SIOFA Sharks

Annex 7: Maturity Ogive Disgnostics for Centroscymnus coelolepis

DELEGATION OF THE EUROPEAN UNION

04 March, 2024

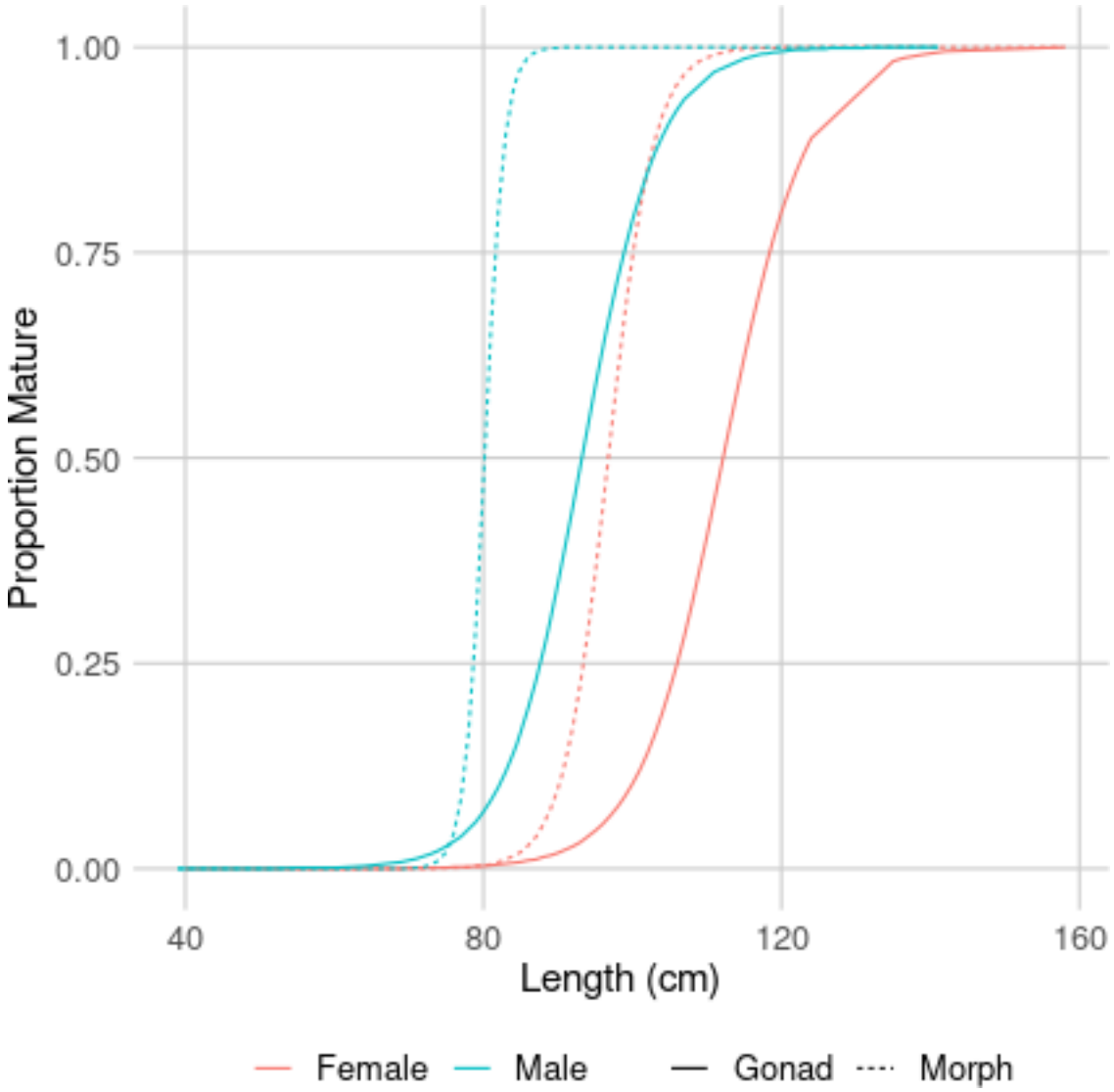
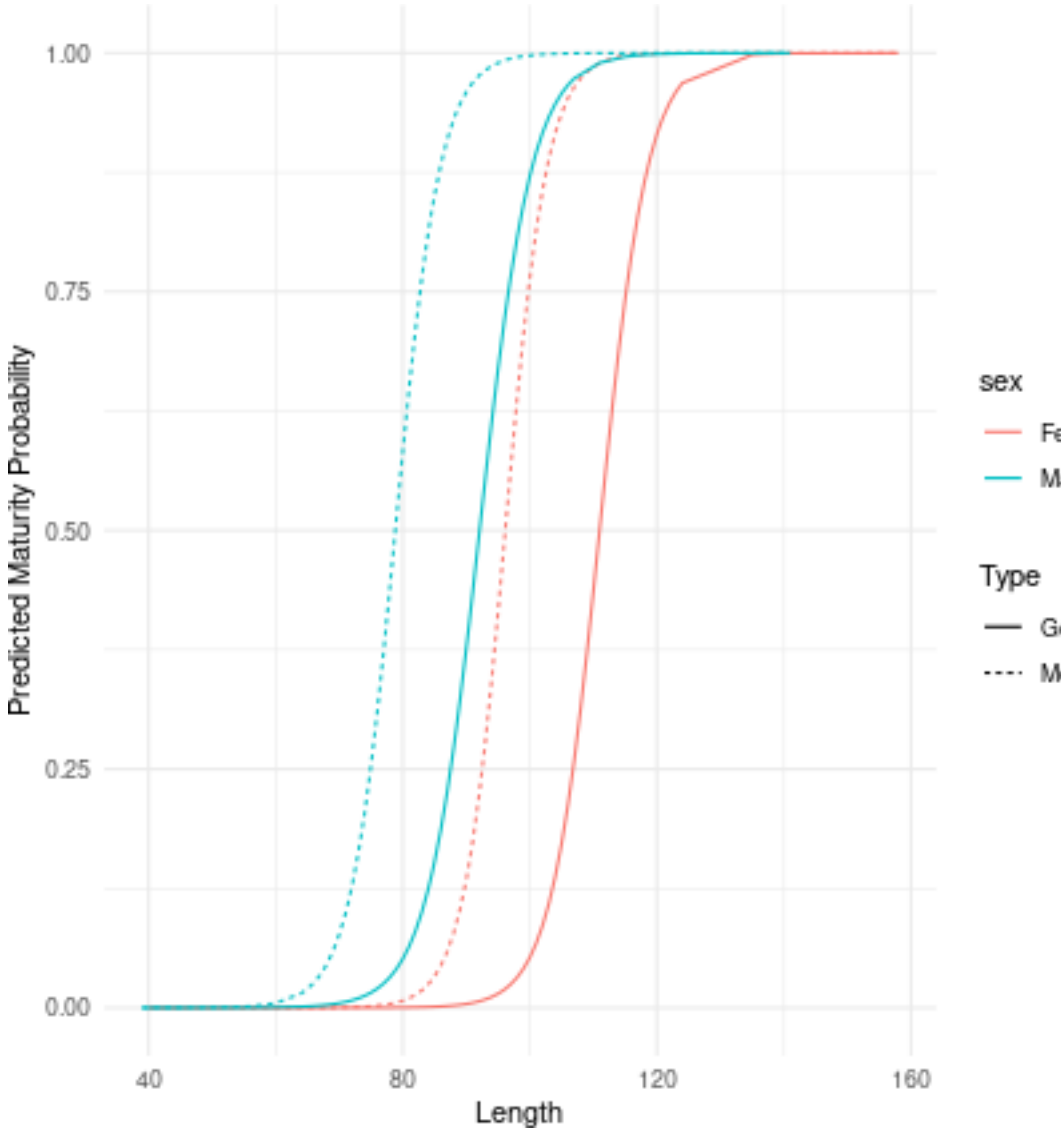


Figure 1 .



Fit logistic regression model

Figure 2 .

