Multiple linear regression assumptions and multicollinearity

In prior videos, you have learned about linear regression assumptions. In this reading, you will build off that knowledge base to extend your understanding of multiple linear regression assumptions. This reading will help you review assumptions that apply to both simple linear regression and multiple linear regression, and will then focus more heavily on the concept of multicollinearity.

Multiple linear regression assumptions

Recall that simple linear regression has four main assumptions that provide validity to the results derived from the analysis. To this list of four assumptions, we add the no multicollinearity assumption when working with multiple linear regression.

1. **Linearity:** Each predictor variable (X_i) is linearly related to the outcome variable (Y).
2. **(Multivariate) normality:** The errors are normally distributed.*
3. **Independent observations:** Each observation in the dataset is independent.
4. **Homoscedasticity:** The variation of the errors is constant or similar across the model.*
5. **No multicollinearity:** No two independent variables (X_i and X_j) can be highly correlated with each other.

***Note on errors and residuals**

As noted earlier, "residuals" and "errors" are sometimes used interchangeably, but there is a difference. We use residuals to estimate errors when we are checking the normality and homoscedasticity assumptions of linear regression.

- **Residuals** are the difference between the predicted and observed values. You can calculate residuals after you build a regression model by subtracting the predicted values from the observed values.
- **Errors** are the natural noise assumed to be in the model.

Extending prior assumptions

Much of what you learned about the first four assumptions with regard to simple linear regression can be directly applied to multiple linear regression. The code might be slightly different or longer, but the rationale is the same.

Linearity

- With multiple linear regression, you need to consider whether each *x* variable has a linear relationship with the *y* variable.
- You can make multiple scatterplots instead of just one, using seaborn's `pairplot` function, or the `scatterplot` function multiple times. Other libraries with plotting capabilities will have similar functions.

Independent observations

- The independent observations assumption is still primarily focused on data collection.
- You can check the validity of the assumption in the same way you would with simple linear regression.

(Multivariate) Normality

- Just as with simple linear regression, you can construct the model, and then create a Q-Q plot of the residuals.
- If you observe a straight diagonal line on the Q-Q plot, then you can proceed in your analysis. You can also plot a histogram of the residuals and check if you observe a normal distribution that way.

Homoscedasticity

- As with simple linear regression, for multiple linear regression, just create a plot of the residuals vs. fitted values.
- If the data points seem to be scattered randomly across the line where residuals equal 0, then you can proceed.

How to check the no multicollinearity assumption

The no multicollinearity assumption is unique to multiple linear regression as it focuses on potential relationships between different independent (X) variables. When assessing the no multicollinearity assumption, you're interested in identifying any linear relationships between the independent (X) variables. X variables that are linearly related could muddle the interpretation of the model's results. If there are X variables that are linearly related, it is usually best to remove some independent variables from the model.

There are a few ways to check the no multicollinearity assumption. This reading will cover two of them. One is purely visual, and the other is numerical in nature. Both can be done prior to building the linear regression model.

Scatterplots or Scatterplot Matrix

A visual way to identify multicollinearity between independent (X) variables is using scatterplots or scatterplot matrices. The process is the same as when you checked the linearity assumption, except now you're just focusing on the X variables, not the relationship between the X variables and the Y variable. If you're using the seaborn library, you can use the `pairplot` function, or the `scatterplot` function multiple times.

Variance Inflation Factors (VIF)

Calculating the variance inflation factor, or VIF, for each independent (X) variable is a way to quantify how much the variance of each variable is "inflated" due to correlation with other X variables. You can read more about VIFs on the Pennsylvania State University's Eberly College of Science website or on the website for Vilnius University's e-book on *Practical Econometrics and Data Science*.

To calculate the VIF for each predictor variable, you can use the `variance_inflation_factor` function from the statsmodels package via the following lines of code.

```
from statsmodels.stats.outliers_influence import variance_inflation_factor
vif["VIF"] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
```

The smallest value a VIF can take on is 1, which would indicate 0 correlation between the X variable in question and the other predictor variables in the model. A high VIF, such as 5 and above, according to the statsmodels documentation, can indicate the presence of multicollinearity.

What to do if there is multicollinearity in your model

Variable Selection

The easiest way to handle multicollinearity is simply to only use a subset of dependent variables in your model. For example, if your multiple linear regression model is something like this:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

But if X_1 and X_3 are highly correlated, then you can choose to include only X_1 or X_3 in your final model, but not both.

There are a few specific statistical techniques you can use to select variables strategically. You'll learn about these more in future videos:

- Forward selection
- Backward elimination

Advanced Techniques

In addition to the techniques listed above that will be covered in -depth in this course, there are more advanced techniques that you may come across in your career as a data professional, such as:

- Ridge regression
- Lasso regression
- Principal component analysis (PCA)

These techniques can result in more accurate and predictive models, but can complicate the interpretation of regression results.

Key Takeaways

- Many of the assumptions of simple linear regression extend readily to multiple linear regression.
- You can use scatterplots and variance inflation factors to check for multicollinearity in a regression model.
- There are different techniques for variable selection to remove multicollinearity in a model.