Stat 6021: Guided Question Set 8

Tom Lever

10/22/22

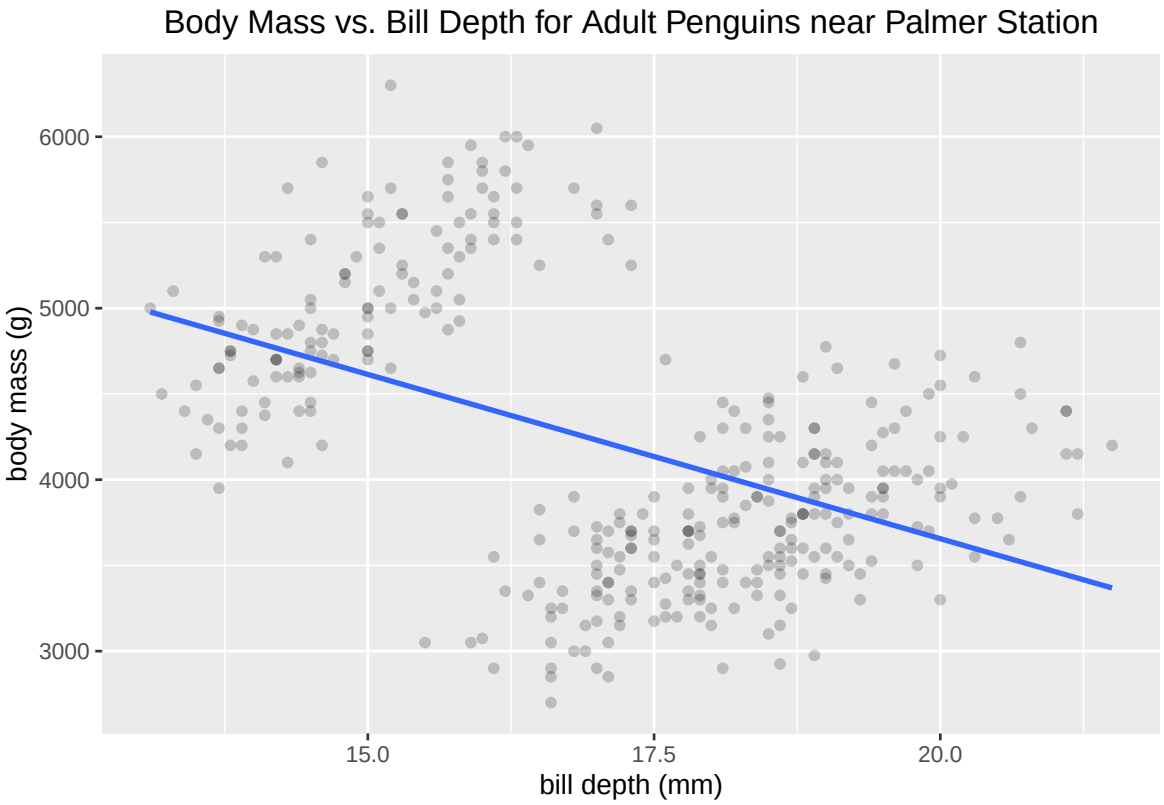
We will revisit the data set `penguins` from the `palmerpenguins` package. The data set contains size measurements for adult foraging penguins near Palmer Station, Antarctica. In this set of questions, we focus on exploring the relationship between body mass y and bill depth x_1 of three species of penguins.

1. Create a scatterplot of the body mass against the bill length of the penguins. How would you describe the relationship between these two variables?

```
library(palmerpenguins)
library(dplyr)
species_bill_depth_and_body_mass <-
    palmerpenguins::penguins %>%
        select(species, bill_depth_mm, body_mass_g) %>%
        filter(!is.na(bill_depth_mm))
head(species_bill_depth_and_body_mass, n = 3)
```

```
## # A tibble: 3 x 3
##   species bill_depth_mm body_mass_g
##   <fct>           <dbl>       <int>
## 1 Adelie           18.7        3750
## 2 Adelie           17.4        3800
## 3 Adelie           18          3250
```