1. Summary of Model Fit

- Use `summary(model)` to get an overview of the model fit, including coefficients, standard errors, z-values, and P-values for each predictor.

```
Call:
glm(formula = mature ~ len + sex + len:sex + len:Type, family = binomial,
    data = mat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.9750  -0.4805  -0.0406   0.3807   4.3521

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.921e+01  5.458e-01 -53.519  < 2e-16 ***
len            2.632e-01  5.203e-03  50.590  < 2e-16 ***
```

```
sexMale        6.960e+00  6.783e-01  10.261  < 2e-16 ***
len:sexMale   -2.159e-02  7.206e-03  -2.996  0.00274 **
len:TypeMorph  4.065e-02  6.248e-04  65.051  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28393  on 20607  degrees of freedom
Residual deviance: 13435  on 20603  degrees of freedom
AIC: 13445

Number of Fisher Scoring iterations: 6
```

2. Residuals and Model Fit

- **Residuals Plot**: Plot residuals to look for patterns. For logistic regression, deviance residuals can be informative.

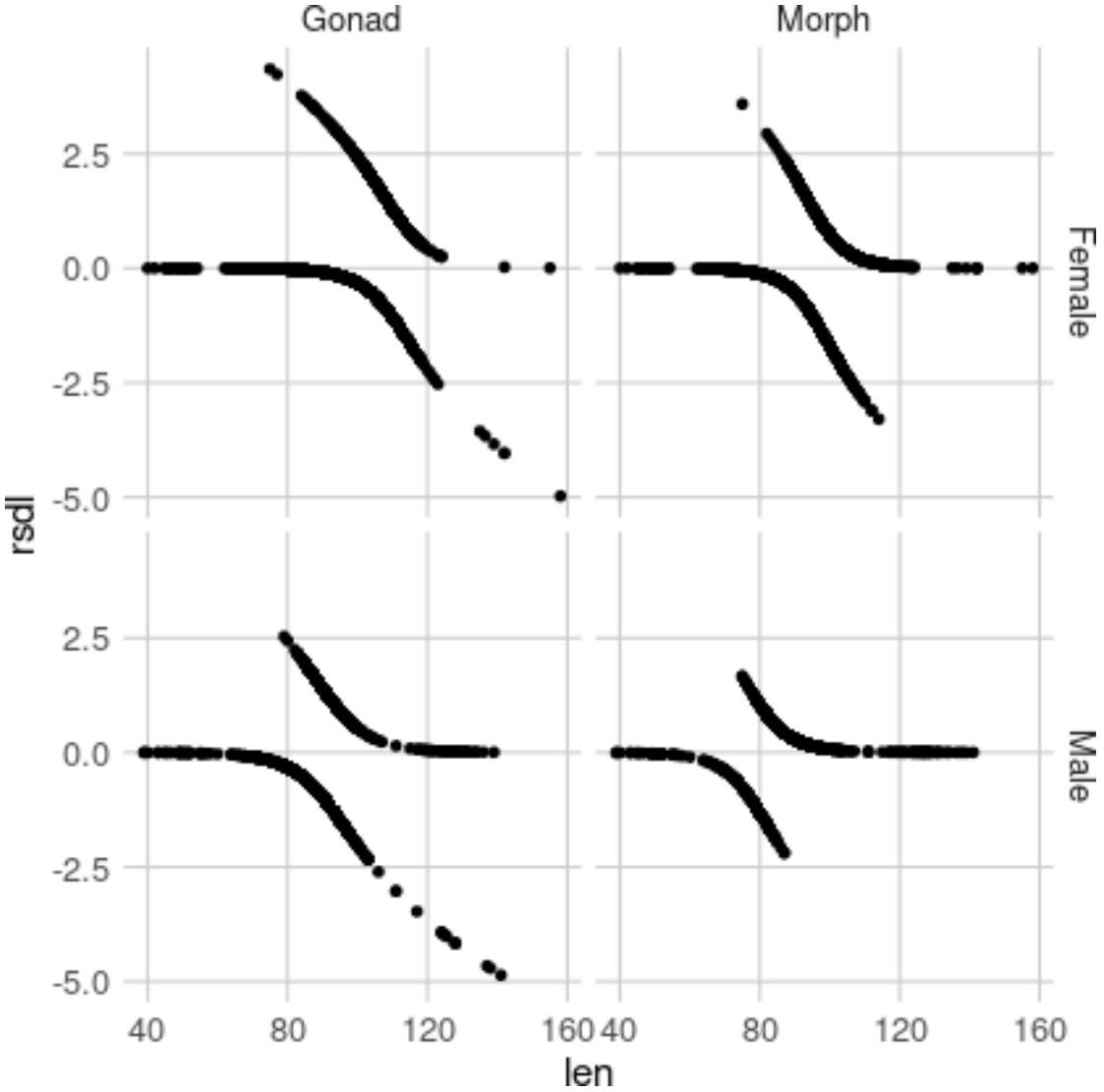


Figure 3 .

- **Hosmer-Lemeshow Test**: Test for goodness of fit specifically designed for logistic regression.

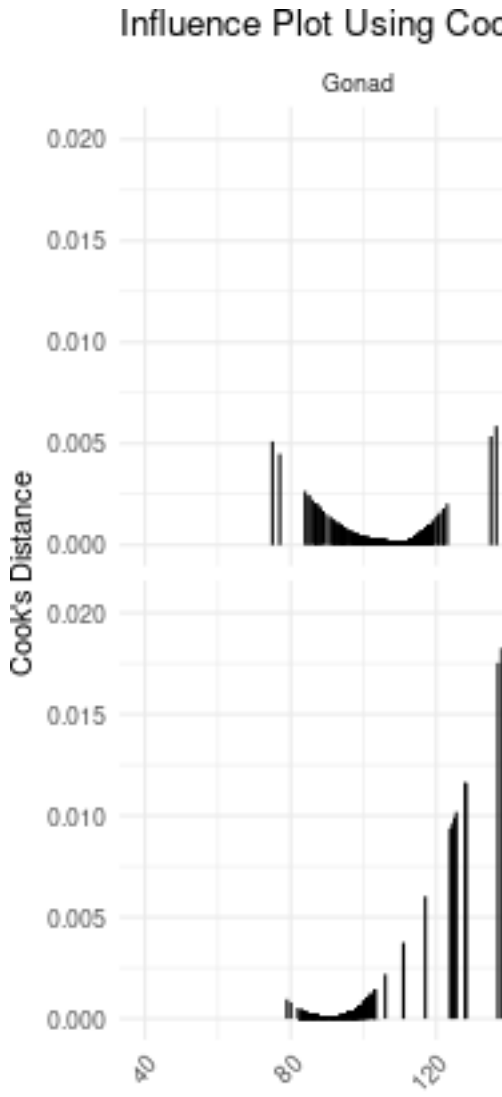
```
Hosmer and Lemeshow goodness of fit (GOF) test

data:  model$y, fitted(model)
X-squared = 122, df = 8, p-value < 2.2e-16
```

3. Influence Measures

- **Cook's Distance**: Identify influential observations based on Cook's distance.

