Stat 6021: HW Set 5

Tom Lever

09/29/22

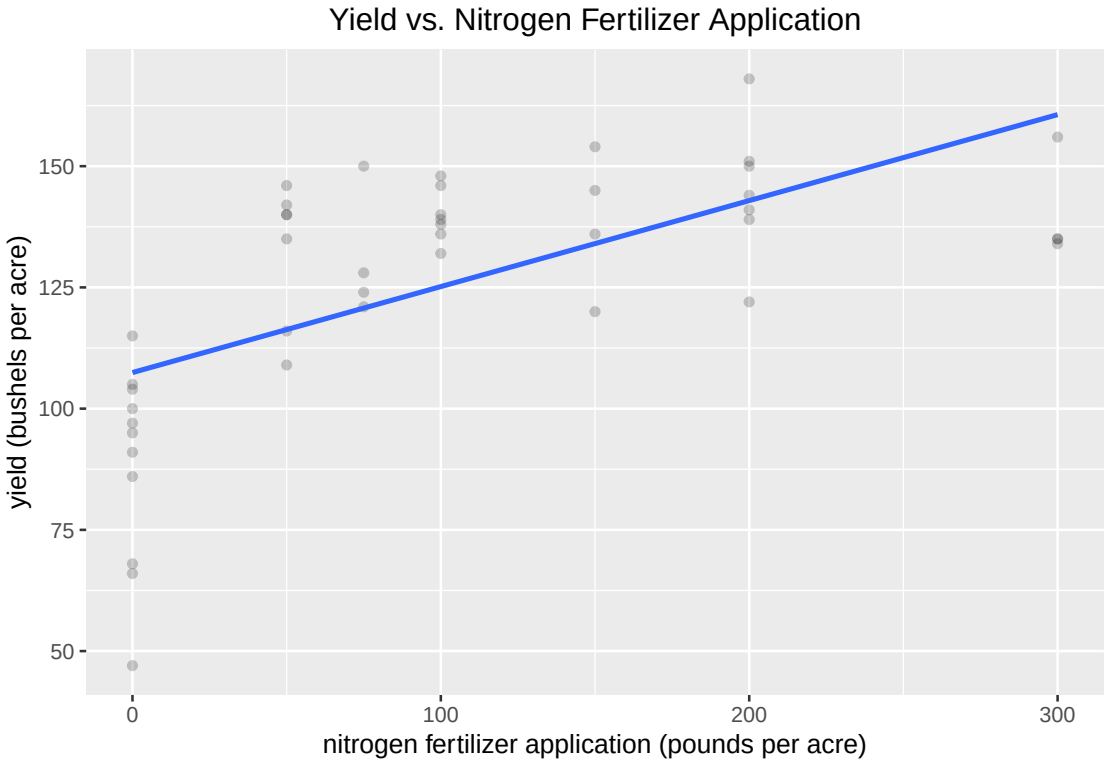
1. For this question, we will use the `cornnit` data set from the `faraway` package. Be sure to install and load the `faraway` package first, and then load the data set. The data explore the relationship between corn yields in bushels per acre and nitrogen fertilizer applications in pounds per acre in a study carried out in Wisconsin.

```
library(faraway)
data_set <- faraway::cornnit
```

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

   The response variable and predictor variable for this study are corn yield in bushels per acre and nitrogen fertilizer applications in pounds per acre, respectively.

```
library(ggplot2)
ggplot(data_set, aes(x = nitrogen, y = yield)) +
    geom_point(alpha = 0.2) +
    geom_smooth(method = "lm", se = FALSE) +
    labs(
        x = "nitrogen fertilizer application (pounds per acre)",
        y = "yield (bushels per acre)",
        title = "Yield vs. Nitrogen Fertilizer Application"
    ) +
    theme(
        plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 0)
    )
```



Generally, there appears to be an increasing association between nitrogen fertilizer applications and yields. The relationship between response / yield y and predictor / regressor / nitrogen fertilizer application x appears nonlinear, perhaps logarithmic.

Assumptions for simple linear regression appear to be not met.

1. The assumption that the relationship between response / yield y and predictor / regressor / nitrogen fertilizer application x is linear, at least approximately, is not met. The relationship appears to be nonlinear.
2. The assumption that the error term ϵ of the linear model has mean 0 is not met. Observations are not scattered evenly around the fitted line.
3. The assumption that the error term ϵ of the linear model has constant variance is not met. The vertical variation of observations is not constant.