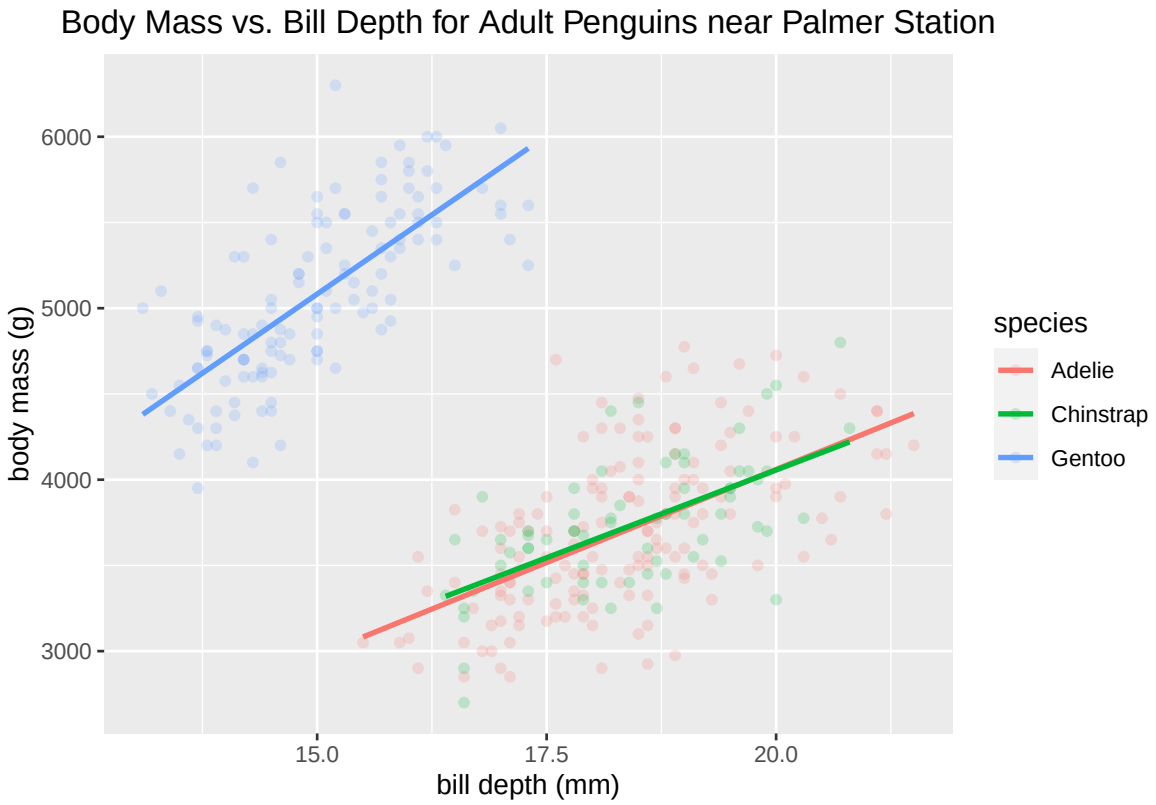
```
library(ggplot2)
ggplot(
    species_bill_depth_and_body_mass,
    aes(x = bill_depth_mm, y = body_mass_g)
) +
    geom_point(alpha = 0.2) +
    geom_smooth(method = "lm", se = FALSE) +
    labs(
        x = "bill depth (mm)",
        y = "body mass (g)",
        title = "Body Mass vs. Bill Depth for Adult Penguins near Palmer Station"
    ) +
    theme(
        plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 0)
    )
```



The relationship between bill depth and body mass for adult penguins near Palmer Station, Antarctica appears clustered, perhaps by species, and nonlinear. A line of best fit has been rendered to aid this determination.

