UNIVERSITY *of* WEST FLORIDA

Model Assumptions and Diagnostics

Statistics for Data Science II

Introduction

When fitting a linear regression model, we may encounter issues:

1. Non-linearity of the response-predictor relationships.
2. Correlation of error terms.
3. Non-constant variance of error terms.
4. Outliers.
5. High-leverage points.
6. Collinearity.

Model building is an *art* rather than a *science*.

Non-Linearity

The linear regression model assumes a straight-line relationship between the outcome and the predictors.

As we stray further away from linearity:

- Any conclusions drawn are questionable.
- The prediction accuracy of the model is reduced.

We will use a residual plot to assess non-linearity.

- We plot the residuals, $e_i = y_i - \hat{y}_i$, against the predicted value, $\hat{y}_i$.
- The presence of a pattern may indicate a problem with some aspect of the model.

Non-Linearity

Example:

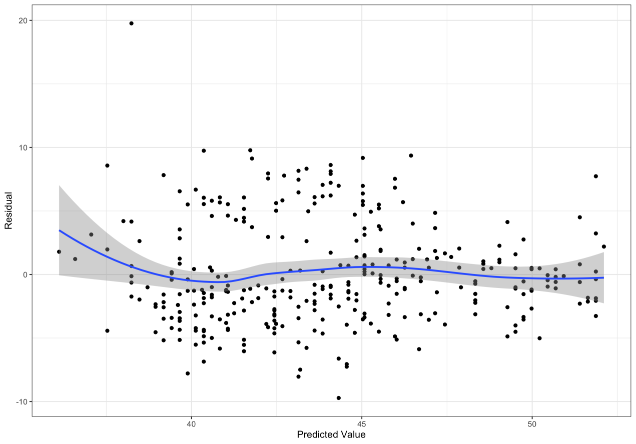
Recall the penguin data. Consider the simple example of modeling bill length as a function of sex and flipper length. Let's look at the residual plot.

```r
m1 <- lm(bill_length_mm ~ flipper_length_mm + sex_male, data = data)
data <- data %>% mutate(
  e = residuals(m1),
  yhat = predict(m1)
  )

p1 <- data %>% ggplot(aes(x = yhat, y = e)) +
  geom_point() +
  geom_smooth() +
  theme_bw() +
  xlab("Predicted Value") +
  ylab("Residual")
```

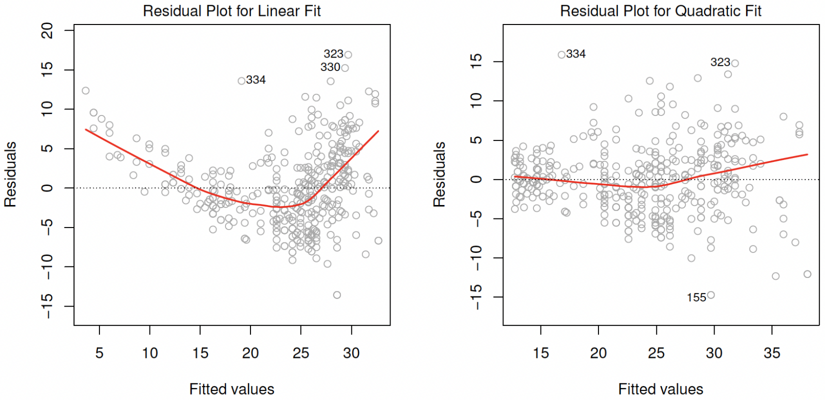
Non-Linearity

Example:



Non-Linearity

Example:



Non-Linearity

What should we do if the residual plot clearly indicates a non-linear relationship?

We can perform transformations on the predictors:

$\log(x)$

$\sqrt{x}$

x^2

etc.

We will deal with this more later on in the course.

Correlation of Error Terms

An important assumption is that the error terms (e_i) are uncorrelated.

If there is correlation among the error terms (repeated measures, time series data, etc.), we must account for it in the model.

> The regression approaches in this course do not account for correlation in error terms.

If we have correlated error terms but do not account for them, our inferences may not be correct – standard errors will be too small.

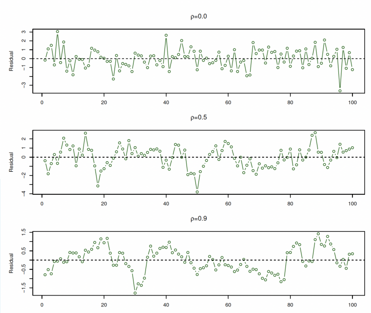
> test statistics are larger than they should be,
>
> p-values are smaller than they should be,
>
> and confidence intervals are narrower than they should be.

Correlation of Error Terms

If time is a component of the analysis, we can plot the residuals against time.

If small correlation, then there should be no pattern.

If positive correlation, then adjacent residuals may have similar values.


ρ=0.0
Residual
ρ=0.5
Residual
ρ=0.9
Residual