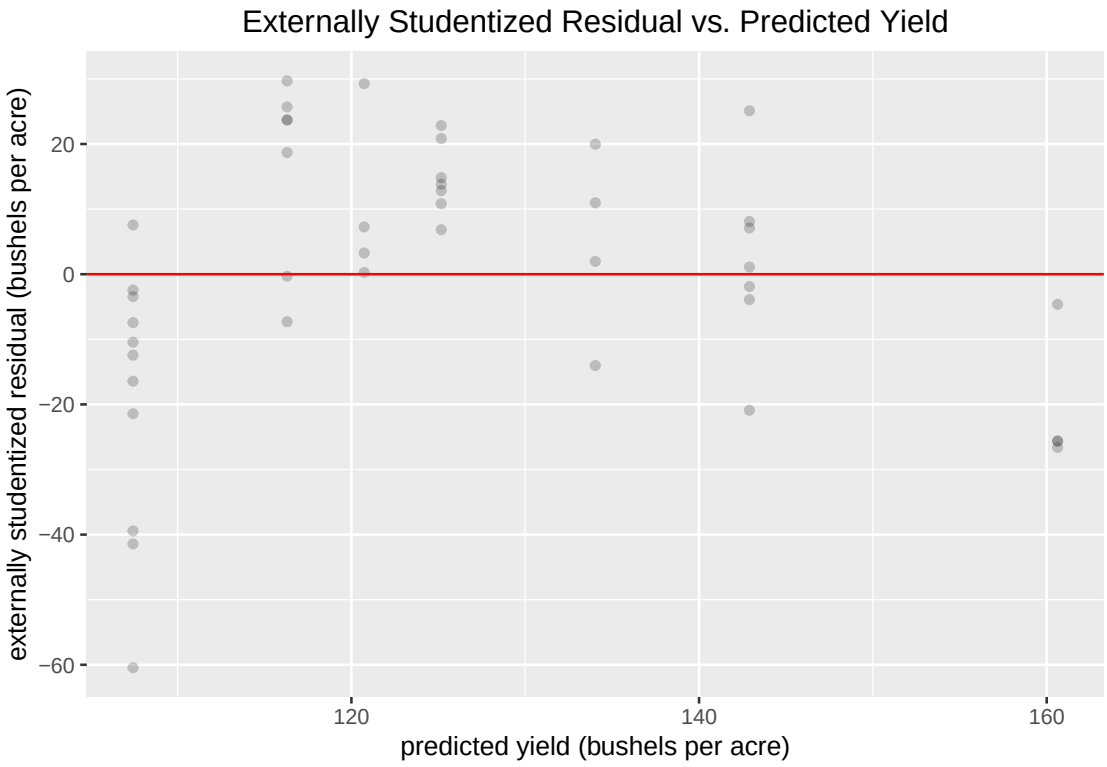
(b) Fit a linear regression without any transformations. Create the corresponding residual plot. Based only on the residual plot, what transformation will you consider first? Be sure to explain your reason.

```
library(TomLeversRPackage)
linear_model_of_yield_and_application <- lm(yield ~ nitrogen, data = data_set)
print(summarize_linear_model(linear_model_of_yield_and_application))
```

```
##
## Call:
## lm(formula = yield ~ nitrogen, data = data_set)
##
## 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
##
## E(y | x) = B_0 + B_1 * x = 107.438643702906 + 0.177296632323543 * x
## Number of observations: 44
## Estimated variance of errors: 421.409192094491
## Multiple R:  0.629456189046164   Adjusted R:  0.617931438617246
```

```
ggplot(
    data.frame(
        externally_studentized_residual =
            linear_model_of_yield_and_application$residuals,
        predicted_yield = linear_model_of_yield_and_application$fitted.values
    ),
    aes(x = predicted_yield, y = externally_studentized_residual)
) +
    geom_point(alpha = 0.2) +
    geom_hline(yintercept = 0, color = "red") +
    labs(
        x = "predicted yield (bushels per acre)",
        y = "externally studentized residual (bushels per acre)",
        title = "Externally Studentized Residual vs. Predicted Yield"
    ) +
theme(
    plot.title = element_text(hjust = 0.5),
    axis.text.x = element_text(angle = 0)
)
```



Assumptions for simple linear regression appear to be not met.