2. Create the same scatterplot but now with differently colored plots for each species. Also be sure to overlay separate regression lines for each species. How would you now describe the relationship between the variables?

```
library(ggplot2)
ggplot(
    species_bill_depth_and_body_mass,
    aes(x = bill_depth_mm, y = body_mass_g, color = species)
) +
    geom_point(alpha = 0.2) +
    geom_smooth(method = "lm", se = FALSE) +
    labs(
        x = "bill depth (mm)",
        y = "body mass (g)",
        title = "Body Mass vs. Bill Depth for Adult Penguins near Palmer Station"
    ) +
    theme(
        plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 0)
    )
```



The relationship between bill depth and body mass for adult penguins near Palmer Station, Antarctica appears linear for each observed species. Lines of best fit have been rendered to aid these determinations. The biases and rates at which body mass increases for increasing bill depth are approximately equal for Adelie and Chinstrap penguins. The bias for Gentoo penguins is substantially greater for Gentoo penguins than for Adelie and Chinstrap penguins. The rate at which body mass increases for increasing bill depth for Gentoo penguins is greater than the rates for Adelie and Chinstrap penguins, indicating that there is an interaction between species and bill depth; the influence of bill depth on body mass may differ between species.

3. Create a multiple linear regression model with interaction between bill depth and species. That is, create a multiple linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 I_1 + \beta_3 I_2 + \beta_4 x_1 I_1 + \beta_5 x_1 I_2 + \epsilon$$

where I_1 and I_2 are indicator variables where $I_1 = 1$ for Chinstrap penguins and 0 otherwise, and $I_2 = 1$ for Gentoo penguins and 0 otherwise. Write down the estimated regression equation for this multiple linear regression model.

Since the categorical variable `species` has three levels, there will be two indicator variables created to represent the various species.

```
library(TomLeversRPackage)
generate_data_frame_for_indicator_variables(species_bill_depth_and_body_mass$species)
```

```
##           I_Chinstrap I_Gentoo
## Adelie              0        0
## Chinstrap           1        0
## Gentoo              0        1
```

A table of indicator variables and their values for each species indicates that the Adelie species is a reference species; both indicator variables take on the value of 0 for species Adelie.

Given the possibility that the rate at which body mass increases for increasing bill depth is greater for Gentoo penguins than for Adelie or Chinstrap penguins, we consider fitting a multiple linear regression model with interaction terms between bill depth and each indicator variable.

```
data_set <- species_bill_depth_and_body_mass
full_model <- lm(body_mass_g ~ bill_depth_mm * species, data = data_set)
summarize_linear_model(full_model)
```

```
##
## Call:
## lm(formula = body_mass_g ~ bill_depth_mm * species, data = data_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -845.89 -254.74  -28.46  228.01 1161.41
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -283.28     437.94  -0.647   0.5182
## bill_depth_mm                    217.15      23.82   9.117   <2e-16 ***
## speciesChinstrap                 247.06     829.77   0.298   0.7661
## speciesGentoo                   -175.71     658.43  -0.267   0.7897
## bill_depth_mm:speciesChinstrap   -12.53      45.01  -0.278   0.7809
## bill_depth_mm:speciesGentoo      152.29      40.49   3.761   0.0002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 354.9 on 336 degrees of freedom
## Multiple R-squared:  0.807,  Adjusted R-squared:  0.8041
## F-statistic:   281 on 5 and 336 DF,  p-value: < 2.2e-16
##
## E(y | x) =
##     B_0 +
##     B_bill_depth_mm * bill_depth_mm +
##     B_speciesChinstrap * speciesChinstrap +
##     B_speciesGentoo * speciesGentoo +
##     B_bill_depth_mm:speciesChinstrap * bill_depth_mm:speciesChinstrap +
##     B_bill_depth_mm:speciesGentoo * bill_depth_mm:speciesGentoo
## E(y | x) =
##     -283.278731161643 +
##     217.151603943763 * bill_depth_mm +
##     247.059540391324 * speciesChinstrap +
##     -175.706460904053 * speciesGentoo +
##     -12.5268993514103 * bill_depth_mm:speciesChinstrap +
##     152.289018274643 * bill_depth_mm:speciesGentoo
## Number of observations: 342
## Estimated variance of errors: 125967.828935121
## Multiple R:  0.898334820255734   Adjusted R:  0.896734914322461
## Critical value t(alpha/2 = 0.05/2, DFRes = 336): 1.9670493839589
## Critical value F(alpha = 0.05, DFR = 5, DFRes = 336): 2.2408542774738
```

