Stat 6021: Addressing Guided Question Set for Module 5: Model Diagnostics and Remedial Measures in SLR

Tom Lever

09/24/22

The data set `mammals` from the `MASS` package contains the average brain and body weights for 62 species of land mammals. We wish to see how body weight x could explain the brain weight y of land mammals.

1. Create a scatter plot of brain weight against body weight of land mammals. Comment on the appearance of the plot. Do any assumptions for simple linear regression appear to be violated? If so, which ones?

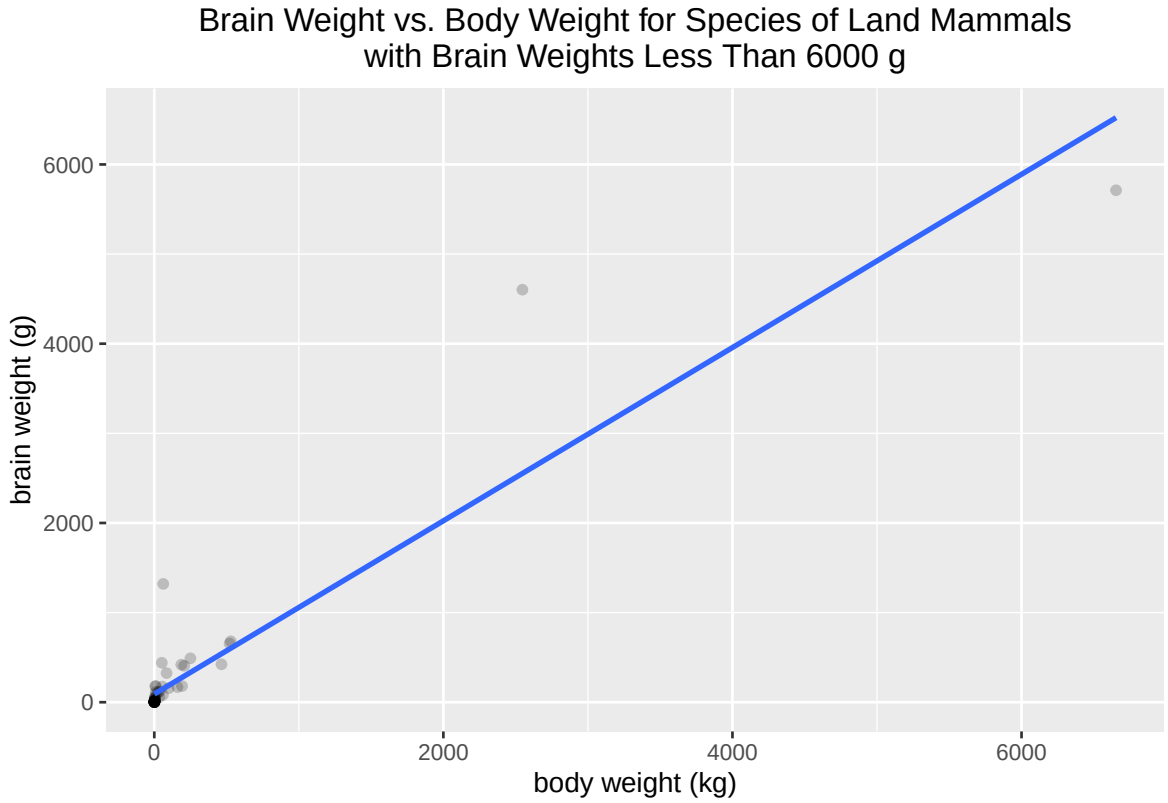
```
library(MASS)
library(dplyr)
library(Dict)
library(ggplot2)
library(TomLeversRPackage)

species_body_weight_and_brain_weight <-
    MASS::mammals %>% rename(body_weight = body, brain_weight = brain)
head(species_body_weight_and_brain_weight, n = 3)
```

```
##                 body_weight brain_weight
## Arctic fox            3.385         44.5
## Owl monkey            0.480         15.5
## Mountain beaver       1.350          8.1
```

```
brain_weight_thresholds_and_data_subsets <- dict(
    '6000 g' = species_body_weight_and_brain_weight %>% filter(brain_weight < 6000),
    '1000 g' = species_body_weight_and_brain_weight %>% filter(brain_weight < 1000),
    '100 g' = species_body_weight_and_brain_weight %>% filter(brain_weight < 100),
    '25 g' = species_body_weight_and_brain_weight %>% filter(brain_weight < 25),
    '5 g' = species_body_weight_and_brain_weight %>% filter(brain_weight < 5)
)

#for (key in brain_weight_thresholds_and_data_subsets$keys) {
for (key in c("6000 g")) {
    data_set <- brain_weight_thresholds_and_data_subsets$get(key)
    plot(
        ggplot(data_set, aes(x = body_weight, y = brain_weight)) +
            geom_point(alpha = 0.2) +
            geom_smooth(method = "lm", se = FALSE) +
            labs(
                x = "body weight (kg)",
                y = "brain weight (g)",
                title = paste(
                    "Brain Weight vs. Body Weight for Species of Land Mammals\n",
                    "with Brain Weights Less Than ",
```

