Regression - II

Regression Diagnostics

Outliers

Standardized residuals = $\frac{y_i - \hat{y}_i}{SE(\text{residuals})}$

- number of std dev away the point is from regressor line
- determine outlier
- can be used in anomaly detection

Influential Values

- ↓ might not be outlier
- ↳ can be determined using hat value or Cook's distance
- ↳ impart significant leverage on Regression

Partial Residual Plots to check non linearity

- partial residual$_i$ = $\underbrace{(y_i - \hat{y}_i)}_{\text{residual}} + \hat{b}_2 x_i$
- relationship between y and x_i taking other x_j into account
- If the partial residual for some x_i is non linear
  - ⇓
  - higher order x_i required

For Validity of Statistical Inference (Hypothesis test/p-val)

- Normally distributed residuals
- have same variance and are ~~independent~~

Heteroskedacity

- lack of constant variance of residuals over the complete range of values
  - may suggest incomplete model
  - may also check independance of errors as they might be correlated → there might be autocorrelations in time or space domain kind of data

Polynomial Regression

$y = b_0 + b_1 x + b_2 x^2 + e$

- ↳ modified least squares can be used to fit

Splines

- ↳ adding higher order terms to whole region may lead to undesirable wiggliness
- ↳ Piecewise Continuous Polynomials
- ↳ Connected at series of time points
  - ↳ KNOTS
- ↳ Coefficient of spline terms are not interpretable