Regression Diagnostics

estimation and inference from the regression model depends on assumptions that need to be checked using regression diagnostics

1) Linearity	relationship between X and $\bar{Y}$ is linear
2) Homoscedasticity	variance of residuals is equal for any x
3) Independence	observations are independent of each other
4) Normality	Y is normally distributed

graphical diagnostics are more flexible but harder to definitively interpret
numerical diagnostics are narrower in scope but require no intuition

model building is iterative because need to keep repeating diagnostics on a succession of models

Theoretical Random Error ($\boldsymbol{\varepsilon}$)

$$E(\boldsymbol{Y}|\boldsymbol{X}) = \boldsymbol{X\beta}$$
$$Var(\boldsymbol{Y}|\boldsymbol{X}) = \sigma^2\boldsymbol{I}$$

$$\boldsymbol{\varepsilon} = \boldsymbol{Y} - E(\boldsymbol{Y}|\boldsymbol{X}) = \boldsymbol{Y} - \boldsymbol{X\beta}$$
$$\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}|\boldsymbol{X}) = 0$$
$$Var(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \sigma^2\boldsymbol{I}$$

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$
$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{HY}$$

Residuals ($\hat{\boldsymbol{e}}$)

$$\boldsymbol{\varepsilon} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X\beta}$$
$$E(\hat{\boldsymbol{e}}|\boldsymbol{X}) = 0$$
$$Var(\hat{\boldsymbol{e}}|\boldsymbol{X}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$$
$$Var(\hat{\boldsymbol{e}}_i|\boldsymbol{X}) = \sigma^2(\boldsymbol{I} - h_{ii})$$

h_{ii} = ith diagonal element of H
cases with large values of h_{ii} have smaller values for $Var(\hat{\boldsymbol{e}}_i|\boldsymbol{X})$
regression diagnostics are based on $\hat{\boldsymbol{e}}$ because residuals are assumed to behave like $\boldsymbol{\varepsilon}$

Hat Matrix

$n \times n$ symmetrical matrix

$$\boldsymbol{HX} = \boldsymbol{X}$$
$$(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X} = 0$$
$$\boldsymbol{HH} = \boldsymbol{H}^2 = \boldsymbol{H}$$
$$Cov(\widehat{\boldsymbol{Y}}, \hat{\boldsymbol{e}}|\boldsymbol{X}) = Cov(\boldsymbol{HY}, (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}|\boldsymbol{X}) = \boldsymbol{\sigma}^2\boldsymbol{H}(\boldsymbol{I} - \boldsymbol{H}) = 0$$

$$h_{ij} = \boldsymbol{x'}_i(\boldsymbol{X'X})^{-1}\boldsymbol{x}_j = \boldsymbol{x'}_j(\boldsymbol{X'X})^{-1}\boldsymbol{x}_i = h_{ji}$$
$$\sum_{i=1}^{n} h_{ii} = p'$$
$$\sum_{i=1}^{n} h_{ij} = \sum_{j=1}^{n} h_{ji} = 1$$
$$\hat{y}_i = \sum_{j=1}^{n} h_{ji}y_j = h_{ii}y_i + \sum_{j\neq 1}^{n} h_{ji}y_j$$

p' = # parameters = # predictors + 1
h_{ii} = leverage of the i[th] case
as h_{ii} approaches 1, $\hat{y}_i$ gets closer to y_i

Checking Error Assumptions

residuals and errors are not interchangeable, $Var(\hat{\boldsymbol{e}}|\boldsymbol{X}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$ only if $Var(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \sigma^2\boldsymbol{I}$
the errors may have equal variance and be uncorrelated, but residuals may be different
errors are not observable, so regression diagnostics are applied to residuals to check assumptions about errors

Graphical Diagnostics

interpretation of plots may be ambiguous, but check to see nothing is seriously wrong with the assumptions

Linearity Assumption

on scatterplot of residuals against fitted values, residuals are randomly scattered and symmetrical around the x-axis

residuals don't follow any curvilinear patterns

Homoscedasticity Assumption

check whether the variance in the residuals is related to some other quantity, $\widehat{\boldsymbol{Y}}$ or X_i

on scatterplot of residuals against fitted value, spread of the residuals should be the same across all fitted values

on scatterplot of residuals against a predictor not in the model, curvilinear patterns indicate that that predictor should be included

including the predictor and refitting the linear model should get rid of the association between residuals and that predictor

range of residuals don't increase or decrease as fitted values increase, forming a triangle shape