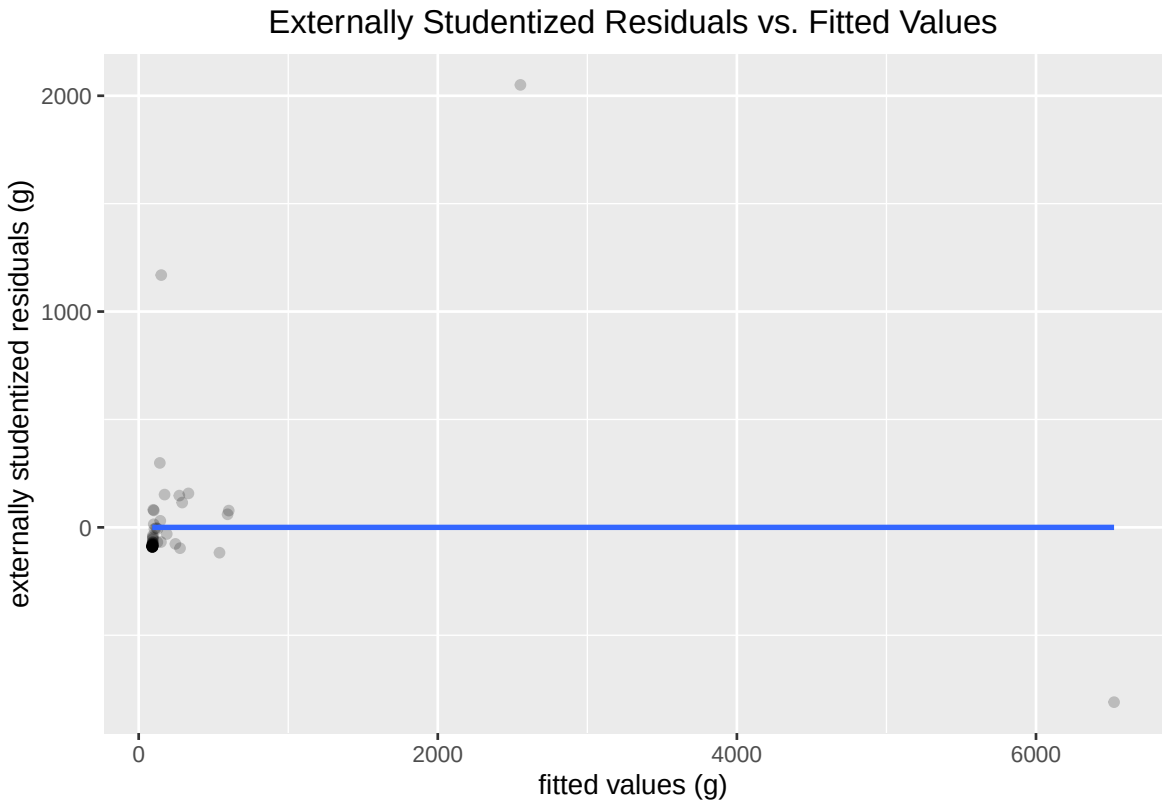
```
                key,
                sep = ""
            )
        ) +
    theme(
        plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 0)
    )
)

linear_model <- lm(brain_weight ~ body_weight, data = data_set)
print(summarize_linear_model(linear_model))

plot(
    ggplot(
        data.frame(
            externally_studentized_residuals = linear_model$residuals,
            fitted_values = linear_model$fitted.values
        ),
        aes(x = fitted_values, y = externally_studentized_residuals)
    ) +
        geom_point(alpha = 0.2) +
        geom_smooth(method = "lm", se = FALSE) +
        labs(
            x = "fitted values (g)",
            y = "externally studentized residuals (g)",
            title = paste(
                "Externally Studentized Residuals vs. Fitted Values",
                sep = ""
            )
        ) +
    theme(
        plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 0)
    )
)

result_of_Box_Cox_Method <- perform_Box_Cox_Method(linear_model)
print(result_of_Box_Cox_Method$maximum_likelihood_estimate_of_parameter_lambda)

data_set <- data_set %>% mutate(brain_weight_subject_to_Box_Cox_Method = result_of_Box_Cox_Meth
plot(
    ggplot(data_set, aes(x = body_weight, y = brain_weight_subject_to_Box_Cox_Method)) +
        geom_point(alpha = 0.2) +
        geom_smooth(method = "lm", se = FALSE) +
        labs(
            x = "body weight (kg)",
            y = "logarithmicized brain weight (g)",
            title = paste(
                "Logarithmicized Brain Weight vs. Body Weight for Species of Land Mammals\n",
                "with Brain Weights Less Than ",
                key,
                sep = ""
```