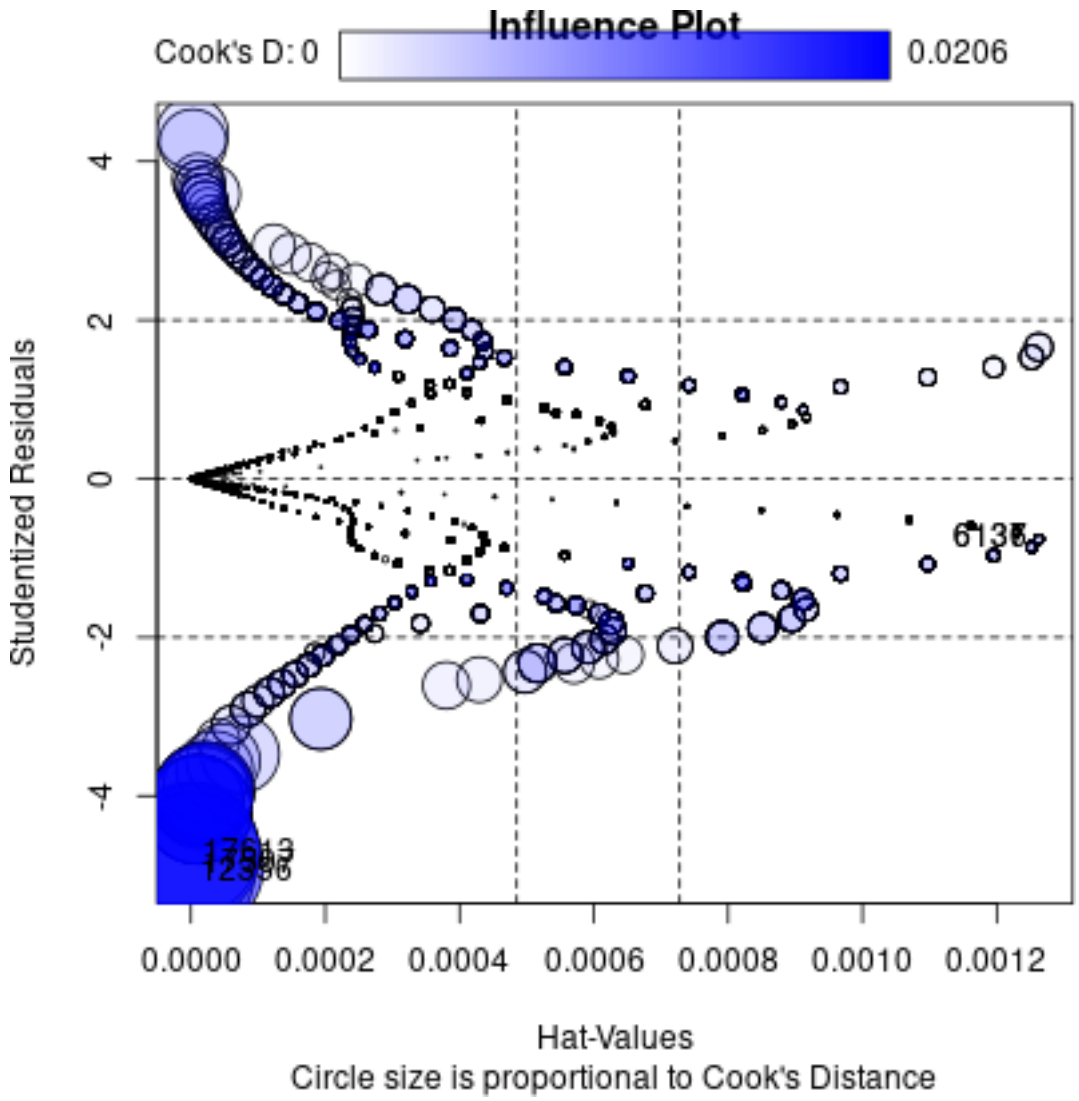
3. Influence Measures



- **Cook's Distance**: Identify influential observations based on Cook's distance.

Figure 4 .

- **Leverage and Influence Plots**: Identify observations with high leverage or influence on the model es-



timation.

```
         StudRes          Hat        CookD
6136  -0.7646307 1.262567e-03 8.588387e-05
6137  -0.7646307 1.262567e-03 8.588387e-05
12356 -4.9829913 3.341237e-07 1.583271e-02
17597 -4.8720466 7.584188e-07 2.055908e-02
17613 -4.7197149 1.391098e-06 1.826662e-02
```

To create an influence plot using `ggplot2` that mimics the functionality of the `car::influencePlot`, you'll need to manually calculate the elements you want to visualize: leverage, standardized residuals, and Cook's distance. Here's how you can do it with a logistic regression model as an example. This approach involves extracting the necessary statistics from the model and then plotting them using `ggplot2`.

First, ensure you have your logistic regression model fitted. For this example, let's assume your model is stored in a variable named `model`. You'll need the `broom` and `ggplot2` packages as well.