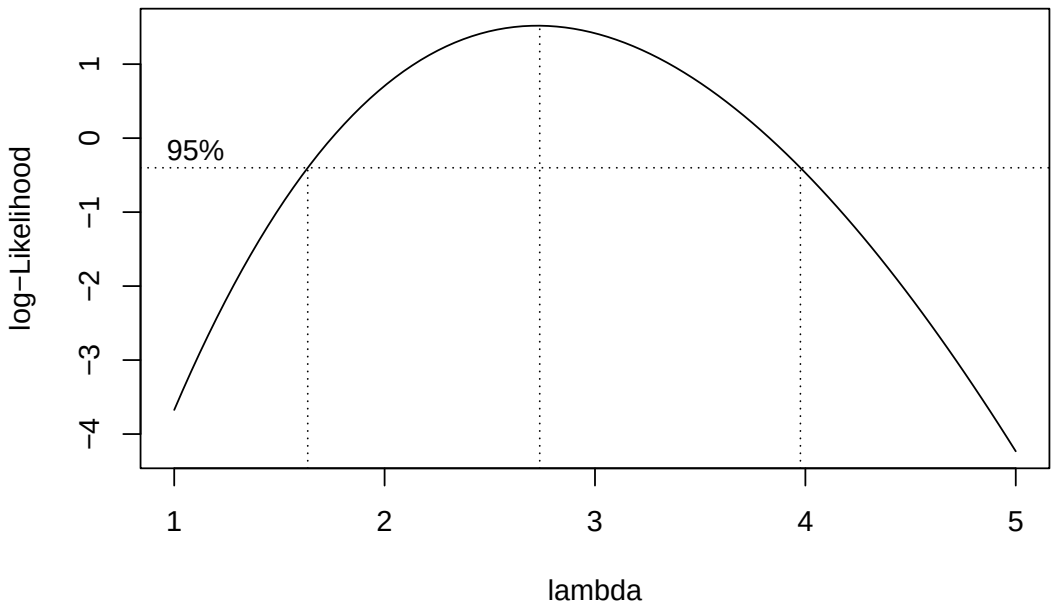
1. The assumption that the relationship between response / yield y and predictor / regressor / nitrogen fertilizer application x is linear, at least approximately, is not met; residuals follow a curve.
2. The assumption that the error term ϵ of the linear model of yield and application has mean 0 is not met. Residuals are not evenly scattered around $e = 0$.
3. The assumption that the error term ϵ of the linear model of yield and application has constant variance is not met. Residuals do not have similar vertical variation across $e = 0$.

Since assumption 3, that the error term ϵ of the linear model of yield and application has constant variance, is not met, we transform the response / yield. Since we are correcting for the error term ϵ's non-constant variance, we use a power transformation y^λ, where λ is a parameter determined below by the Box-Cox Method.

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

A Box-Cox plot suggests a maximum-likelihood estimate of parameter λ around 3, not 1.

```
result_of_Box_Cox_Method <- perform_Box_Cox_Method(
    linear_model_of_yield_and_application,
    vector_of_values_of_lambda = seq(1, 5, 0.1)
)
```



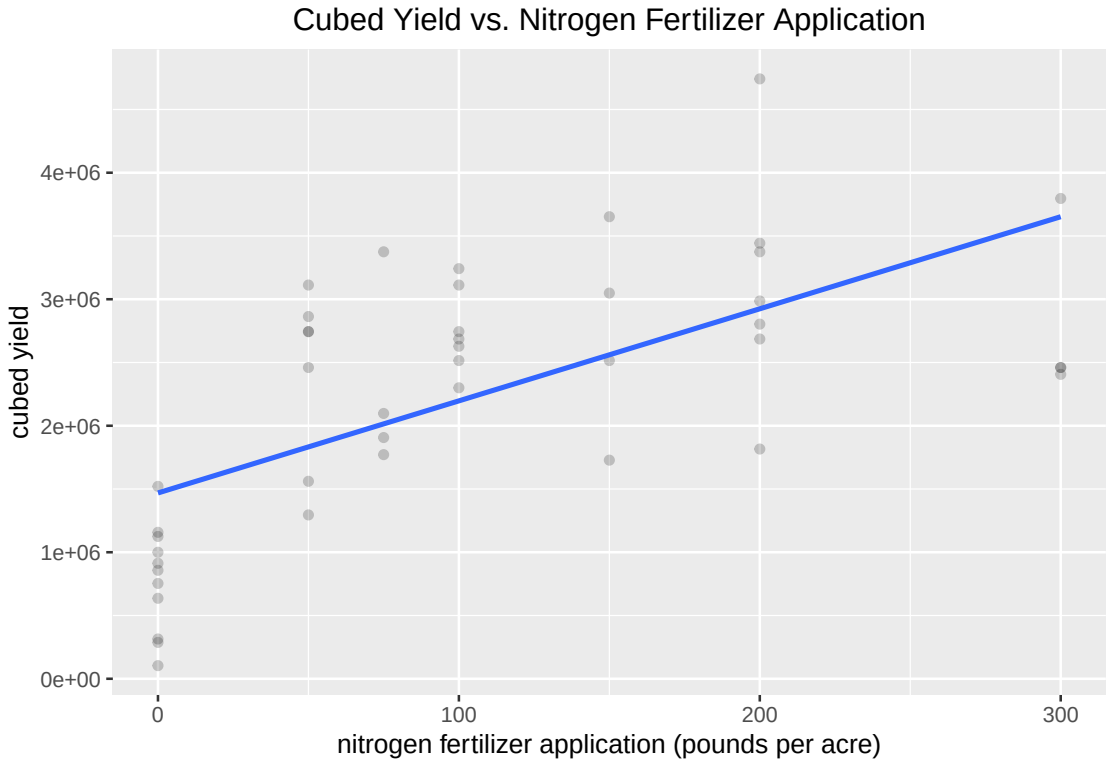

```
print(result_of_Box_Cox_Method$maximum_likelihood_estimate_of_parameter_lambda)
```

```
## [1] 2.737374
```