4. Carry out the relevant hypothesis test to see if the interaction terms can be dropped. What is the conclusion?

```
reduced_model <- lm(body_mass_g ~ bill_depth_mm + species, data = data_set)
analyze_variance_for_reduced_and_full_linear_models(reduced_model, full_model)
```

```
## Analysis of Variance Table
##
## Model 1: body_mass_g ~ bill_depth_mm + species
## Model 2: body_mass_g ~ bill_depth_mm * species
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1    338 44399670
## 2    336 42325191  2   2074479 8.2342 0.0003227 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## test statistic F0 for Partial F Test: 8.23416263648462
## Fc(alpha = 0.05, predictors_dropped = 2, DFRes(full) = 336) = 3.02260130221784
## P(F > F0) for Partial F Test: 0.000322717591739742
## significance level: 0.05
```

We conduct the Partial F Test to investigate if the interaction terms $\beta_4 x_1 I_1$ and $\beta_5 x_1 I_2$ are 0 / jointly insignificant in the context of the full multiple linear model and all predictors / can be dropped from the full model. The test statistic for the Partial F Test $F_0 = 8.234$. Since the test statistic is greater than a critical value $F_c = 3.023$, we have sufficient evidence to reject a null hypothesis that the regression coefficients for the interaction terms are 0 / the interaction terms are jointly insignificant / the interaction terms can be dropped from the full model. We have sufficient evidence to support an alternate hypothesis that a regression coefficient for a dropped predictor is not 0 / the interaction terms are jointly significant / the interaction terms cannot be dropped from the full model. The full model should be used. There is sufficient evidence that the rate at which body mass increases for increasing bill depth for Gentoo penguins is different that the rates for Adelie and Chinstrap penguins.

5. Based on your answer in part 4, write out the estimated regression equations relating body mass and bill depth, for each species of penguins.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 I_1 + \beta_3 I_2 + \beta_4 x_1 I_1 + \beta_5 x_1 I_2 + \epsilon$$
$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 I_1 + \beta_3 I_2 + \beta_4 x_1 I_1 + \beta_5 x_1 I_2$$

