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

- Q & A
- A few comments about Module 5
- Working through Guided Question Set 5

Q & A

Any questions?

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

From section 2 of notes.

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.

Consequences of Not Meeting Regression Assumptions

From sections 2.1 to 2.4 of notes. 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.

General Rule for Data Transformation

From section 3.5 of notes.

- 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.

Box Cox Transformation

From section 4 of notes. 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

From section 5. (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 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.

Please see sections 4.3, 5.2, and 5.3 of notes for specifics on how to interpret coefficients with log transformed variables.

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.

Going Over Guided Question Set

What is coming up...

- No meeting next week. Mar 2 to 10 is UVa Spring Break.
- Office hours: Skye on Thursday Feb 29 (not on Mar 7). I will hold mine on Mar 11 (not on Mar 4).
- HW 5 due on Mar 11. Note that Question 1 has several ways to solve the problem, so be sure to clearly document your thought process.
- Next module: Multiple linear regression (multiple predictors). As you go through the material, note the similarities and differences with SLR.

What is coming up...

- Project 1 is up. Covers Units 1 to 5. Grouping with Module 5 to 8. I have created videos explaining the main info. Find in between Modules 5 and 6. I will provide time on Mar 12 class for you all to get started on the Group Expectations.
- Do not pressure group members to work on project during Spring Break.