For the set of observations of yield and nitrogen fertilizer application, the maximum-likelihood estimate of λ is close to a whole, within-confidence-interval parameter $\lambda = 3$. Since parameter $\lambda = 3 \neq 0$, we use the transformation $y' = y^{\lambda} = y^3$.

(d) Perform the necessary transformation of the data. Refit the regression with the transformed variable 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.

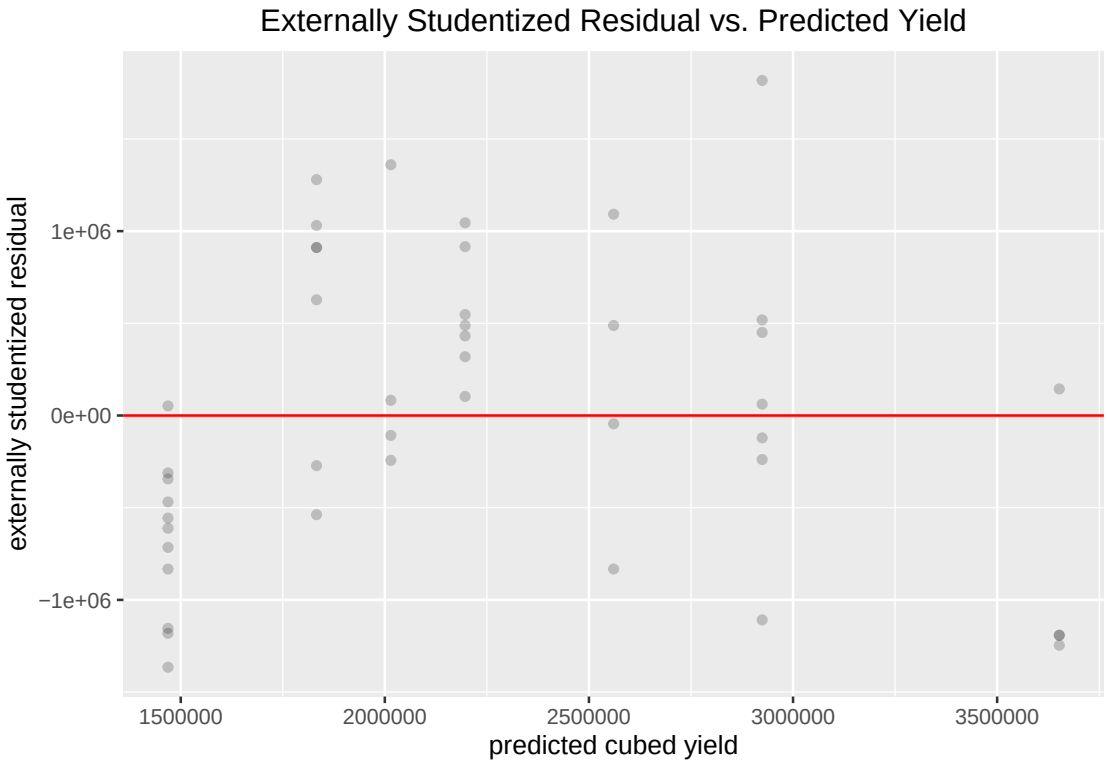
```
library(dplyr)
data_set <- data_set %>% mutate(cubed_yield = yield^3)
ggplot(data_set, aes(x = nitrogen, y = cubed_yield)) +
    geom_point(alpha = 0.2) +
    geom_smooth(method = "lm", se = FALSE) +
    labs(
        x = "nitrogen fertilizer application (pounds per acre)",
        y = "cubed yield",
        title = "Cubed Yield vs. Nitrogen Fertilizer Application"
    ) +
theme(
    plot.title = element_text(hjust = 0.5),
    axis.text.x = element_text(angle = 0)
)
```




```
linear_model_of_cubed_yield_and_application <-
    lm(cubed_yield ~ nitrogen, data = data_set)
print(summarize_linear_model(linear_model_of_cubed_yield_and_application))
```

```
##
## Call:
## lm(formula = cubed_yield ~ nitrogen, data = data_set)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1365038  -570012     3471   525741  1817216
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1468861     186746   7.866 8.63e-10 ***
## nitrogen        7278       1352   5.385 3.03e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 821600 on 42 degrees of freedom
## Multiple R-squared:  0.4084, Adjusted R-squared:  0.3943
## F-statistic: 28.99 on 1 and 42 DF,  p-value: 3.029e-06
##
## E(y | x) = B_0 + B_1 * x = 1468860.92088267 + 7277.77658926649 * x
## Number of observations: 44
## Estimated variance of errors: 674960731055.44
## Multiple R:  0.639067540671304   Adjusted R:  0.627950461084573
```

```
ggplot(
    data.frame(
        externally_studentized_residual =
            linear_model_of_cubed_yield_and_application$residual,
        predicted_cubed_yield =
            linear_model_of_cubed_yield_and_application$fitted.values
    ),
    aes(x = predicted_cubed_yield, y = externally_studentized_residual)
) +
    geom_point(alpha = 0.2) +
    geom_hline(yintercept = 0, color = "red") +
    labs(
        x = "predicted cubed yield",
        y = "externally studentized residual",
        title = "Externally Studentized Residual vs. Predicted Yield"
    ) +
theme(
    plot.title = element_text(hjust = 0.5),
    axis.text.x = element_text(angle = 0)
)
```



