time.