Normality Assumption

Q-Q plot compares residuals to ideal normal observations

on scatterplot of quantiles of the residuals against quantiles of a standard normal distribution, residuals should follow the $y = x$ line and fall within the 95% confidence interval boundaries

histogram of residuals should show a bell curve

Numerical Diagnostics

Shapiro-Wilk Normality Test

recommended to use Shapiro-Wilk normality test in conjunction with Q-Q plot

Shapiro-Wilk normality test lacks power for small sample sizes

mild deviations from normality may be detected, but don't abandon linear model if the effects of non-normality are mitigated by large sample sizes

Shapiro and Wilk W-statistic is the square of the correlation between the observed order statistics and expected order statistics

H_0: The residuals follow a normal distribution.

H_1: The residuals do not follow a normal distribution.

Testing for Curvature

U = regressor or combination of regressor

refit the linear model with an addition regressor U^2 added and test if the β-coefficient for U^2 is 0

if U doesn't depend on estimated coefficients, then a t-test can be used

use Turkey's Test for Non-additivity to see if there's a curvature relationship between response variable and any of the predictors

Unusual Observations

outlier = observation that has a large residual, observed value for the point is much different than what is predicted by the regression model

leverage point = observation whose x-value is far away from $\bar{x}$

influential point = observation that substantially changes the slope of the line, thus having a large influence on the fit of the model

to find influential points, compare the fit of the model with and without each observation

Leverage Points

h_{ii} = leverage

leverage only depends on X, not Y

$$Var(\hat{\boldsymbol{e}}_i|\boldsymbol{X}) = \sigma^2(\boldsymbol{I} - h_{ii})$$

a large leverage will make $Var(\hat{\boldsymbol{e}}_i|\boldsymbol{X})$ small

fit will be forced closer to y_i

$$\sum_{i=1}^{n} h_{ii} = p'$$

$$\bar{h}_{ii} = \frac{p'}{n}$$

if leverage > $\frac{2p'}{n}$, observation should be looked at more closely

Outliers

an outlier in one model may not be an outlier in another where the variables have been changed or transformed
individual outliers are less of a problem in larger datasets
a single point won't have the leverage to affect the fit very much
only need to worry about clusters of outliers in larger datasets because they're less likely to occur by chance and more likely to represent actual structure

if i[th] case is suspected to be an outlier, exclude point i to form a reduced dataset
recompute estimates to get $\hat{\beta}_{(i)}$ and $\hat{\sigma}_{(i)}$

Standardized Residuals/Internal Studentized Residual:

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

Studentized Residual/External Studentized Residuals/Jackknife Residuals

$$t_i = \frac{\hat{e}_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}} = r_i \sqrt{\frac{n - p' - 1}{n - p' - {r_i}^2}}, t(n - p' - 1)\ df$$

need to fit a new model for each removed observation to calculate jackknife residuals
jackknife residuals can be approximated by standardized residuals
implicitly testing all cases so implement Bonferroni correction, a multiple testing correction
for a significance level of α, $\frac{\alpha}{n}$ correction should be used in each test

How to Handle Outliers
check for data entry error
examine physical context of data collection
exclude the point from analysis, but try to include it if the model is changed
always report the existence of outliers even if they're not included in final model

Influential Points

Difference in Fits

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{\hat{\sigma}_{(i)} h_{ii}}}$$

if $DFFITS_i > 2\sqrt{\frac{p'+1}{n-p'-1}}$, observation should be looked at more closely

Cook's Distance

see if slope changes significantly if you exclude the influential observation
reduce information to a single value for each case

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta}_i)'(X'X)(\hat{\beta}_{(i)} - \hat{\beta}_i)}{p'\hat{\sigma}^2} = \frac{(\hat{Y}_{(i)} - Y_i)'(\hat{Y}_{(i)} - \hat{Y}_i)}{p'\hat{\sigma}^2} = \frac{1}{p'} r_i{}^2 \frac{h_{ii}}{1 - h_{ii}} \sim F(p', n - p')\ df$$

if $D_i > \frac{4}{n}$, observation should be looked at more closely
if $D_i > 0.5$, observation may be influential
if $D_i \geq 1$, observation is most likely influential
if D_i sticks out from others, observation is almost certainly influential

use the F-distribution to calculate potential influential point's percentile value
if < 10-20%, observation has little influence
if > 50% observation is a highly influential point
if in between those thresholds, judgement is ambiguous