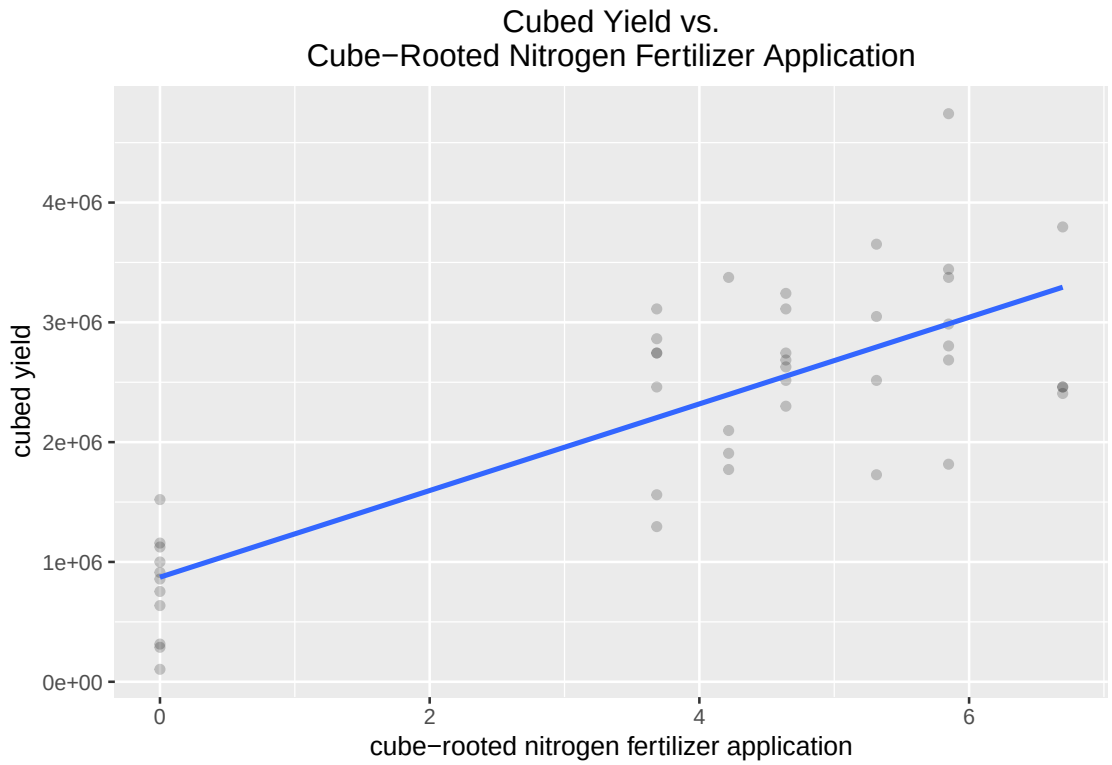
Assumptions for simple linear regression appear to be not met.

1. The assumption that the relationship between response / yield y and predictor / regressor / nitrogen fertilizer application x is linear, at least approximately, is not met. The relationship appears to be nonlinear.
2. The assumption that the error term ϵ of the linear model of cubed yield and application has mean 0 is not met. Observations are not scattered evenly around the fitted line. Residuals are not evenly scattered aroun $e = 0$.

3. The assumption that the error term ϵ of the linear model of cubed yield and application has constant variance is not met. The vertical variation of observations is not constant. Residuals are not evenly scattered aroun $e = 0$.

