DATA SCIENCE WITH R

REGRESSION ANALYSIS

Overview

Simple Linear Regression

Multiple Linear Regression

Regression Assumptions

Implementation in SAS

Regression Assumptions

1. Model is *linear in parameters*

Regression Assumptions

1. Model is *linear in parameters*

2. The data are a *random sample* of the population

Regression Assumptions

1. Model is *linear in parameters*

2. The data are a *random sample* of the population
   – The errors are *statistically independent* from one another

Regression Assumptions

1. Model is *linear in parameters*

2. The data are a *random sample* of the population
   – The errors are *statistically independent* from one another

3. The expected value of the errors is always zero

Regression Assumptions

1. Model is *linear in parameters*

2. The data are a *random sample* of the population
   - The errors are *statistically independent* from one another

3. The expected value of the errors is always zero

4. The independent variables are not too strongly *collinear*

Regression Assumptions

1. Model is *linear in parameters*
2. The data are a *random sample* of the population
   - The errors are *statistically independent* from one another
3. The expected value of the errors is always zero
4. The independent variables are not too strongly *collinear*
5. The independent variables are measured *precisely*

Regression Assumptions

1. Model is *linear in parameters*
2. The data are a *random sample* of the population
   - The errors are *statistically independent* from one another
3. The expected value of the errors is always zero
4. The independent variables are not too strongly *collinear*
5. The independent variables are measured *precisely*
6. The residuals have *constant variance*

Regression Assumptions

7. The errors are normally distributed

Regression Assumptions

7. The errors are normally distributed

8. The model is correctly specified

Regression Assumptions

7. The errors are normally distributed

8. The model is correctly specified

If all these conditions hold, then OLS estimators are **BLUE** –

Regression Assumptions

7. The errors are normally distributed

8. The model is correctly specified

If all these conditions hold, then OLS estimators are **BLUE** – Best Linear Unbiased Estimators

Regression Assumptions

In real life, all the conditions may not be met.

Regression Assumptions

In real life, all the conditions may not be met.

We need to:

1. Check if the assumptions are holding up

Regression Assumptions

In real life, all the conditions may not be met.

We need to:

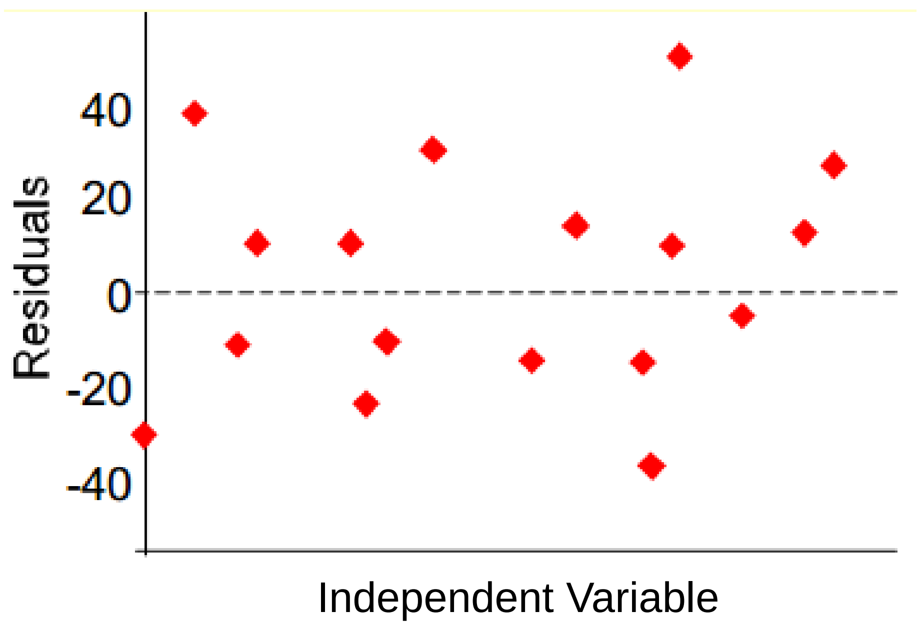
1. Check if the assumptions are holding up
2. If not, assess how to correct for violations