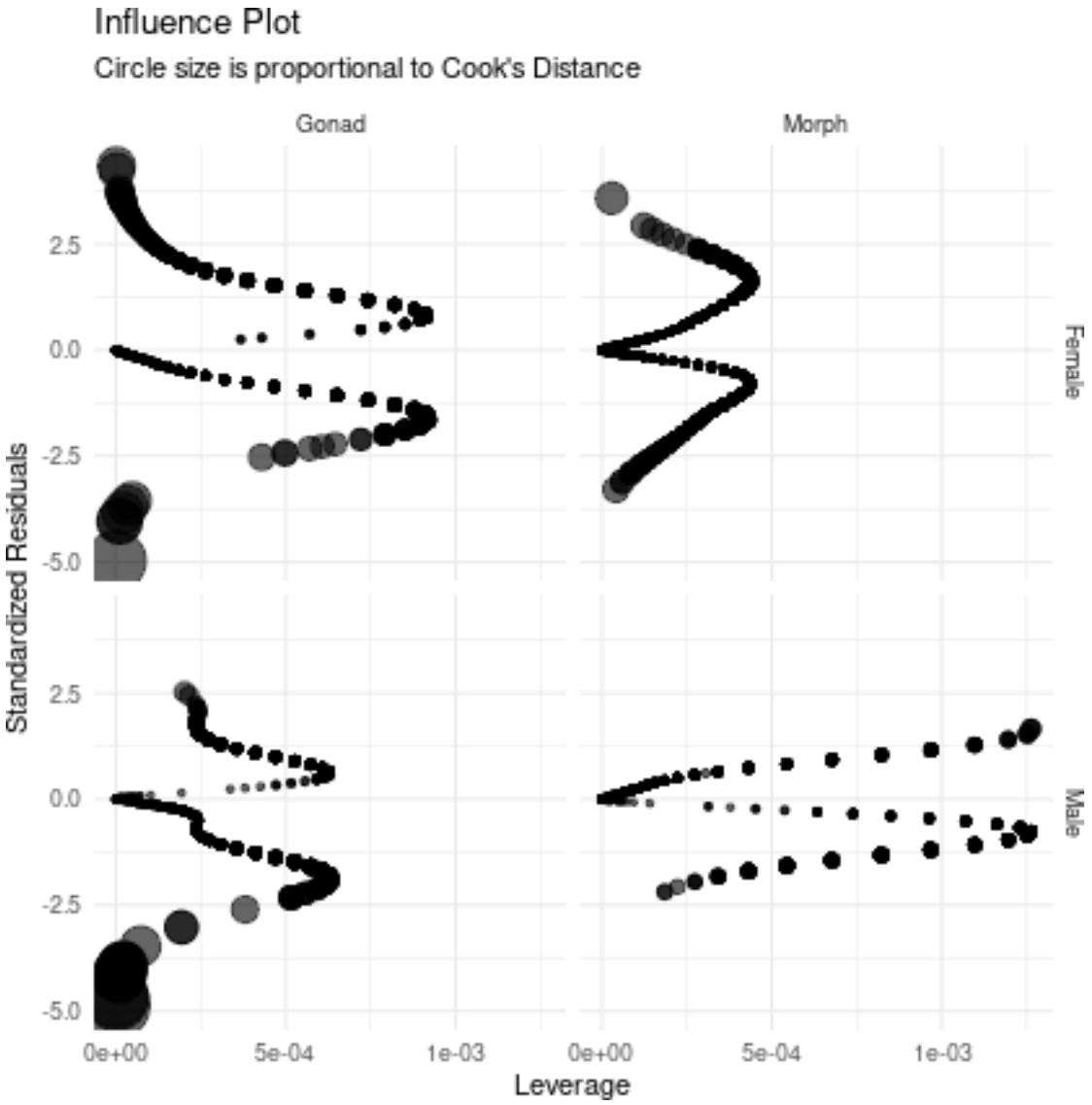


Figure 5 .

This script does the following: - Calculates the leverage, standardized residuals, and Cook's distance for each observation in your logistic regression model. - Constructs a dataframe `influence_data` that contains these values. - Uses `ggplot2` to create a scatter plot where the x-axis is leverage, the y-axis is standardized residuals, and the size of each point represents Cook's distance.