Because a cube-root function roughly fits the above scatterplot of cubed yields vs. nitrogen fertilizer applications, we transform the predictor / nitrogen fertilizer application x.

```
library(dplyr)
data_set <- data_set %>% mutate(cube_rooted_application = nitrogen^(1/3))
ggplot(data_set, aes(x = cube_rooted_application, y = cubed_yield)) +
    geom_point(alpha = 0.2) +
    geom_smooth(method = "lm", se = FALSE) +
    labs(
        x = "cube-rooted nitrogen fertilizer application",
        y = "cubed yield",
        title = paste(
            "Cubed Yield vs.\n",
            "Cube-Rooted Nitrogen Fertilizer Application",
            sep = ""
        )
    ) +
theme(
    plot.title = element_text(hjust = 0.5),
    axis.text.x = element_text(angle = 0)
)
```

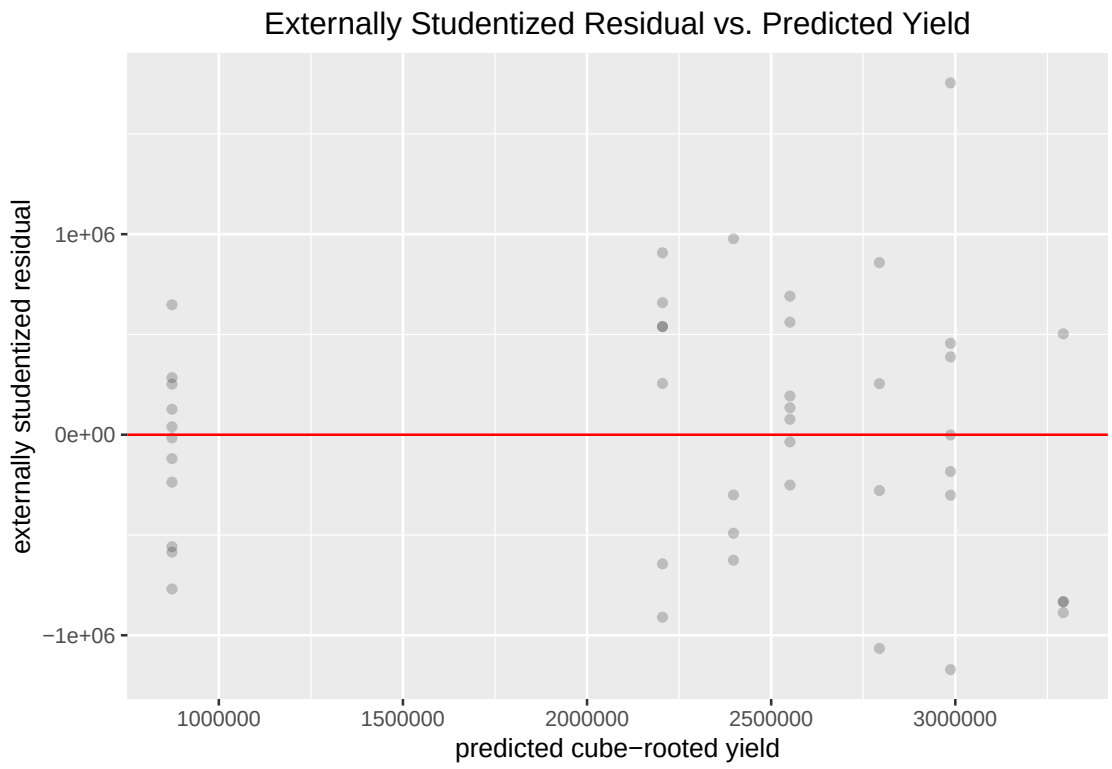



```
linear_model_with_cubed_yield_and_cubed_rooted_application <-
    lm(cubed_yield ~ cube_rooted_application, data = data_set)
print(
    summarize_linear_model(
```