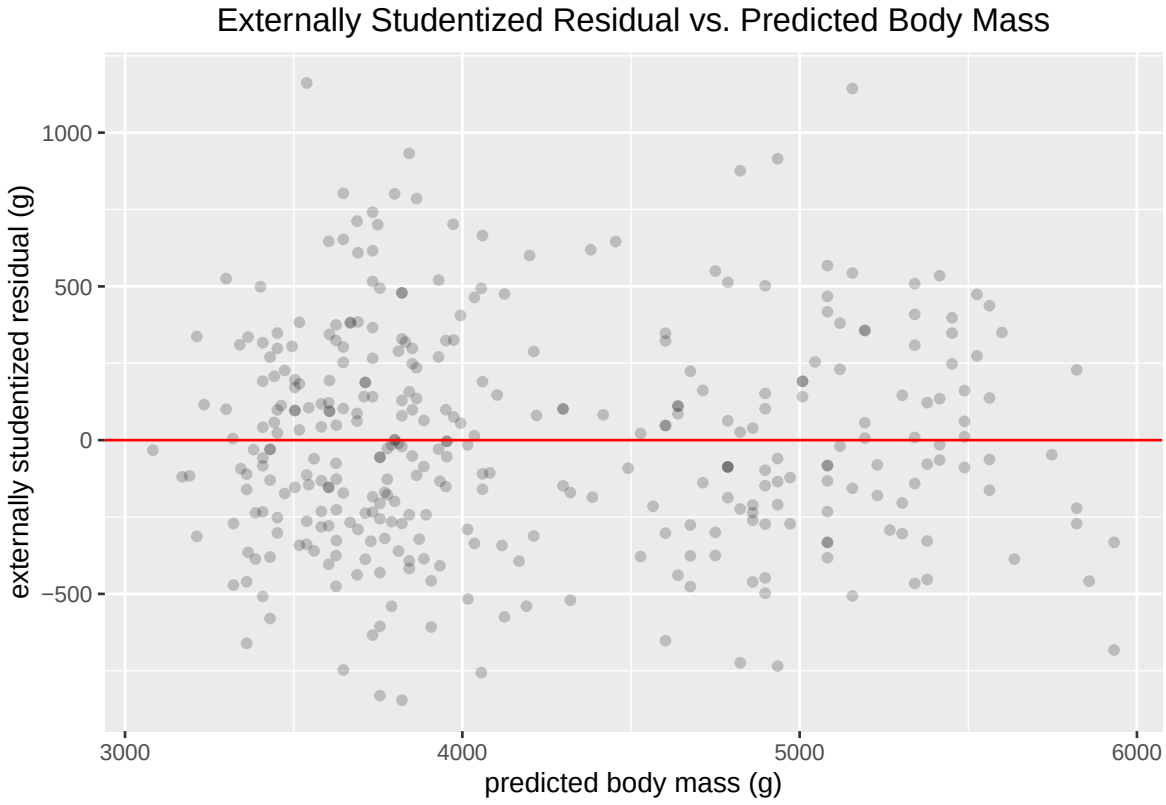
$$E(y|species = Adelie) = \beta_0 + \beta_1 x_1 + \beta_2(0) + \beta_3(0) + \beta_4 x_1(0) + \beta_5 x_1(0)$$
$$E(y|species = Adelie) = \beta_0 + \beta_1 x_1$$

$$E(y|species = Chinstrap) = \beta_0 + \beta_1 x_1 + \beta_2(1) + \beta_3(0) + \beta_4 x_1(1) + \beta_5 x_1(0)$$
$$E(y|species = Chinstrap) = \beta_0 + \beta_1 x_1 + \beta_2 + \beta_4 x_1$$
$$E(y|species = Chinstrap) = (\beta_0 + \beta_2) + (\beta_1 + \beta_4)\, x_1$$

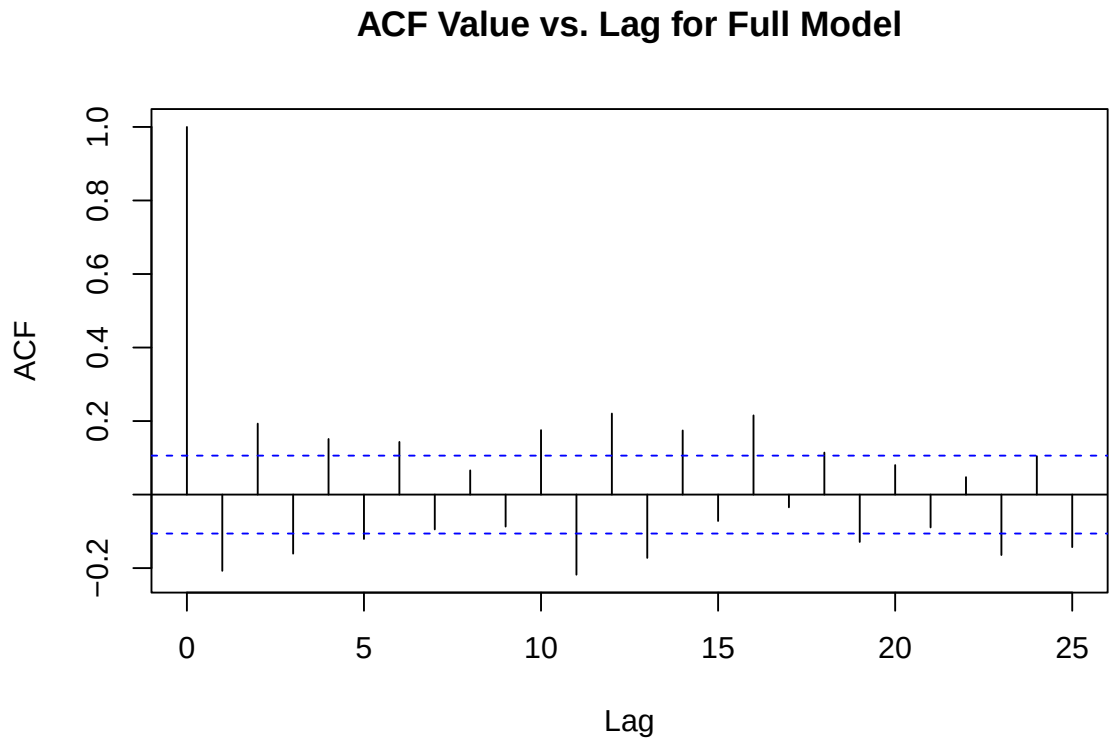
$$E(y|species = Gentoo) = \beta_0 + \beta_1 x_1 + \beta_2(0) + \beta_3(1) + \beta_4 x_1(0) + \beta_5 x_1(1)$$
$$E(y|species = Gentoo) = \beta_0 + \beta_1 x_1 + \beta_3 + \beta_5 x_1$$
$$E(y|species = Gentoo) = (\beta_0 + \beta_3) + (\beta_1 + \beta_5)\, x_1$$

