Solutions to Assumption Violations

Sean Hellingman ©

Introduction to Statistical Data Analysis (ADSC1000)

shellingman@tru.ca

Fall 2024

THOMPSON RIVERS UNIVERSITY

Topics

2. Introduction
3. Some Solutions
4. Linearity
5. Independence
6. Homoscedasticity
7. Normality
8. Regression Outliers
9. Exercises and References

Introduction

- We have covered the assumptions and diagnostics for linear regression models.
- In practice, violations of these assumptions are common.
- If we want to **make meaningful inferences** from the models we need methods to address any violations.

Some Solutions

Example 1

- Using the *wage2.csv* data from Linear Regression II:

  1. Estimate the following linear regression model: *wage* (Y) with *age* (X_1), *iq* (X_2), *educ* (X_3), *meduc* (X_4), *hours* (X_5), *tenure* (X_6), *kww* (X_7), *feduc* (X_8), and *exper* (X_9) as explanatory variables.

  2. Check the validity of the linear regression assumptions.

Linearity

1. Apply a nonlinear transformation to the dependent and/or the independent variables.
   - Logarithmic transformations only to the dependent/response variable: *Assumes that the response grows/decays exponentially as a function of the independent variables.*
   - Logarithmic transformation of both the dependent and the independent variables implies that the effects are multiplicative rather than additive. *small percentage change in one of the independent variables induces a proportional percentage change in the expected value of the dependent variable*

2. Consider adding *nonlinear* functions of the explanatory variables X_j^z
   - Can use visuals (scatterplots with LOWESS curves to identify degree of polynomial)

3. Identify missing variables (including interactions) and add them to your regression model.

Example 2

- Using the variables from the model in Example 1:

  1. Load the required function to your workspace.
  2. Examine the scatterplots for any polynomials.
  3. Examine the scatterplots for any heteroscedasticity.
  4. What steps could we take to improve the model?

Independence

1. Examine the Variance Inflation Factors (VIF) and consider removing one or more of the variables identified as having a high VIF.
    - Consider removing when the result of `vif()`: $GVIF \wedge (1/(2 * Df)) > 5$
2. Re-consider any transformations that you have already made to your data.
    - Transformations *may* actually increase the chance of violating the independence assumption.
3. If your data contain time-dependent variables, consider time-series/econometric models.
    - May be able to include lagged values of dependent variable in your model (not generally encouraged).

Example 3

- Using the model in Example 1:
  1. Check the variance inflation factors of your model.
  2. Change the dependent variable *wage* to *lwage*.
  3. Do these values change by a lot?

Homoscedasticity

1. Box-Cox Power Transformation
   - If the variance increases with the mean, choose $\lambda < 1$.
   - If the variance decreases as the mean increases, choose $\lambda > 1$
2. Weighted Least Squares (If the variance is not constant and is **not** related to the factor-level means we can use weighted least squares.)
   - A weight is assigned to each observation based on the variability.

     $W_{ii} = \frac{1}{\sigma_i^2}$
3. Try modelling with more complex regression models
   - *We will cover some of these models next term*

Box-Cox Transformation in R

- We can use the *MASS* package to obtain the *best* λ.
- In R:
  - `bc <- boxcox(model)`
  - `lambda <- bc$x[which.max(bc$y)]`

    To transform your data:
  - `transform.df <- transform(df, variable = variable^(lambda))`
- Then you can estimate your model again.
- *The Box-Cox transformation can usually help with the normality assumption.*

Weighted Least Squares in R

-
  ```
  model <- lm(response ∼ explanatory1 + ..., data = data)
  # Estimate Model
  ```
- **Perform diagnostics to determine this is needed.**

-
  ```
  wt <- 1/lm(abs(model$residuals) ∼
      model$fitted.values)$fitted.values∧2 # Obtain
  weights
  ```

-
  ```
  wls_model <- lm(response ∼ explanatory1 + ...,
      data = data, weights = wt) # Estimate WLS model
  ```

-
  ```
  summary(wls_model) # Check results
  ```
- *Be sure to perform diagnostics on your new model*

Example 4

- *We know that the model from Example 1 violates the homoscedasticity assumption*:

    1. Use the Box-Cox Power Transformation to estimate the model.
        - Did it help?
    2. Use the WLS method to estimate the model.
        - Did it help?
    3. Which method do you prefer?

Normality

1. Apply a nonlinear transformation to the dependent and/or the independent variables.
    - *The Box-Cox transformation can usually help with the normality assumption.*
2. Examine your data for outliers.
    - Sometimes, outlying observations may contribute heavily to a violation of this assumption.
3. Try modelling with more complex regression models
    - *We will cover some of these models later in the course*

Outliers in Linear Regression

- **Regression outliers** are those observations whose values (of the response and explanatory variables) deviate from the regression relationship which holds for the majority of observations.
- **Cook's distance** is a measure of influence which measures the difference between the fitted values in the model with all the observations and the model with observation *i* removed.
  - We can plot in R: `plot(model,3)` AND `plot(model,4)`
  - To get the numeric values in R: `cooks.distance(model)`
- The `plot(model)` visualizations in R usually do a good job of identifying possible outliers.

Removing Outliers in R

- Can just remove specific rows by number: `data[-c(1,4,6),]`
- By name: `rows.to.remove <- c("row1","row2")`
  data[!(row.names(data) %in% rows.to.remove),]
- OR, you can save the Cook's distance as a variable and `filter` rows based on a threshold.

Example 5

1. Did the Box-Cox transformation correct the normality assumption violation in Example 1?

2. Check the model estimated in Example 1 for potential regression outliers.

3. Remove any identified regression outliers and re-estimate the model.

Example 6

1. With the tools you now possess, estimate a linear regression model for the wage variable that does not violate any of the assumptions.

2. **Interpret the estimated model results.**

Exercise 1

- Using the *Cars93* data estimate a linear regression model that explains the average price of a car (*price*). Be sure to try potential interactions and make corrections to violated assumptions as needed.

Exercise 2

- Take some time to estimate some multiple linear regression models using numeric variables from your project data (keep in mind that causality matters). Complete all of the required diagnostics, take steps to correct your model as required, comment on the estimates, and comment on the fit ($\bar{R}^2$).

References & Resources

1. Evans, J. R., Olson, D. L., & Olson, D. L. (2007). *Statistics, data analysis, and decision modeling.* Upper Saddle River, NJ: Pearson/Prentice Hall.

2. Devore, J. L., Berk, K. N., & Carlton, M. A. (2012). *Modern mathematical statistics with applications (Second Edition).* New York: Springer.

- `vif()`
- `boxcox()`
- `cooks.distance()`