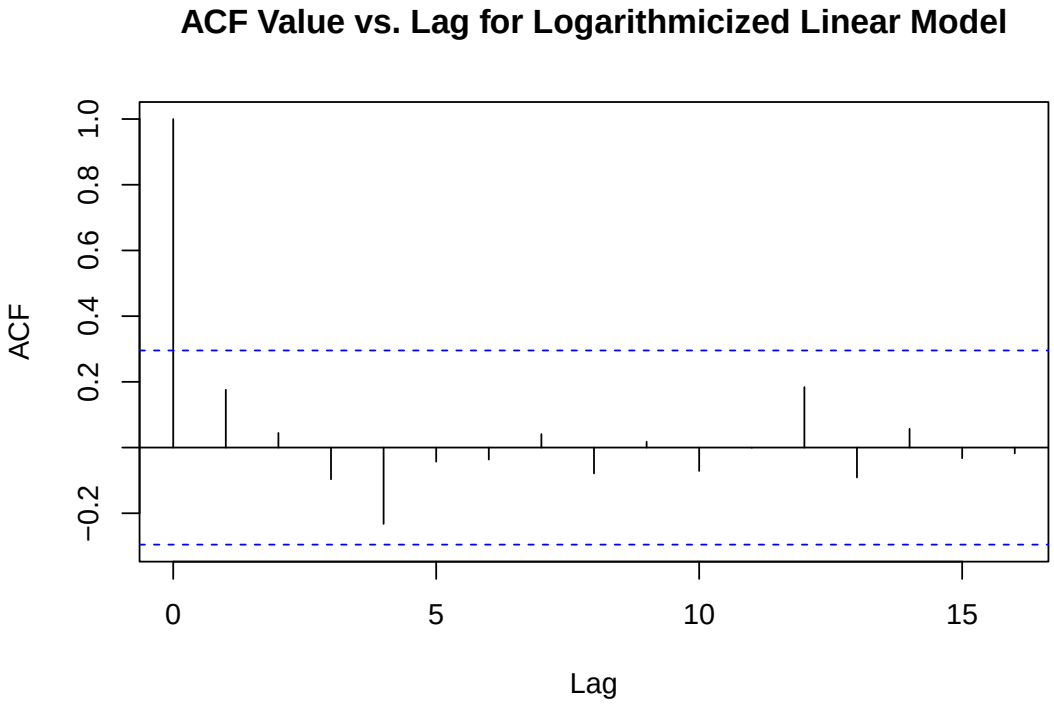
```
        linear_model_with_cubed_yield_and_cubed_rooted_application
    )
)
```

```
##
## Call:
## lm(formula = cubed_yield ~ cube_rooted_application, data = data_set)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1171423  -507871    19240   467551  1754361
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                872906     181799   4.801 2.02e-05 ***
## cube_rooted_application    361552      41398   8.733 5.44e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 636500 on 42 degrees of freedom
## Multiple R-squared:  0.6449, Adjusted R-squared:  0.6364
## F-statistic: 76.27 on 1 and 42 DF,  p-value: 5.436e-11
##
## E(y | x) = B_0 + B_1 * x = 872905.511650767 + 361551.460218236 * x
## Number of observations: 44
## Estimated variance of errors: 405152132237.357
## Multiple R:  0.803050683448782   Adjusted R:  0.79776902027832
```

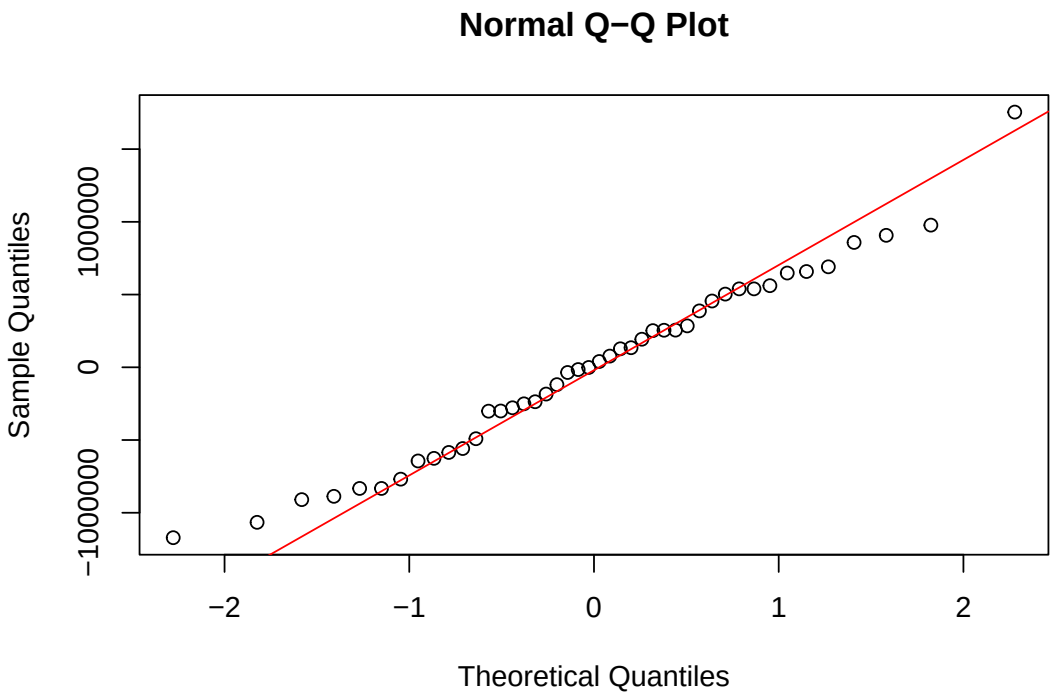
```
ggplot(
    data.frame(
        externally_studentized_residual =
            linear_model_with_cubed_yield_and_cubed_rooted_application$residuals,
        predicted_cube_rooted_yield =
            linear_model_with_cubed_yield_and_cubed_rooted_application$
                fitted.values
    ),
    aes(x = predicted_cube_rooted_yield, y = externally_studentized_residual)
) +
    geom_point(alpha = 0.2) +
    geom_hline(yintercept = 0, color = "red") +
    labs(
        x = "predicted cube-rooted yield",
        y = "externally studentized residual",
        title = "Externally Studentized Residual vs. Predicted Yield"
    ) +
theme(
    plot.title = element_text(hjust = 0.5),
    axis.text.x = element_text(angle = 0)
)
```




```
acf(
    linear_model_with_cubed_yield_and_cubed_rooted_application$residuals,
    main = "ACF Value vs. Lag for Logarithmicized Linear Model"
)
```



```
qqnorm(linear_model_with_cubed_yield_and_cubed_rooted_application$residuals)
qqline(linear_model_with_cubed_yield_and_cubed_rooted_application$residuals, col = "red")
```



