Assessment of Model Fit in Linear Regression

Dr. Alex Marsella

2024-03-11

Three Learning Goals for Today

The basics of:

- Assessing a model's fit.
- Assessing the errors (residuals) for normality.
- Checking for multicollinearity.

Advanced mathematical/econometric understanding of this is moreso for a higher level class, but I will teach you the basic applications today and how to code them in R.

Understanding our error term.

- Residuals, the difference between the observed and predicted values, are key to evaluating model performance.
- Recall that $y - \hat{y} = \epsilon$
- That error term in our model: $Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \epsilon$ is assumed to be
  1. Normally distributed around 0 for all values of Y.
  2. with constant variance.
- In other words, most residuals should be near zero (point 1), and the size of the residuals should not change across different values of y (point 2).

Exercise: Residuals Plot

- Let's create a residuals plot using the Boston housing dataset.

The Boston Dataset

```r
# Overview of the Boston dataset
```

Simple Linear Regression Model

```r
# Fitting a multiple regression model
```

Creating a Residual Plot

- Note that this residual plot we create will plot fitted values (values on the regression line) against their error.
- We will not be plotting a regular scatterplot since **we have multiple independent variables**.
  - You could easily plot a meaningful (regular X-Y) scatterplot if we had only one X.

Reading Residual Plots

- A well-fitted model shows a random spread of residuals with an average of 0.
  - Don't worry about the "spread" on the x-axis, we interpret the y-axis on this plot.
- Most Y values should be close to 0, few should be far.
- At each value of X, the Y values should be similarly distant from the red line AND *should usually not be more spread out than they are for other values of X*.

Understanding Multicollinearity with VIF

- Multicollinearity is when multiple independent variables in a model are highly correlated (positive or negative).
- Multicollinearity **inflated coefficients and can undermine statistical significance**
- We use the Variance Inflation Factor (VIF) to quantify multicollinearity.
  - Quantifies the inflation of the coefficients caused by multicollinearity.

Detecting Multicollinearity in Boston Dataset

```
# Calculating VIF for the model
```

VIF Interpretation

- VIF values greater than 4 suggest potential multicollinearity.
  - We can then correlate that variable with the other variables to see how serious it is.
  - High VIF values require model reassessment and possible adjustments.
  - If one variable is highly correlated with others, it may be redundant and problematic to keep it.
  - It's up to us to think about why these correlations exist and if it makes sense to exclude a variable.
- VIF values greater than 10 mean there is serious multicollinearity.
  - You must fix this, in general, it's not up to your discretion like in the case of VIF > 4.
  - This will happen if variables are close to perfectly correlated.
    - e.g: A model of life expectancy where you include hourly wage AND yearly earnings. Obviously these are going to be almost perfectly correlated

Exercise on your own

Load in any data you want

Info on this here: http://www.sthda.com/english/wiki/r-built-in-data-sets#list-of-pre-loaded-data

Fit a multiple regression model of your own design

Plot the fitted values against the residuals and analyze it

Calculate the VIF and determine if there is any problematic multicollinearity.