Module 5: Model Diagnostics and Remedial Measures in SLR

Jeffrey Woo

MSDS, University of Virginia

Welcome

- Remind me to record the live session!
- Recommended: put yourself on mute unless you want to speak.
- Reminder: the raise hand button can be found under “Reactions”.

Agenda

- A few comments about Module 5
- Working through Guided Question Set 5
- Project 1

Linear Regression Model

- Regression model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
- $y = f(x) + \epsilon$, where $f(x) = \beta_0 + \beta_1 x$.
- Assumptions about ϵ:
  - In mathematical form: $\epsilon_1, \ldots, \epsilon_n$ i.i.d. $\sim \mathrm{N}(0, \sigma^2)$ (i.i.d. means independent and identically distributed)

Assumptions for Linear Regression Model

1. The errors, for each fixed value of x, have mean 0. This implies that the relationship as specified in the regression equation, $y = f(x)$, is appropriate.
2. The errors, for each fixed value of x, have constant variance. That is, the variation in the errors is theoretically the same regardless of the value of x (or $\hat{y}$).
3. The errors are independent.
4. The errors, for each fixed value of x, follow a normal distribution.

Goals of Statistical Models

- Prediction: given x, predict y.
- Inference: how does the y variable relate to the x variable?

Consequences of Not Meeting Regression Assumptions

Generally, the "regression is unreliable". More specifically:

1. Wrong functional form for $f(x)$. So predictions will systematically over- or under-predict. Estimated coefficients are biased.
2. Results from hypothesis tests and intervals and interpreting various measures such as R^2 are unreliable.
3. Results from hypothesis tests and intervals and interpreting various measures such as R^2 are unreliable.
4. Model is fairly robust to this assumption not being met. Prediction intervals may get affected, but other results are still reliable.

Depending on which of the two goals you are focused on, some consequences may be more / less serious than others.

General Rule for Data Transformation

Residual plots can help detect issues 1 and 2. Data transformations can be used to deal with issues 1 and 2.

- Transforming the response is performed to handle issue 2. A successful transformation of the response will result in a residual plot with constant variance.
- Transforming the response may also influence issue 1; however, the choice of how to transform the response is chosen to solve issue 2.
- Transforming the predictor is performed to handle issue 1. Transforming the predictor does not, theoretically, influence issue 2.
- When both issues 1 and 2 are present, we transform the response first, to handle issue 2. Then we transform the predictor if issue 1 is still present.

Non Constant Variance

Two main kinds of non constant variance we can solve with transforming the response variable.

Box Cox Transformation

Residual plot is an empirical way to evaluate assumptions. The Box Cox method is an analytical way to transform the response to deal with the constant variance and normality assumptions.

- $y^{(\lambda)} = y^{\lambda}$ if $\lambda \neq 0$
- $y^{(\lambda)} = \log y$ if $\lambda = 0$

Note: In most statistics literature, log is log base e or ln. Same thing with R, if no base is stated, base e is assumed.

Transforming the Predictor

(Assuming the constant variance assumption is dealt with) Use shape of scatterplot to guide decision on how to transform predictor.
`https://www.mathsisfun.com/sets/functions-common.html`

Interpreting Transformed Variables

If your goal is inference, and interpreting the regression coefficients is important, then log transformed variables are preferred as we can still interpret the coefficients. Any other type of transformation leads to coefficients that are difficult / impossible to interpret.

