# REGRESSION & PREDICTIONS

Madhurima Nath

Data Scientist, Slalom

# Contents

# Linear Regression

- establishes a linear relationship between the predictor variable(s) and the response variable
- estimates the value of the dependent variable *y*, when only the predictors, *x* are known
- given *n* data points, the linear model with *p*-vector of predictors *x* is

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ \ldots + \beta_p \ x_{ip} + \epsilon_i$$

In matrix notation, $\boldsymbol{y} = \boldsymbol{X}\,\boldsymbol{\beta} + \ \boldsymbol{\epsilon}$

- best fitted model – one which minimizes the prediction errors (least squared errors)

# Assumptions of Linear Regression

- errors/residuals, $\epsilon_i$ are identically and independently distributed

- Errors/residuals, $\epsilon_i$ are normally distributed, with mean 0 and equal (unknown) variance - homoscedasticity.

# Simple Linear Regression

- best fitted line: straight line with regression coefficients – intercept and slope
- $y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, \dots, n$
- regression coefficients are:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- $\hat{\beta}_0$ tells estimated equation at $x = 0$
- $\hat{\beta}_1$: amount by which mean response vary for a unit increase in $x$.

# Multiple Linear Regression

- generalized version of simple regression, more than one predictor used
- solution:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\,\boldsymbol{y}$$

- generally, all real-world problems involve multiple predictors

# Real-world Example

**Data: New York air quality data in R** (details about data)

- Ozone: Mean ozone in parts per billion from 1300 to 1500 hours at Roosevelt Island
- Solar.R: Solar radiation in Langleys in the frequency band 4000–7700 Angstroms from 0800 to 1200 hours at Central Park
- Wind: Average wind sped in miles per hour at 0700 and 1000 hours at LaGuardia Airport
- Temp: Maximum daily temperature in degrees Fahrenheit at LaGuardia Airport
- Month: Numeric value of Month (1–12)
- Day: Day of month (1–31).

Our target → attribute 'ozone'.

# R notebook

# Factor Variables

- Categorical variables
- Both numeric and character variables can be made into factors
- 'Month' in our dataset
- Ordered factor variable: when the categories of the factor variables have a particular order, and it is used to perform comparisons

**Ways to handle Factor Variables**:

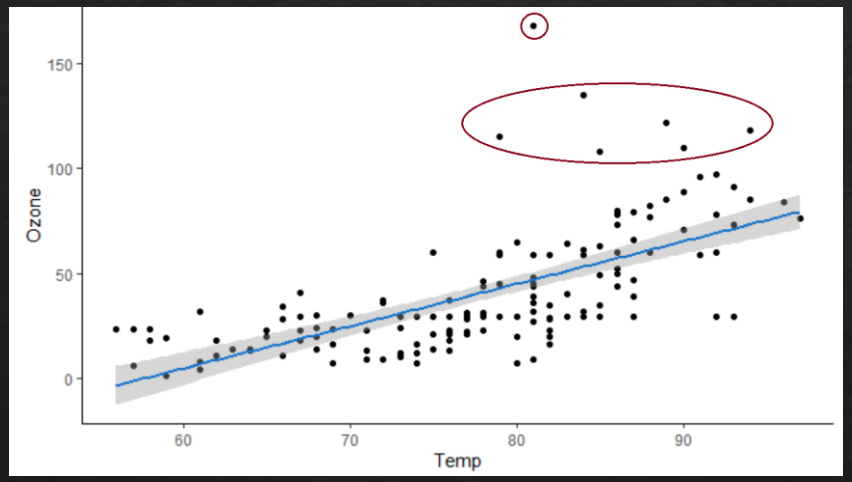
- Dummy variables: 0 or 1, creates *k-1* variables for *k* categories, e.g.: Jan: 1 if Jan, 0 otherwise
- One hot encoding: 0 or 1, creates *k* variables for *k* categories, e.g.:

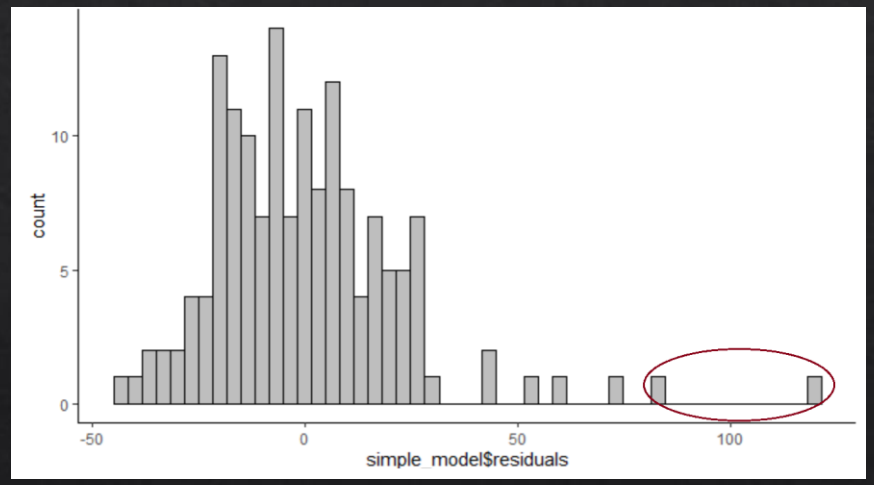
| $Jan$ | $Feb$ |
|---|---|
| 1 | 0 |
| 0 | 1 |

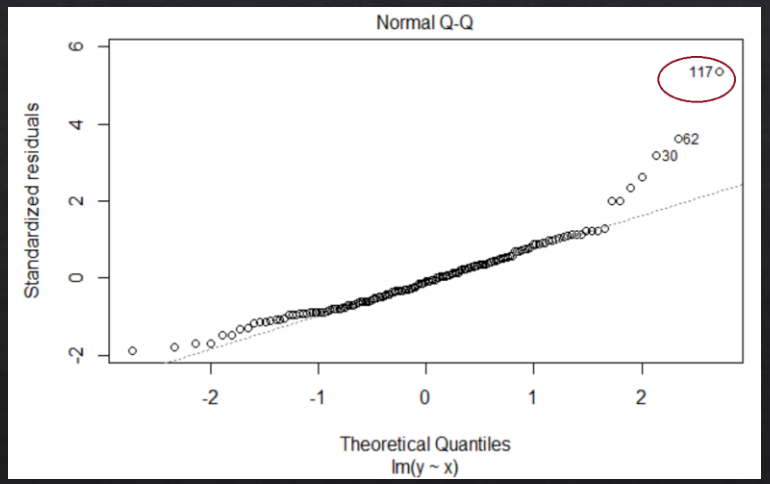
- Reference coding: a reference category is named and identified as a category of comparison for the other categories, i.e., the other categories are *compared to* the reference.

# Regression Diagnostics - Outliers

**Outliers**: an observation which has a response value that is different from the predicted values based on the model.



# Regression Diagnostics – Influential values

**Influential values**: a data point which unduly influences any part of a regression analysis, such as the predicted responses, the estimated slope coefficients, or the hypothesis test results.