Adjust the `scale_size_continuous` function as needed to better fit your data. The `guide='none'` argument hides the size legend since it generally clutters the plot. This approach gives you a detailed and customizable influence plot using `ggplot2`.

Creating an influence plot with `ggplot2` requires calculating influence measures like Cook's distance, leveraging the model you've fitted, and then plotting these measures. The influence plot visually identifies points that have a significant impact on the model's estimates. Here's how you can create an influence plot for your logistic regression model using `ggplot2` in R:

1. **Calculate Cook's Distance**: Cook's distance is a measure used to estimate the influence of each data point. In logistic regression, it identifies points that, if removed, would change the model's parameters significantly.

2. **Prepare the Data**: Combine the Cook's distance with your original dataset or create a new dataframe with the Cook's distance and leverage values for each observation.

3. **Plot Using `ggplot2`**: Create the plot using `ggplot2`, highlighting observations with high influence.

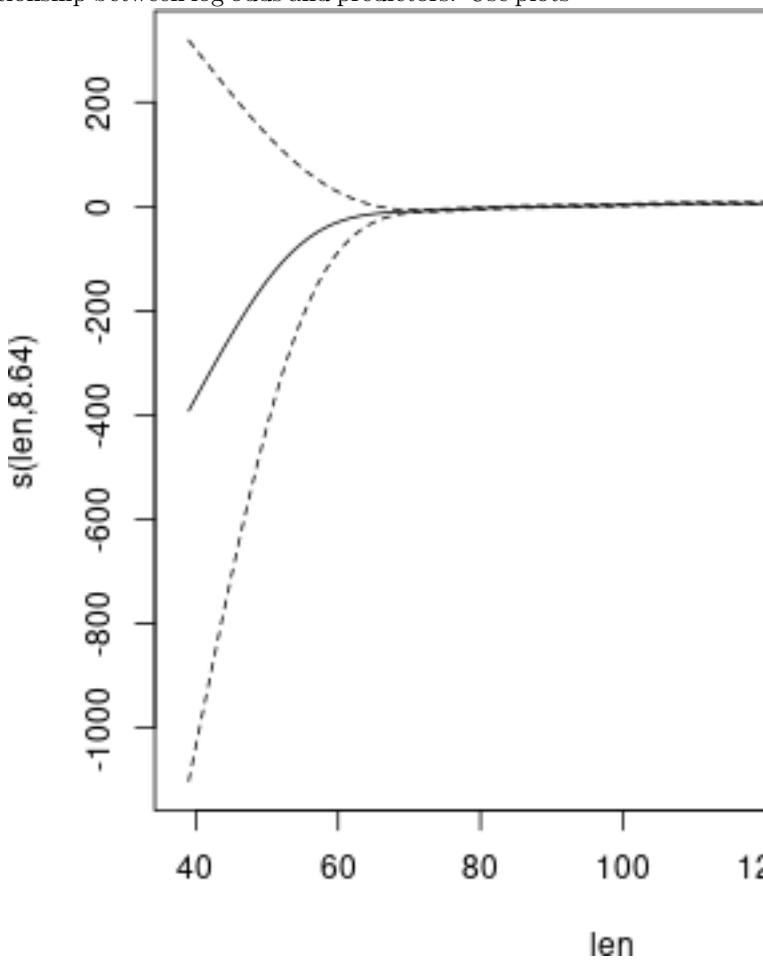
Here's an example code snippet that demonstrates this process:

4. Multicollinearity - **Variance Inflation Factor (VIF)**: Check for multicollinearity among predictors.

```
      len        sex    len:sex   len:Type
 4.526954 244.565253 216.956171   1.773010
```

5. Model Assumptions and Fit

- **Check Linearity**: The logit link assumes a linear relationship between log odds and predictors. Use plots



or Generalized Additive Models (GAMs) to assess this.