Some assumptions for simple linear regression appear to be met.

1. The assumption that the relationship between response / yield y and predictor / regressor / nitrogen fertilizer application x is linear, at least approximately, is met. The relationship appears to be linear.
2. The assumption that the error term ϵ of the linear model of cubed yield and cube-rooted application has mean 0 is met. Observations are scattered evenly around the fitted line. Residuals are evenly scattered aroun $e = 0$.
3. The assumption that the error term ϵ of the linear model of cubed yield and cube-rooted application has constant variance is met. The vertical variation of observations is constant. Residuals are evenly scattered aroun $e = 0$.
4. The assumption that the errors ϵ_i / residuals e_i are uncorrelated is met. The ACF value for lag 0 is always 1; the correlation of the vector of residuals with itself is always 1. Since all ACF values are insignificant, we have insufficient evidence to reject a null hypothesis that the residuals of the linear model of cubed yield and cube-rooted application are uncorrelated. We have insufficient evidence to conclude that the residuals of the linear model of cubed yield and cube-rooted application are correlated. We have insufficient evidence to conclude that the assumption that the errors ϵ_i / residuals e_i are uncorrelated is not met.
5. Assumptions that the errors ϵ_i / residuals e_i are normally distributed is not met. A linear model is robust to these assumptions. Given moderate flattening at extremes of a plot of sample quantiles versus theoretical quantiles for the residuals of the linear model of cubed yield and cube-rooted application, the tails of the probability vs. externally studentized residuals plot / distribution are too heavy for this distribution to be considered normal. The assumption that the residuals are normally distributed is not met.

2. A chemist studied the concentration of a solution y over time x by fitting a simple linear regression. The scatterplot of the dataset and the residual plot from the regression model are shown in the prompt

for this homework.

(a) Based on these two scatterplots, would you recommend transforming the predictor x or the response y first?

The assumption that the error term ϵ of the linear model has constant variance is not met. Residuals do not have similar vertical variation across $e = 0$.

Since the assumption that the error term ϵ of the linear model has constant variance is not met, we transform the response / concentration first. Transforming the response / concentration may also influence the linear model's compliance with an assumption that the mean of errors is 0. After transforming the response / concentration, if ensuring the linear model's compliance with an assumption that the mean of errors is 0 requires a transformation, transform the predictor / time.

(b) The profile log-likelihoods for the parameter λ of the Box-Cox power transformation is shown in the prompt for this homework. 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.

Yes. For the set of observations of concentrations and times, the maximum-likelihood estimate of λ is close to a whole, within-confidence-interval parameter $\lambda = 0$. Since parameter $\lambda = 0$, we use the transformation $y' = ln\,(y)$.

(c) Your classmate is adamant on applying the log transformation to the response variable and fits the regression model. The R output is shown in the prompt for this homework. Write down the estimated regression equation for this model. How do we interpret the regression coefficients $\hat{\beta}_1$ and $\hat{\beta}_0$ in context?

$$\hat{\beta}_0 = 1.50792$$

$$\hat{\beta}_1 = -0.44993 \ \frac{1}{time\ unit}$$

$$ln\,(\hat{y}) = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\hat{y} = exp\left(\hat{\beta}_0 + \hat{\beta}_1 x\right)$$

$$\hat{y}_+ = exp\left(\hat{\beta}_0 + \hat{\beta}_1 (x+1)\right)$$

$$\hat{y}_+ = exp\left(\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_1\right)$$

$$\hat{y}_+ = exp\left(\hat{\beta}_0 + \hat{\beta}_1 x\right) exp\left(\hat{\beta}_1\right)$$

$$\hat{y}_+ = \hat{y}\ exp\left(\hat{\beta}_1\right)$$

When predictor variable / time x increases by one unit, the predicted response / concentration $\hat{y}$ increases by a factor of $exp\left(\hat{\beta}_1\right)$.