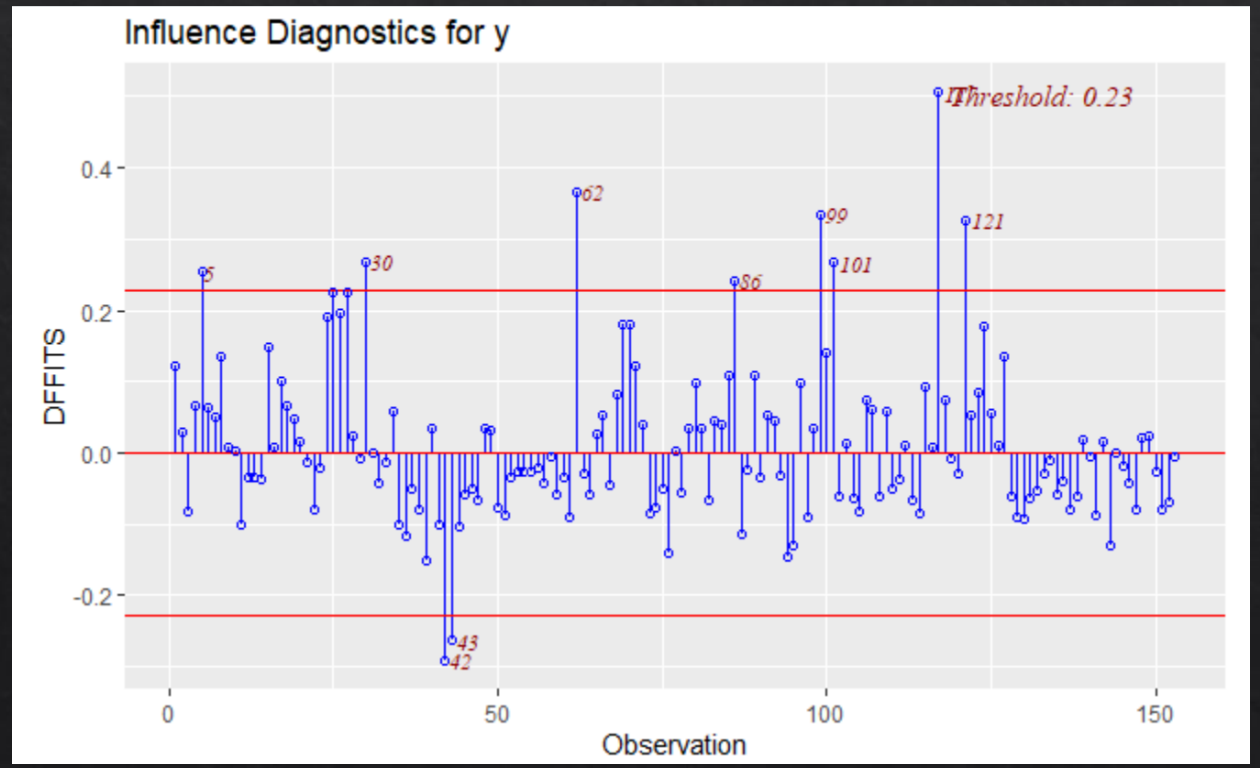
Methods to identify:

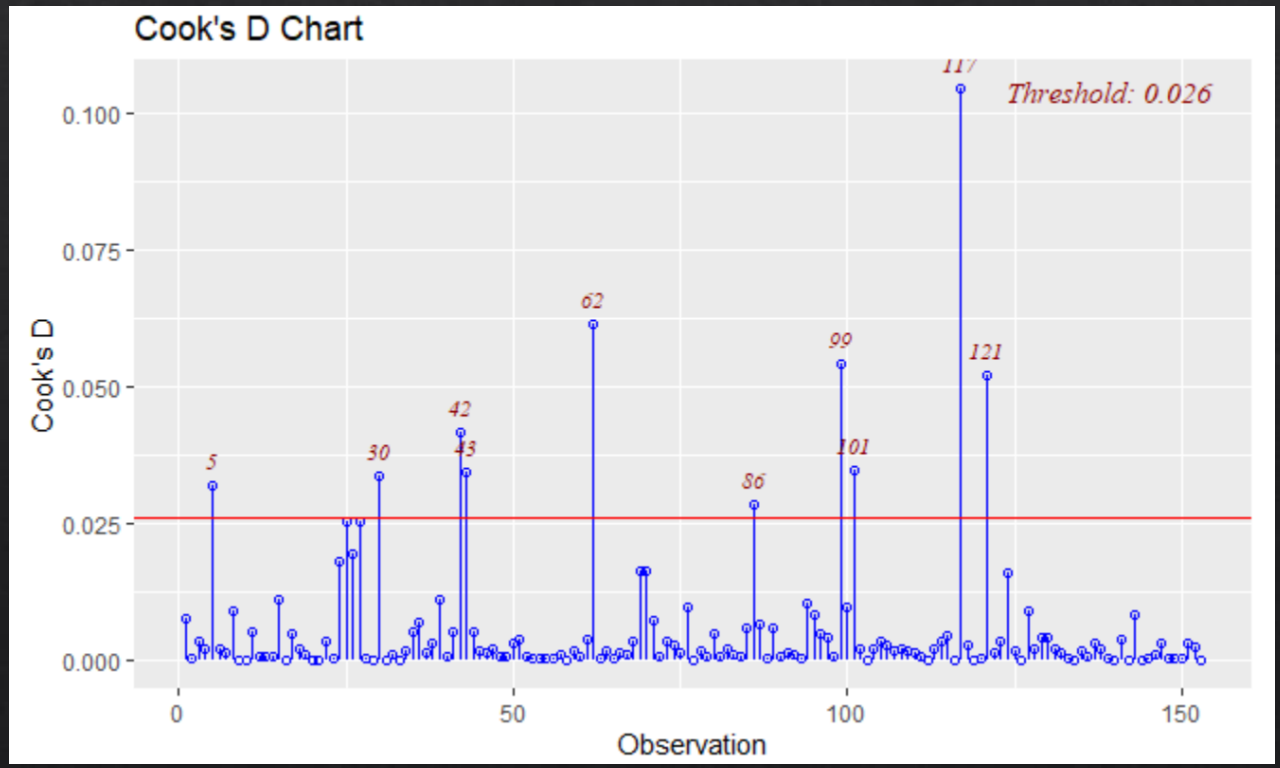
Delete the observations one at a time, refit the regression model on the remaining *n-1* observations each time. Then, compare the results using all *n* observations to the results with the $i^{th}$observation deleted to see how much influence the observation has on the analysis. This enables to assess the potential impact each data point has on the regression analysis.

1. Difference in Fits (DFFITS) – if $|DFFITS| > 2\ \frac{\sqrt{m+1}}{(n-m-1)}$, strong influencer
   ($m$: # of predictors including intercept, $n$: # of observations)