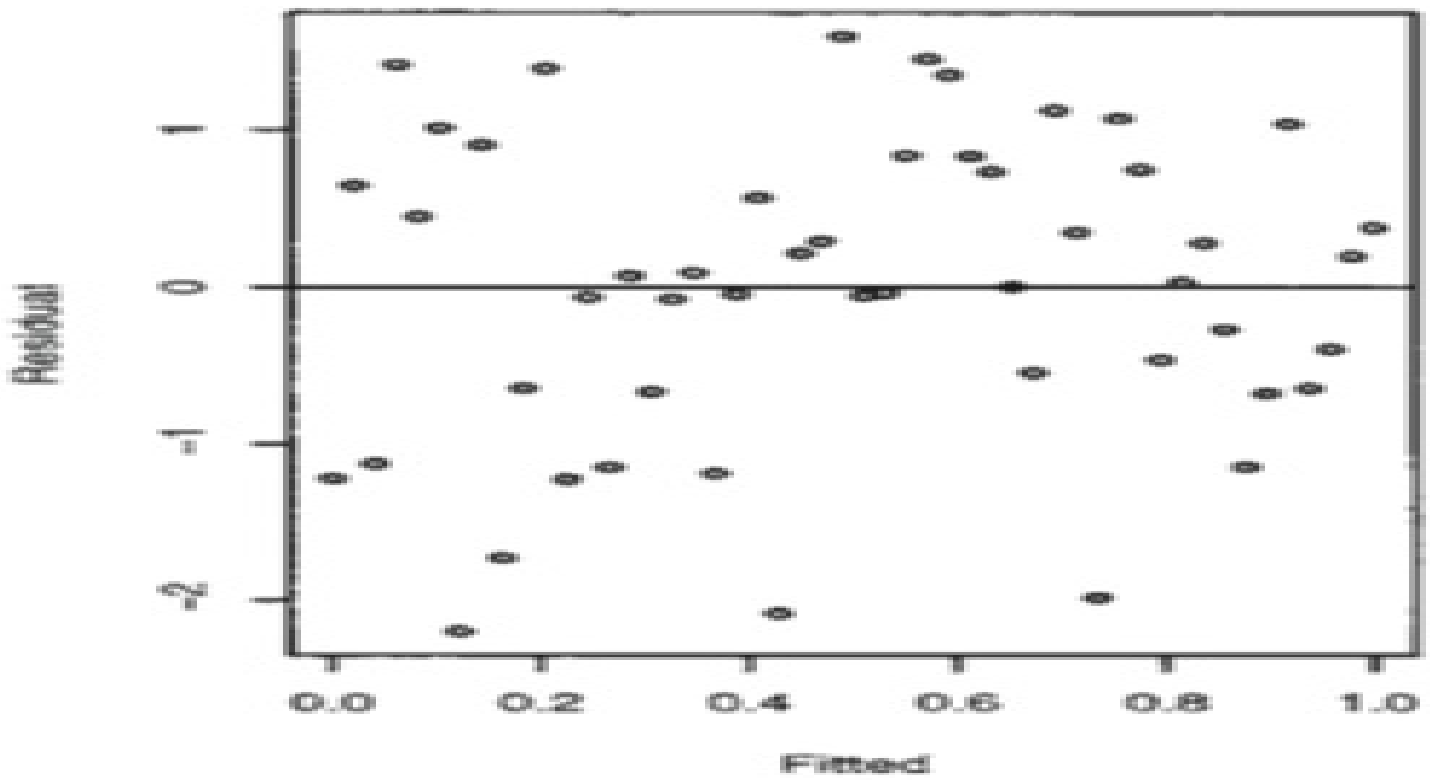
Hierarchical Principle

- If you think the relationship between the response and predictor is of a higher order polynomial (e.g. quadratic, cubic), the hierarchical principle states that the lower order terms should remain in the model.
- For example, if the relationship is of order h, fit $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_h x^h + \epsilon$. Fit in a multiple predictor framework.
- Linear regression models should be invariant to scaling and shifting.

General Comments on Residual Plot

- When assessing the assumptions with a residual plot, we are assessing if the assumptions are reasonably / approximately met.
- With real data, assumptions are rarely met 100%.
- If unsure, proceed with model building, and test how model performs on new data. If poor performance, go back to residual plot to assess what transformation will be appropriate.

General Comments on Residual Plot



Going Over Guided Question Set

What is coming up...

- HW 5 on Oct 3, as usual. Note that Question 2 has several ways to solve the problem, so be sure to clearly document your thought process.
- Project 1 due on Oct 16. Grouping with Module 5 to 8.
- No meeting on Oct 4, due to UVa Fall Break.
- I will still hold office hours on Monday, Oct 3 and 10.
- Next module (Oct 11): Multiple linear regression (multiple predictors). As you go through the material, note the similarities and differences with SLR.