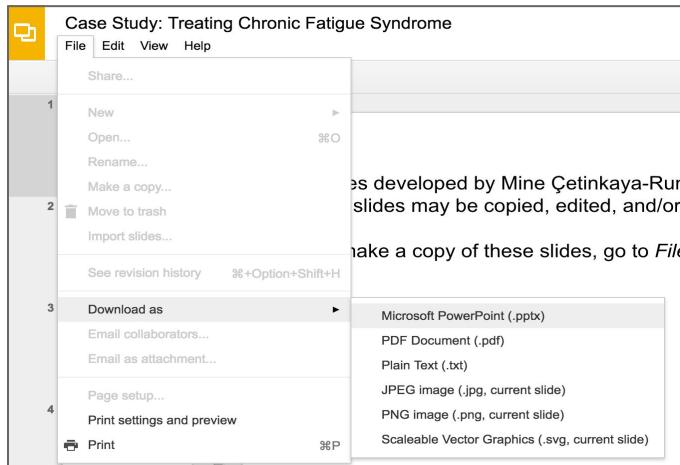


Slides developed by Mine Çetinkaya-Rundel of OpenIntro

Translated from LaTeX to Google Slides by Curry W. Hilton of OpenIntro.

The slides may be copied, edited, and/or shared via the [CC BY-SA license](#)

To make a copy of these slides, go to *File > Download as > [option]*, as shown below. Or if you are logged into a Google account, you can choose *Make a copy...* to create your own version in Google Drive.



Checking model conditions using graphs

Modeling conditions

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

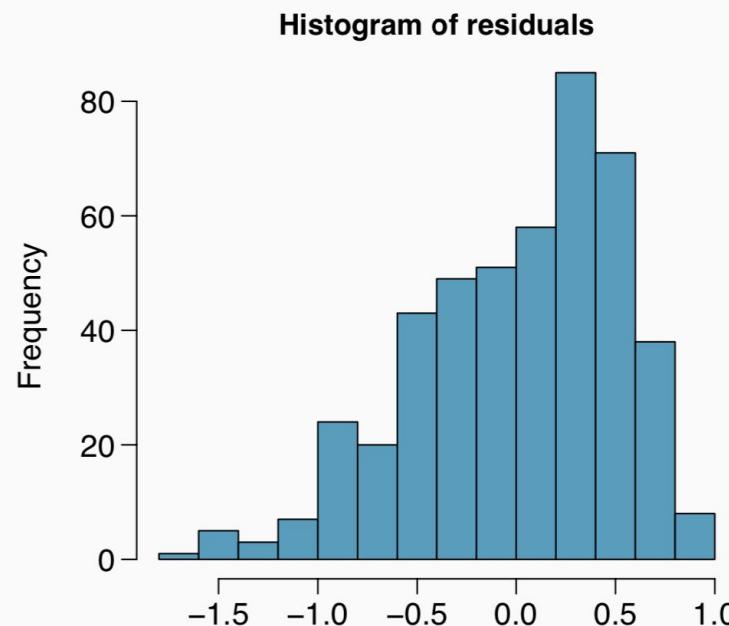
The model depends on the following conditions

1. residuals are nearly normal (less important for larger data sets)
2. residuals have constant variability
3. residuals are independent
4. each variable is linearly related to the outcome

We often use graphical methods to check the validity of these conditions, which we will go through in detail in the following slides.