6. Assess if the regression assumptions are met for the model you recommend to use based on part 4. Also, be sure to carry out Levene's test of equality of variances since we have a categorical predictor.

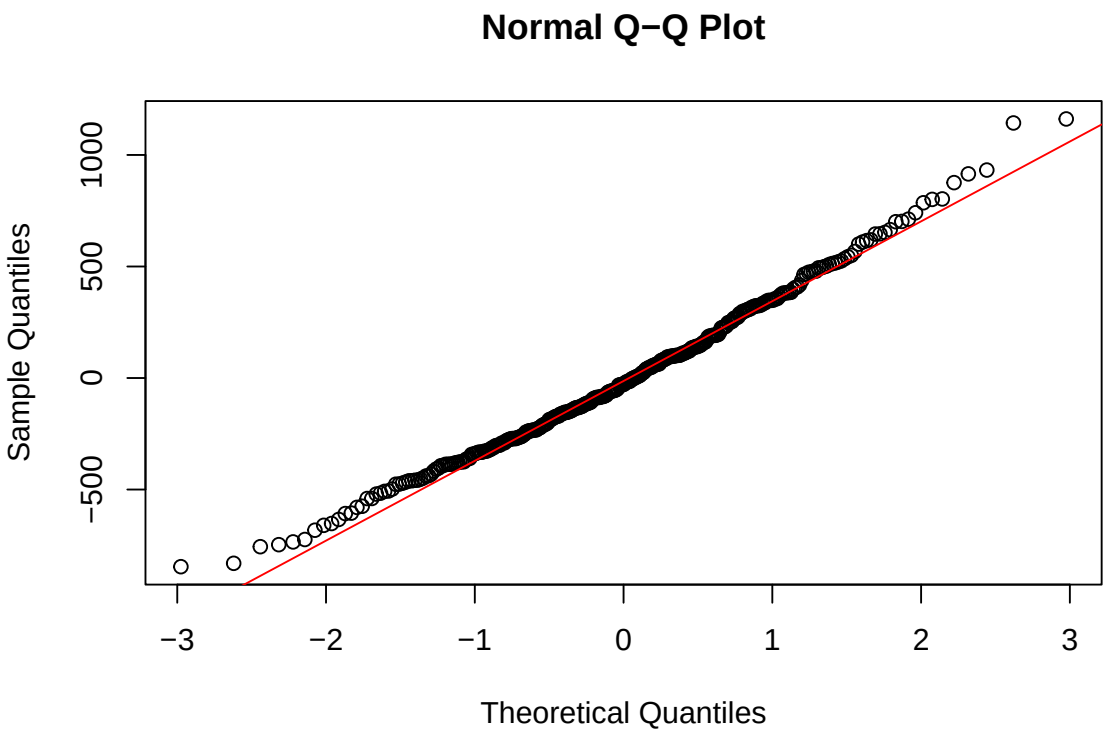
```
library(ggplot2)
ggplot(
    data.frame(
        externally_studentized_residual = full_model$residuals,
        predicted_body_mass = full_model$fitted.values
    ),
    aes(x = predicted_body_mass, y = externally_studentized_residual)
) +
    geom_point(alpha = 0.2) +
    geom_hline(yintercept = 0, color = "red") +
    labs(
        x = "predicted body mass (g)",
        y = "externally studentized residual (g)",
        title = "Externally Studentized Residual vs. Predicted Body Mass"
    ) +
    theme(
        plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 0)
    )
```




```
acf(full_model$residuals, main = "ACF Value vs. Lag for Full Model")
```