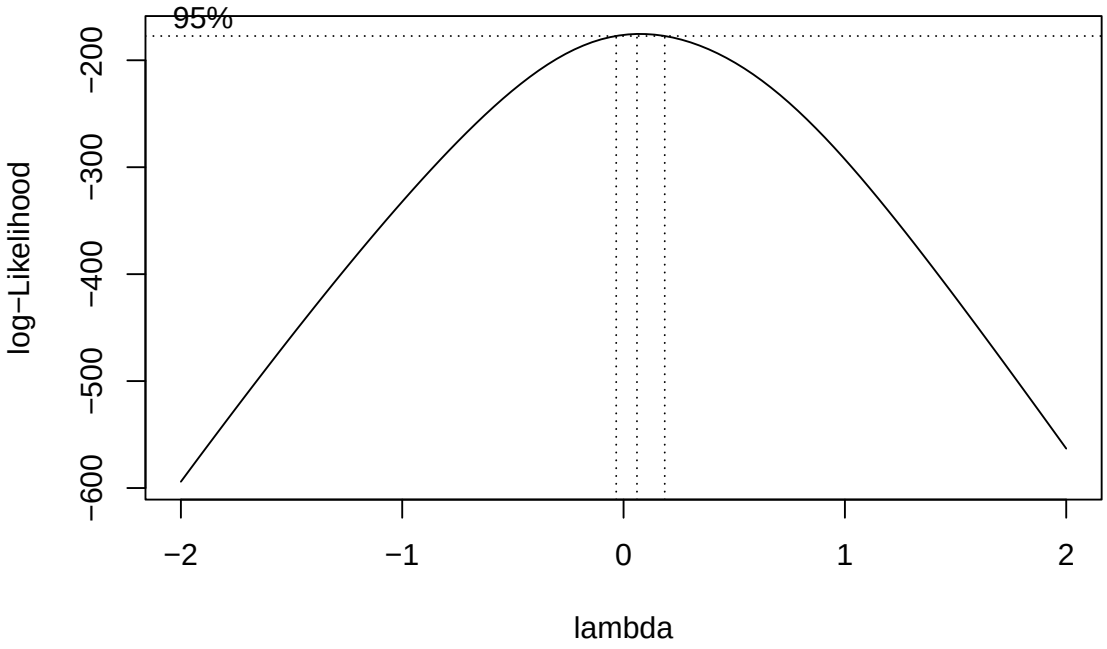
```
                )
            ) +
        theme(
            plot.title = element_text(hjust = 0.5),
            axis.text.x = element_text(angle = 0)
        )
    )
}
```