2. Cook's Distance – large Cook's distance, strong influencer

# Regression Diagnostics – Influential values

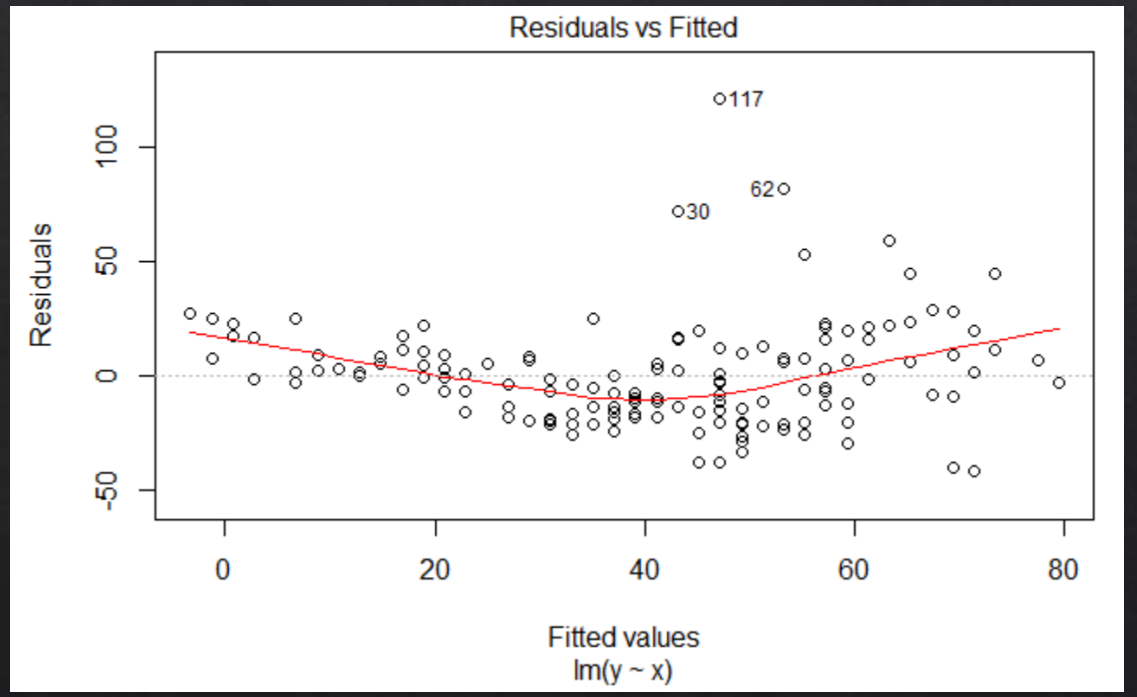


# Regression Diagnostics – Correlated errors

**Correlated errors**: The assumptions of linear regression state that the errors/residuals are uncorrelated. If correlation exists, it means that the model has not taken into account the additional information present in the data.

Easiest method to check: Plot the residuals to identify any non-random trends.



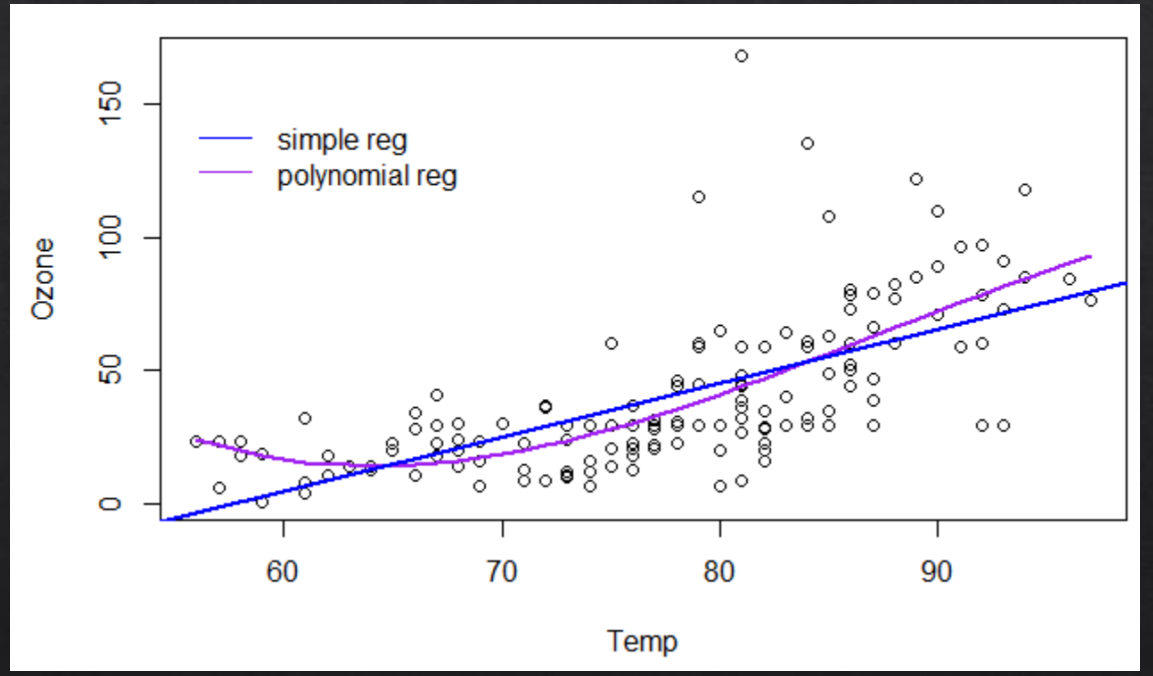
# Polynomial Regression

- Fits a polynomial model between the dependent variable, $y$ and the independent variable $x$.
- Such a model with a single predictor is

$$y_i = \beta_0 + \beta_1 x + \beta_2 x^2 + \ \ldots + \beta_h x^h + \epsilon_i$$