(1) nearly normal residuals

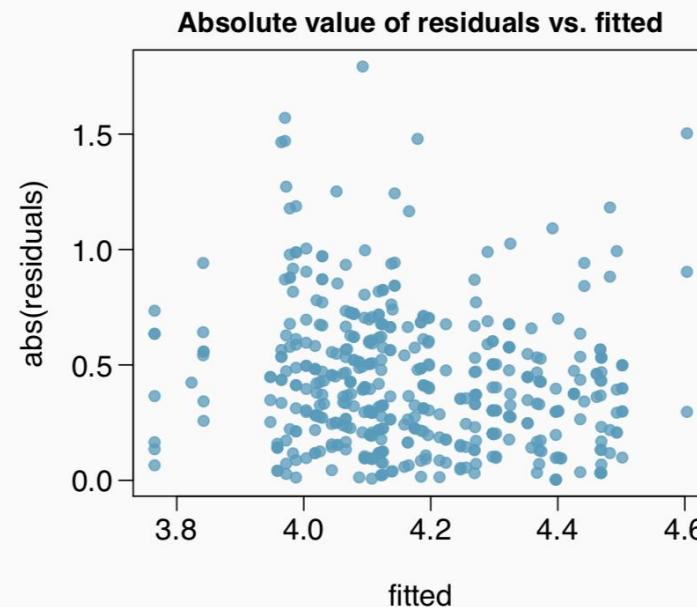
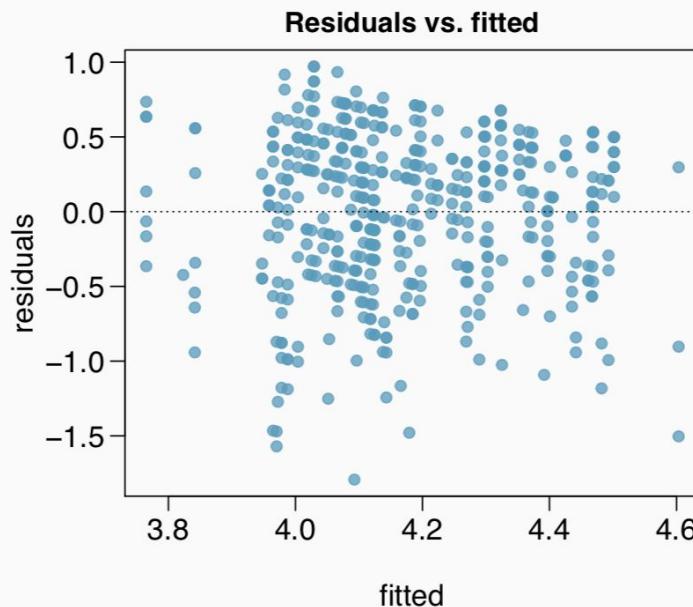
Histogram of the residuals.



Does this condition appear to be satisfied?

(2) constant variability in residuals

Scatterplot of residuals and/or absolute value of residuals vs. fitted (predicted).



Does this condition appear to be satisfied?

Checking constant variance - recap

When we did simple linear regression (one explanatory variable) we checked the constant variance condition using a plot of *residuals vs. x*.

With multiple linear regression (2+ explanatory variables) we checked the constant variance condition using a plot of *residuals vs. fitted*.

Why are we using different plots?

Checking constant variance - recap

When we did simple linear regression (one explanatory variable) we checked the constant variance condition using a plot of *residuals vs. x*.

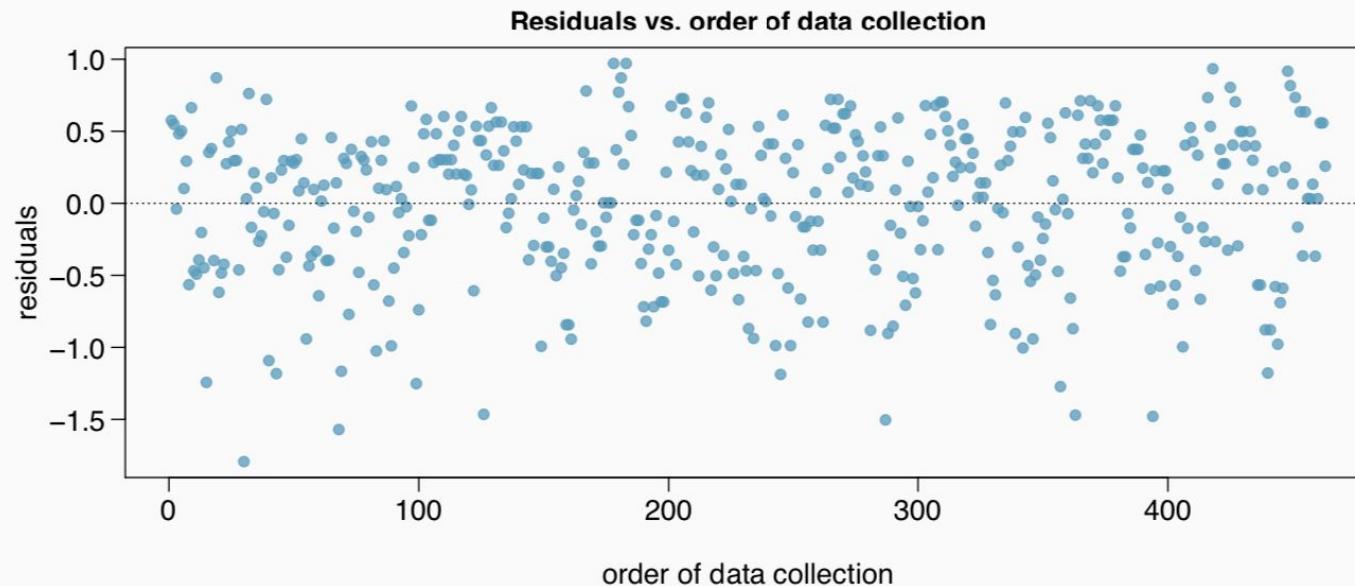
With multiple linear regression (2+ explanatory variables) we checked the constant variance condition using a plot of *residuals vs. fitted*.