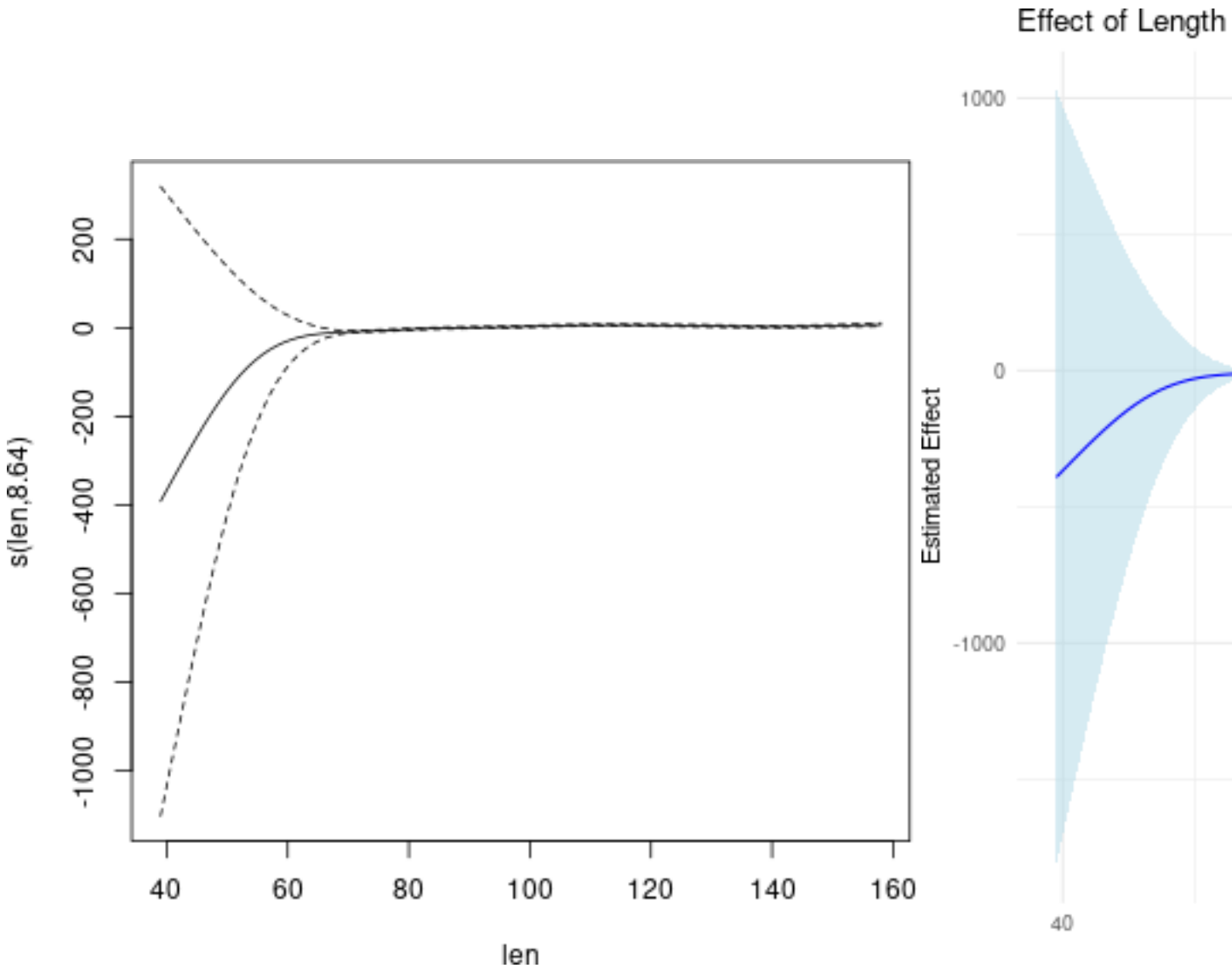


Figure 6 .

- **Check for Overdispersion**: This is less of a concern in binary logistic regression but can be relevant for count data.

6. Predicted vs. Observed

- **Predicted vs. Observed**: Compare the predicted probabilities to the observed outcomes to evaluate the model's performance.

Figure 7 .