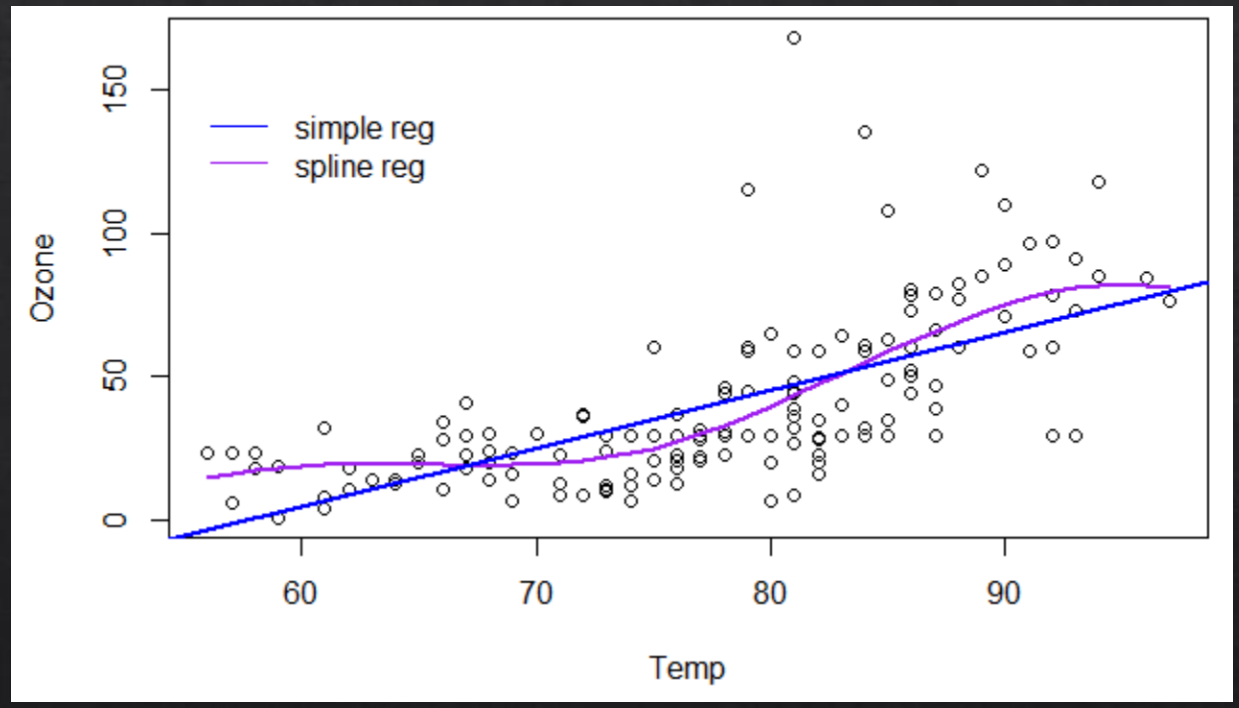
where $h$ is the degree of the polynomial, i.e., $h=2$ → quadratic, $h=3$ → cubic etc.

Caution: Polynomials are powerful tools, but they can lead to over-fitting.



# Spline Regression

- One method for testing non-linearity in the predictor variables and for modeling non-linear functions.
- A spline is a function constructed piece-wise from polynomial functions.
- Instead of building one model for the full dataset, spline regression divides the data into multiple bins and fits each bin with separate model.
- The points where the division occurs is called knots.



# Additional Points to Remember

**Multi-collinearity**: when two or more of the predictors in a regression model are moderately or highly correlated with one another.

Computing the ***Variance Inflation Factor*** (VIF) for each independent variable helps detect multi-collinearity. A variance inflation factor (*VIF*) quantifies how much the variance is inflated. It exists for each of the predictors in a multiple regression model.

**Bias-Variance Trade off (Under-fitting vs Over-fitting)**: two extreme ends of fitting a model to the data observations.

Bias: it is the error from incorrect assumptions in the learning algorithm – Under-fitting

Variance: it is the error from sensitivity to small fluctuations in the training set – Over-fitting