Why are we using different plots?

In multiple linear regression there are many explanatory variables, so a plot of residuals vs. one of them wouldn't give us the complete picture.

(3) independent residuals

Scatterplot of residuals vs. order of data collection.



Does this condition appear to be satisfied?

More on the condition of independent residuals

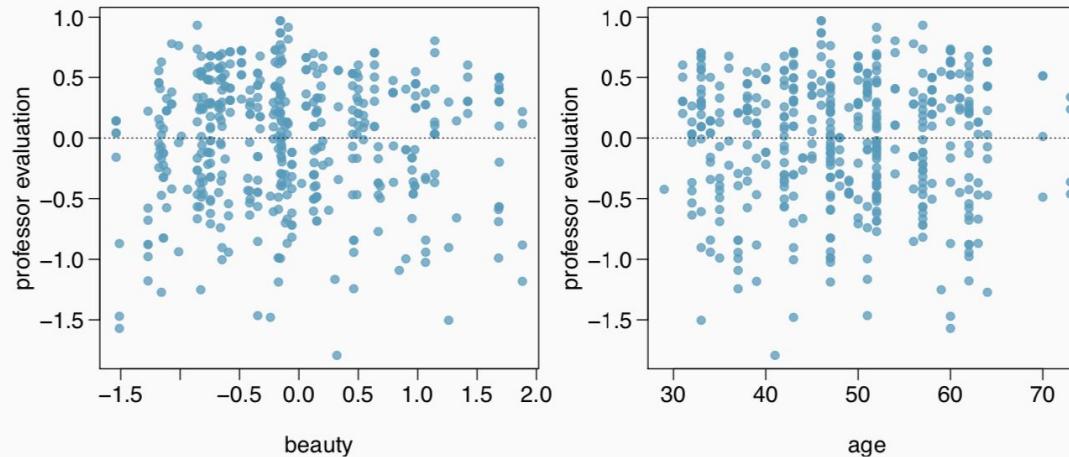
Checking for independent residuals allows us to indirectly check for independent observations.

If observations and residuals are independent, we would not expect to see an increasing or decreasing trend in the scatterplot of residuals vs. order of data collection.

This condition is often violated when we have time series data. Such data require more advanced time series regression techniques for proper analysis.

(4) linear relationships

Scatterplot of residuals vs. each (numerical) explanatory variable.



Does this condition appear to be satisfied?

Note: We use residuals instead of the predictors on the y-axis so that we can still check for linearity without worrying about other possible violations like collinearity between the predictors.

Several options for improving a model

Transforming variables

Seeking out additional variables to fill model gaps

Using more advanced methods that would account for challenges around inconsistent variability or nonlinear relationships between predictors and the outcome

Transformations

If the concern with the model is non-linear relationships between the explanatory variable(s) and the response variable, transforming the response variable can be helpful.

- Log transformation ($\log y$)
- Square root transformation (\sqrt{y})
- Inverse transformation ($1/y$)
- Truncation (cap the max value possible)

It is also possible to apply transformations to the explanatory variable(s), however such transformations tend to make the model coefficients even harder to interpret.

Models can be wrong, but useful

All models are wrong, but some are useful.

- George Box

No model is perfect, but even imperfect models can be useful, as long as we are clear and report the model's shortcomings.

If conditions are grossly violated, we should not report the model results, but instead consider a new model, even if it means learning more statistical methods or hiring someone who can help.

Find more resources at openintro.org/os, including

- Slides
- Videos
- Statistical Software Labs
- Discussion Forums (free support for students and teachers)
- Learning Objectives

Teachers only content is also available for [Verified Teachers](#), including

- Exercise solutions
- Sample exams
- Ability to request a free desk copy for a course
- Statistics Teachers email group

Questions? [Contact us](#).