Regression Results: Assumptions

CHECKING IF ASSUMPTIONS ARE VALID

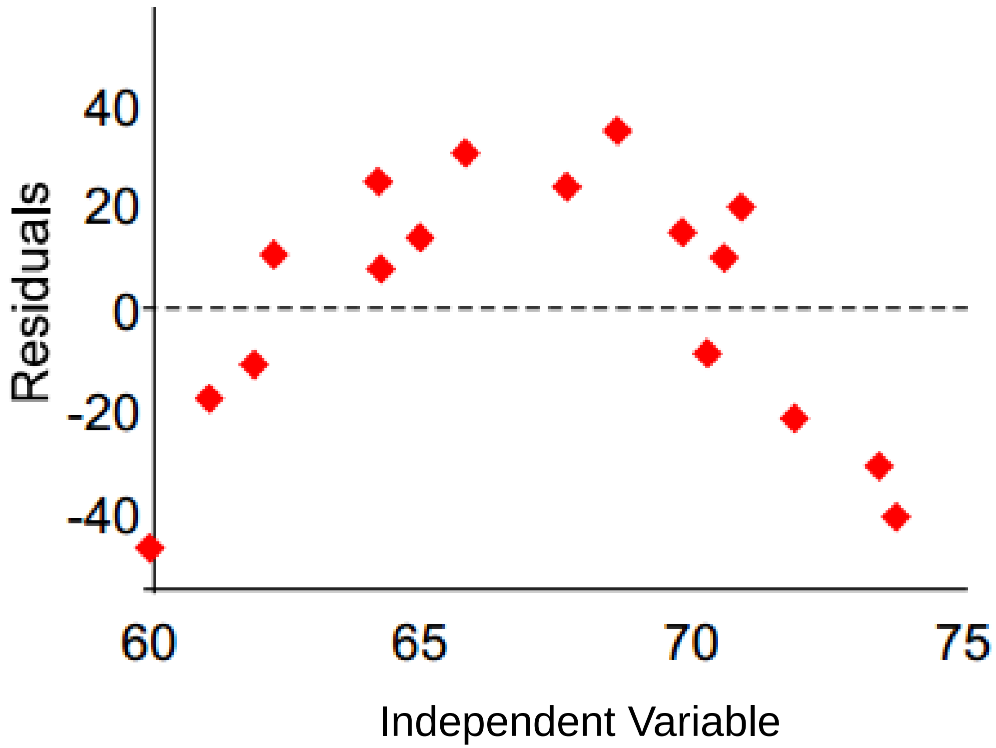
1. **Check for linearity – plot the residuals against each IV**

   - If data is linearly related, we should see no pattern in the plot



Regression Results: Assumptions

If relationship is non–linear?



Regression Results: Assumptions

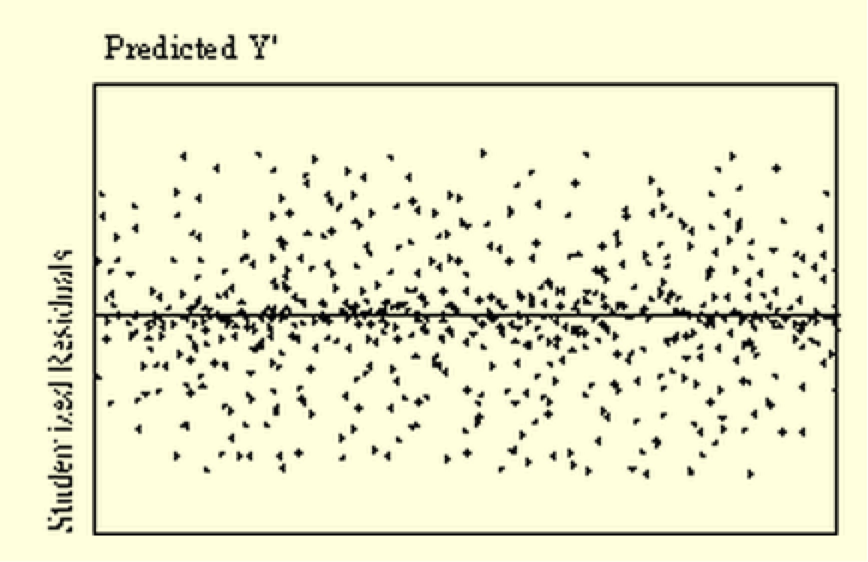
2. **The residuals should have constant variance – homoscedasticity**