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



I recommend the full model of body mass versus bill depth, indicator variables for the various species, and interaction terms between bill depth and indicator variables.

1. The assumption that the relationship between response and predictors is linear, at least approximately, is met cannot be addressed.
2. The assumption that the residuals of the linear model have mean 0 is met. Residuals are evenly scattered around $e = 0$ at random.
3. The assumption that the distributions of residuals of the linear model have constant variance is met. Residuals are evenly scattered around $e = 0$ with constant vertical variance.
4. The assumption that the residuals of the linear model are uncorrelated is not met. The ACF value for lag 0 is always 1; the correlation of the vector of residuals with itself is always 1. Since the ACF value for the first seven lags at least are significant, we have sufficient evidence to reject a null hypothesis that the residuals of the linear model are uncorrelated. We have sufficient evidence to conclude that the residuals of the linear model are correlated. We have sufficient evidence to conclude that the assumption that the residuals of the linear model are uncorrelated is not met.
5. The assumption that the residuals of the linear model are normally distributed is met. A linear model is robust to these assumptions. Considering a plot of sample quantiles versus theoretical quantiles for the residuals of the linear model, since observations lie near the line of best fit / their theoretical values, a probability vs. externally studentized residuals plot / distribution is normal.

We conduct Levene's test for equality of variances of body mass across all species.

```
ggplot(species_bill_depth_and_body_mass, aes(x = species, y = body_mass_g)) +
    geom_boxplot(fill = "Blue", color = "Orange") +
    labs(
        title = "Distributions of Body Mass by Species",
        y = "body mass (g)",
        x = "species"
    ) +
```

```
    theme(
        plot.title = element_text(hjust = 0.5, size = 11),
        axis.text.x = element_text(angle = 0)
    )
```

The spread of body mass seems to decrease across species Gentoo, Adelie, and Chinstrap.

```
library(lawstat)
levene.test(
    species_bill_depth_and_body_mass$body_mass_g,
    species_bill_depth_and_body_mass$species
)
```

```
##
##  Modified robust Brown-Forsythe Levene-type test based on the absolute
##  deviations from the median
##
## data:  species_bill_depth_and_body_mass$body_mass_g
## Test Statistic = 5.1203, p-value = 0.006445
```

The null hypothesis for Levene's test is that the variances of body mass across all species are equal. Since the p-value is less than significance level $\alpha = 0.05$, we have sufficient evidence to reject the null hypothesis; we have evidence that this assumption is not met.

7. Briefly explain if we can conduct pairwise comparisons for the difference in mean body mass among all pairs of species for given values for given values of bill depth. These pairs of species are (Adelie, Chinstrap), (Adelie, Gentoo), and (Chinstrap, Gentoo). If we are able, conduct Tukey's multiple comparsions and contextually interpret the results of these hypothesis tests.

   If we had a model with no interactions, we could interpret the coefficents of the indicator variable I_i as

the difference in the mean body mass, given values of bill depth, between the species for which that indicator variable is 1 and the reference species *Adelie*. We have a model with interactions.