Advanced Statistics
DS2003 (BDS-4A)
Lecture 22

Instructor: Dr. Syed Mohammad Irteza

Assistant Professor, Department of Computer Science, FAST

12 May, 2022

Previous Lecture

- More in detail workings of multiple linear regression
  - Calculating the slopes for each independent variable (X_1, X_2, …, X_n)
  - Calculating the intercept (b_0)
- Preference for no multicollinearity
  - We don’t want correlation between different explanatory (independent) variables
  - We want to witness a correlation between explanatory variables with the response (dependent) variable

Today

- More on multicollinearity
  - We don’t want correlation between different explanatory (independent) variables
  - We want to witness a correlation between explanatory variables with the response (dependent) variable
- R^2 measure

Adding More Explanatory Variables

- Adding more independent (or explanatory) variables to a multiple linear regression does not mean the regression will be “better” or offer better predictions; in fact, it can make things worse.
  - This is called *overfitting*
- The addition of more independent variables creates more relationships among them
  - So not only are independent variables related to the explanatory (or dependent) variables
  - Explanatory variables may potentially be related to each other
  - When this happens, it is known as *multi-collinearity*

Interpreting a Linear Regression Model

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
- Y = quantity demanded of a commodity
- X_1 = price of the commodity
- X_2 = consumer income

Interpreting a Linear Regression Model

- $Y = 27 + 9X_1 + 12X_2$
- Y = Predicted Sales ($, thousands)
- X_1 = Capital Expenditure ($, thousands)
- X_2 = Marketing Expenditure ($, thousands)

Multicollinearity

- Multicollinearity occurs when independent variables in a *regression* model are correlated. This *correlation* is a problem because independent variables should be independent. If the degree of correlation between variables is high enough, it can cause problems when you fit the model and interpret the results.

Why is Multicollinearity a Potential Problem?

- A key goal of regression analysis is to isolate the relationship between each independent variable and the dependent variable.
- The interpretation of a regression coefficient is that it represents the mean change in the dependent variable for each 1 unit change in an independent variable when you hold all of the other independent variables constant.

Types of Multicollinearity

- ***Structural***: This type occurs when we create a model term using other terms.
  - In other words, it's a byproduct of the model that we specify rather than being present in the data itself.
  - For example, if you square term X to model curvature, clearly there is a correlation between X and X^2.
- ***Data***: This type of multicollinearity is present in the data itself rather than being an artifact of our model.
  - Observational experiments are more likely to exhibit this kind of multicollinearity.

Do I Have to Fix Multicollinearity?

- Multicollinearity makes it hard to interpret your coefficients, and it reduces the power of your model to identify independent variables that are statistically significant.
- These are definitely serious problems. However, the *good news* is that you don’t always have to find a way to fix multicollinearity.

Keep three points in mind

- The severity of the problems increases with the degree of the multicollinearity.
  - Therefore, if you have only moderate multicollinearity, you may not need to resolve it.
- Multicollinearity affects only the specific independent variables that are correlated.
  - Therefore, if multicollinearity is not present for the independent variables that you are particularly interested in, you may not need to resolve it.
- Multicollinearity affects the *coefficients* and *p-values*, but it does not influence the predictions, precision of the predictions, and the goodness-of-fit statistics.
  - If your primary goal is to make predictions, and you don’t need to understand the role of each independent variable, you don’t need to reduce severe multicollinearity.

Testing for Multicollinearity with Variance Inflation Factors (VIF)

- If you can identify which variables are affected by multicollinearity and the strength of the correlation, you're well on your way to determining whether you need to fix it.
- Fortunately, there is a very simple test to assess multicollinearity in your regression model.
- The variance inflation factor (VIF) identifies correlation between independent variables and the strength of that correlation.

Testing for Multicollinearity with Variance Inflation Factors (VIF)

- Statistical software calculates a VIF for each independent variable. VIFs start at 1 and have no upper limit.
  - A *value of 1 indicates* that there is no correlation between this independent variable and any others.
  - *VIFs between 1 and 5 suggest* that there is a moderate correlation, but it is not severe enough to warrant corrective measures.
  - *VIFs greater than 5* represent critical levels of multicollinearity where the coefficients are poorly estimated, and the p-values are questionable.

Regression Analysis: Femoral Neck versus %Fat, Weight kg, Activity

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	0.555785	0.138946	27.95	0.000
%Fat	1	0.009240	0.009240	1.86	0.176
Weight kg	1	0.127942	0.127942	25.73	0.000
Activity	1	0.047027	0.047027	9.46	0.003
%Fat*Weight kg	1	0.041745	0.041745	8.40	0.005
Error	87	0.432557	0.004972		
Total	91	0.988342			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0705118	56.23%	54.22%	50.48%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.155	0.132	1.18	0.243	
%Fat	0.00557	0.00409	1.36	0.176	14.93
Weight kg	0.01447	0.00285	5.07	0.000	33.95
Activity	0.000022	0.000007	3.08	0.003	1.05
%Fat*Weight kg	-0.000214	0.000074	-2.90	0.005	75.06

Useful Links & Resources

- **Source**:
  - https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/

- **Reference**:
  - openintro.org/os (Chapter 9, Section 9.1)