    - Plot the residuals against Predicted Y

Regression Results: Assumptions

2. **The residuals should have constant variance – homoscedasticity**

   - Plot the residuals against Predicted Y

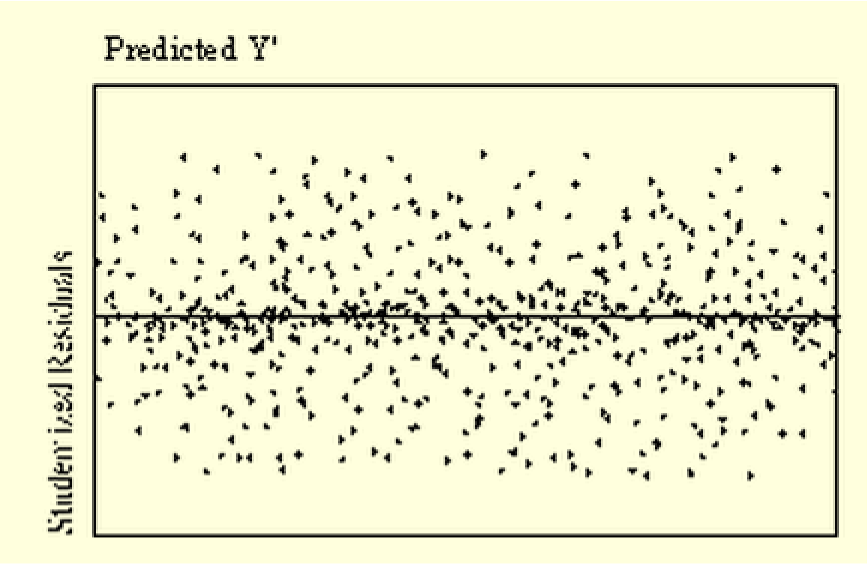


Homoscedasticity

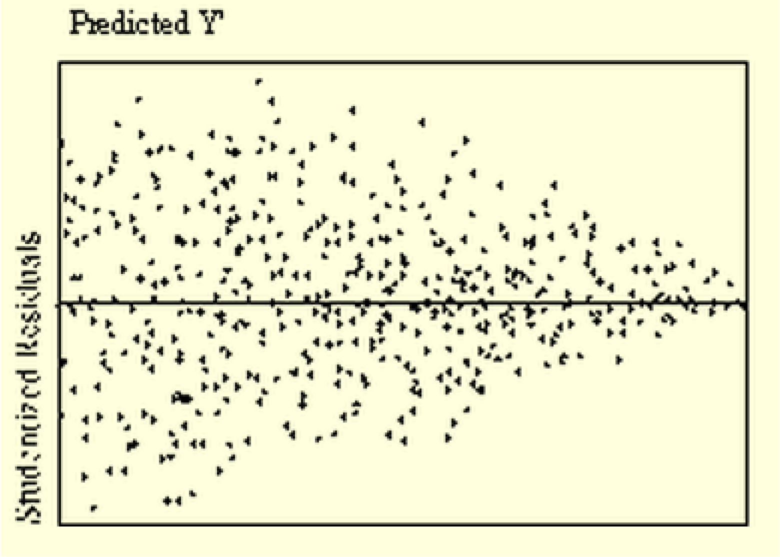
Regression Results: Assumptions

2. The residuals should have constant variance – homoscedasticity

- Plot the residuals against Predicted Y



Homoscedasticity



Heteroscedasticity

Regression Results: Assumptions

If errors are heteroscedastic?

Regression Results: Assumptions

If errors are heteroscedastic?

- The presence of heteroscedasticity does not imply bias in the estimates

Regression Results: Assumptions

If errors are heteroscedastic?

- The presence of heteroscedasticity does not imply bias in the estimates
- Heteroscedasticty leads to bias in the standard errors, leading to issues with hypothesis testing and confidence intervals

Regression Results: Assumptions

If errors are heteroscedastic?

