Model Checking
Testing the Linearity Assumption

- A residual (error) is the difference between a predicted and an observed value.
- A residual plot is a scatter plot of residuals versus predicted (fitted) values.
- The residual plot of a linear regression model should look like a formless cloud.
- If there is a discernable pattern, our linear model is missing something key.

In case it isn’t obvious, these plots are NOT formless clouds!
A residual plot for ice cream sales

- This residual plot is a formless cloud.
- Thus, linear regression is a suitable model.

```r
fit <- lm(sales ~ temperature)
plot(residuals(fit) ~ fitted(fit), ylab = "Residuals", xlab = "Fitted Values") + abline(a = 0, b = 0)
```
Scatter Plot of Old Faithful

- The scatter plot shows two clusters, one with low eruptions and low waiting time, and another with high eruptions and high waiting time
- The scatter plot is not a formless cloud
- There are diagonal stripes, suggesting a nonlinear pattern in the data
Summary

These are characteristics of a well-behaved residual vs. fits plot:

- The residuals "bounce randomly" around the x-axis, implying a decent fit, so that the assumption of linearity is valid.
- No residuals "stand out" from the others, meaning there are no outliers.
Pitfalls
Regression Pitfalls

- Is the relationship truly linear?
- Extrapolating beyond sample data
- Correlation does not imply causation
- Reverse causality
  - Life-long smokers quit smoking after being diagnosed with cancer…
  - And then they die! Did they die because they quit smoking?  
- Omitted variable bias
  - Golfers are prone to heart disease! (Age is likely an omitted variable!)
- Correlated explanatory variables
  - Parents’ education levels, when modeling attributes of children (e.g., SAT scores)
- Data mining: Maybe you should just throw in everything in the kitchen sink!
Violations are not always easy to detect!

- Assume a nonlinear relationship:
  - $x \leftarrow \text{rnorm}(25, 10, 2)$
  - $y \leftarrow x^2 + \text{rnorm}(25, 0, 1)$

- And build a linear model:
  - `summary(lm(y ~ x))$coefficients`

- R tells us that the $x$ coefficient is near .5, and that this coefficient is statistically significant, but a linear model is incorrect!