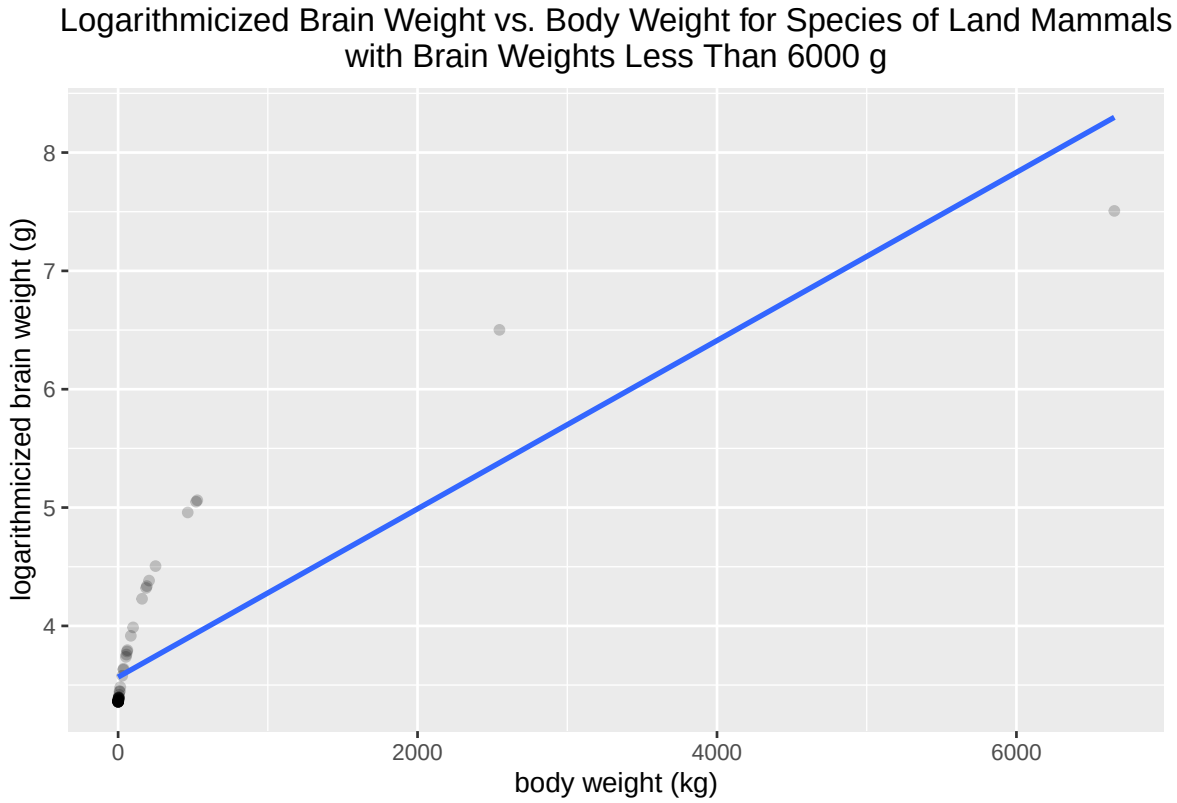

```
##
## Call:
## lm(formula = brain_weight ~ body_weight, data = data_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -810.07  -88.52  -79.64  -13.02 2050.33
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 91.00440   43.55258    2.09   0.0409 *
## body_weight  0.96650    0.04766   20.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 334.7 on 60 degrees of freedom
## Multiple R-squared:  0.8727, Adjusted R-squared:  0.8705
## F-statistic: 411.2 on 1 and 60 DF,  p-value: < 2.2e-16
```

```
##
## E(y | x) = B_0 + B_1 * x = 91.00440 + 0.96650 * x
## Number of observations: 62
## Estimated variance of errors: 112024.09
## Multiple R:  0.934163842323355   Adjusted R:  0.933005894943864
```




95%
log−Likelihood
−200
−300
−400
−500
−600
−2
−1
0
1
2
lambda

```
## [1] 0.06060606
```


Logarithmicized Brain Weight vs. Body Weight for Species of Land Mammals
with Brain Weights Less Than 6000 g
logarithmicized brain weight (g)
8
7
6
5
4
0
2000
4000
6000
body weight (kg)

For sets of observations of brain weight and body weight with maximum brain weights of 6000 g, 1000 g, 100 g, 25 g, and 5 g, taken individually and collectively, the relationship(s) between response $y = brain\ weight$ and predictor / regressor $x = body\ weight$ appears to be nonlinear / logarithmic.

Assumptions for simple linear regression appear to be violated for sets of observations, taken individually and collectively.

a. Assumptions that the relationship between response $y = brainweight$ and predictor / regressor $x = bodyweight$ is linear, at least approximately, are violated. The relationship(s) appears to be nonlinear / logarithmic. Residuals are not evenly scattered aroun $e = 0$.
b. Assumptions that the error term ϵ of the linear model has mean 0 are violated. Observations are not scattered evenly around the fitted line(s). Residuals are not evenly scattered around $e = 0$.
c. Assumptions that the error term ϵ of the linear model has constant variance are violated. The vertical variation of observations is not constant. Residuals do not have similar vertical variation across $e = 0$.
d. Assumptions that the errors ϵ_i / residuals e_i are uncorrelated are violated. ?
e. Assumptions that the errors ϵ_i / residuals e_i are normally distributed are violated. ?

2. Fit a simple linear regression to the data, and create the corresponding residual plot. Do any assumptions for simple linear regression appear to be violated? If so, which ones?

   See above.

3. Based on your answers to parts 1 and 2, do we need to transform at least one of the variables?

   Yes. Because assumptions that the error term ϵ of the linear model has constant variance are violated, we will transform response $\hat{y} = estimated\ brain\ weight$. Because assumptions that the error term ϵ of the linear model has mean 0 are violated, we will transform predictor $x = body\ weight$ after transforming response $\hat{y} = estimated\ brain\ weight$ if necessary.

4. For the simple linear regression in part 2, create a Box-Cox plot. What transformation, if any, would you apply to the response variable?

   For the set of observations of brain weight and body weight with maximum brain weights of 6000 g, the maximum-likelihood estimate of λ is close to parameter $\lambda = 0$. Since parameter $\lambda = 0$, we use the transformation $\hat{y}' = ln(\hat{y})$.

5. Apply the transformation you specified in part 4, and let $y*$ denote the transformed response variable. Create a scatterplot of $y*$ against x. Comment on the appearance of the plot. Do any assumptions for simple linear regression appear to be violated?