- The presence of heteroscedasticity does not imply bias in the estimates
- Heteroscedasticty leads to bias in the standard errors, leading to issues with hypothesis testing and confidence intervals
  - Std error is a measure of variance, and therefore if standard errors are biased, then hypothesis test results will be biased leading to wrong inferences

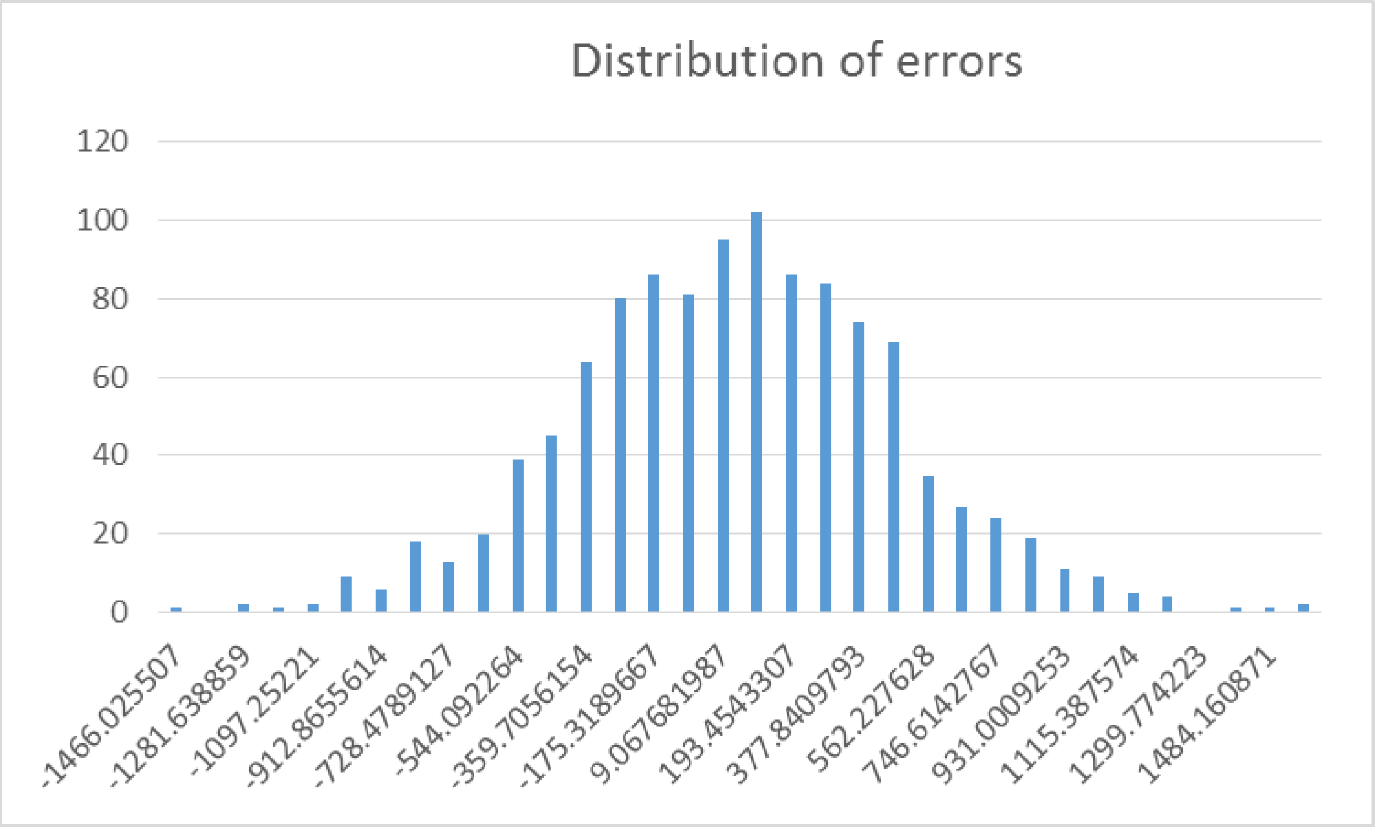
Regression Results: Assumptions

3. **The residuals are normally distributed**

Regression Results: Assumptions

3. The residuals are normally distributed

- Histogram or probability plot for the residuals



Regression Results: Assumptions

If residuals are not normally distributed?

Regression Results: Assumptions

If residuals are not normally distributed?

Hypothesis test outcomes may be invalid, though less of an issue with large samples

Regression Results: Assumptions

4. **The IVs are not too correlated - multicollinearity**

    - The IVs should not be highly correlated to one another

Regression Results: Assumptions

4. The IVs are not too correlated - multicollinearity

- The IVs should not be highly correlated to one another
- Check pairwise correlations, or generate VIF

Regression Results: Assumptions

4. If some of the IVs are highly correlated?

Regression Results: Assumptions

4. **If some of the IVs are highly correlated?**

   - Estimates are not biased, but the standard errors are inflated, leading to misleading hypothesis test outcomes –

Regression Results: Assumptions

4. If some of the IVs are highly correlated?

- Estimates are not biased, but the standard errors are inflated, leading to misleading hypothesis test outcomes –

  Either drop the correlated variables, or combine them

Multiple Linear Regression

MODELING TECHNIQUES

Running a model given data is an easy task given that all the computation is done via SAS or Excel

Multiple Linear Regression

MODELING TECHNIQUES

Running a model given data is an easy task given that all the computation is done via SAS or Excel

The skill of an analyst lies in generating the right model to understand and solve for the business issue at hand

Multiple Linear Regression

MODELING TECHNIQUES

Running a model given data is an easy task given that all the computation is done via SAS or Excel

The skill of an analyst lies in generating the right model to understand and solve for the business issue at hand

The first model, or naïve model that is generated from data is usually used as a starting point

Multiple Linear Regression

MODELING TECHNIQUES

Running a model given data is an easy task given that all the computation is done via SAS or Excel

The skill of an analyst lies in generating the right model to understand and solve for the business issue at hand

The first model, or naïve model that is generated from data is usually used as a starting point

- Depending on domain understanding and modeling techniques knowledge, typically many models are run before arriving at a final model

Multiple Linear Regression

MODELING TECHNIQUES

Running a model given data is an easy task given that all the computation is done via SAS or Excel

The skill of an analyst lies in generating the right model to understand and solve for the business issue at hand

The first model, or naïve model that is generated from data is usually used as a starting point

- Depending on domain understanding and modeling techniques knowledge, typically many models are run before arriving at a final model

How can multiple models be run using the same data?

Multiple Linear Regression

MODELING TECHNIQUES

Step-wise Regression

It may be useful to run models adding (or removing) one variable at a time

Multiple Linear Regression

MODELING TECHNIQUES

Step-wise Regression

It may be useful to run models adding (or removing) one variable at a time

Two types of step-wise regressions:

Multiple Linear Regression

MODELING TECHNIQUES

Step-wise Regression

It may be useful to run models adding (or removing) one variable at a time

Two types of step-wise regressions:

- Forward – Add one variable at a time

Multiple Linear Regression

MODELING TECHNIQUES

Step-wise Regression

It may be useful to run models adding (or removing) one variable at a time

Two types of step-wise regressions:

- Forward – Add one variable at a time
- Backward – Remove one variable at a time

Multiple Linear Regression

MODELING TECHNIQUES

Forward Step-wise regression

Variables are added one at a time until the model cannot be significantly improved by adding another variable

- Note that the variable order we use to add has an impact, so multiple step-wise forward regression models could be run before arriving at a best model

Backward Step-wise regression

This approach is the reverse, where we start with a model that has all explanatory variables, and variables are dropped one by one based on p-value (highest p-value dropped first).

- Re-run model without the variable dropped, and then drop next variable with highest p-value. Continue till no other variables can be dropped based on a pre-determined cut-off value (5%, 10%)

Multiple Linear Regression

MODELING TECHNIQUES

Other modeling techniques could include:

Multiple Linear Regression

MODELING TECHNIQUES

Other modeling techniques could include:

- Transforming variables – use log transformation for example

Multiple Linear Regression

MODELING TECHNIQUES

Other modeling techniques could include:

- Transforming variables – use log transformation for example
- Creating interaction variables – trying to capture impact of variable A and Variable B together

Multiple Linear Regression

MODELING TECHNIQUES

Other modeling techniques could include:

- Transforming variables – use log transformation for example
- Creating interaction variables – trying to capture impact of variable A and Variable B together
- Aggregating or disaggregating variables – Adding up marketing, or disaggregating promotions

Multiple Linear Regression

MODELING TECHNIQUES

Other modeling techniques could include:

- Transforming variables – use log transformation for example
- Creating interaction variables – trying to capture impact of variable A and Variable B together
- Aggregating or disaggregating variables – Adding up marketing, or disaggregating promotions
- Creating stock variables

Recap

- Simple Linear Regression
- Multiple Linear Regression
- Assumptions

Coming Up

Regression Analysis

Case Study

THANK YOU