Diagnostics and Remedial Measures

3.1 diagnostics for predictor variables

3.2 Properties of Residuals

Definition. ***Definition and Properties of residuals***
residual e_i is the difference between the observed value y_i and fitted value $\hat{y}_i$

$$e_i = y_i - \hat{y}_i$$

*so residual is a random variable and can be regarded as the **observed error**, in comparison to unknown true error ϵ_i in the regression model*

$$\epsilon_i = y_i - \mathbb{E}(y_i)$$

*Note ϵ_i are assumed to be independent normal random variable with mean 0 and variance σ^2. If model is appropriate for the data at hand, then observed residual e_i should reflect properties assumed for ϵ_i. This underlies **residual analysis**, which is used to examine aptness of a statistical model*

1. ***mean*** *From taking partial of RSS (when evaluating LS estimator)*

$$\bar{e} = \frac{\sum e_i}{n} = 0$$

   Note this does not imply $\mathbb{E}(\epsilon_i) = 0$

2. ***Variance***

$$s^2 = \frac{\sum (e_i - \bar{e})^2}{n-2} = \frac{\sum e_i^2}{n-2} = \frac{RSS}{n-2} = MSE$$

   if model is ok, MSE is unbiased estimator of $\sigma^2(e_i)$

3. ***nonindependence*** *residual e_i are not independent random variables because they involve $\hat{y}_i$ which are based on the same fitted regression line. So residuals subject to*

$$\sum e_i = 0 \quad \textit{and} \quad \sum x_i e_i = 0$$

   If sample size is large, dependency effect of e_i may be ignored

Definition. ***semistudentized residuals*** *Want to standardize residual*

$$e_i^* = \frac{e_i - \bar{e}}{\sqrt{MSE}} = \frac{e_i}{\sqrt{MSE}}$$

*if $\sqrt{MSE}$ is an estimate of $se(e_i)$, then e_i^* is a **studentized residual**. But standard deviation of e_i is complex, and $\sqrt{MSE}$ is only an approximation, so e_i^* is **semistudentized residual***

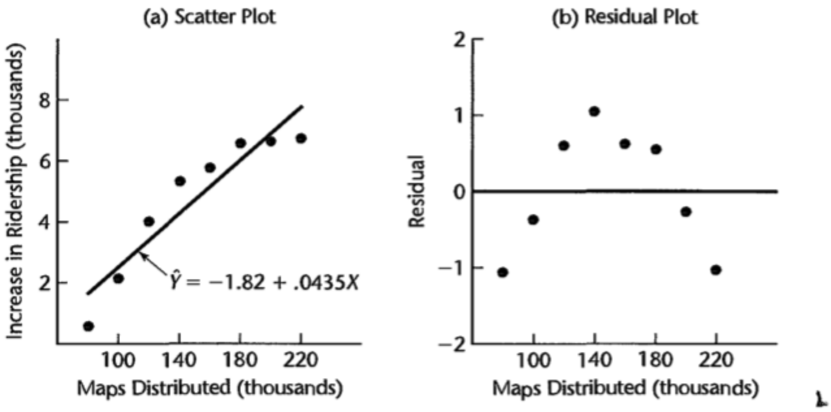
Definition. ***Possible departure from model with residuals***

1. *regression fucntion is not linear*
2. *error term does not have constant variance (not homoscedastic)*
3. *error term not independent*
4. *model fits all but one or a few outlier observations*
5. *error term not normally distributed*
6. *one or several important predictor variables have been omitted from the model*

3.3 Diagnostics for residuals

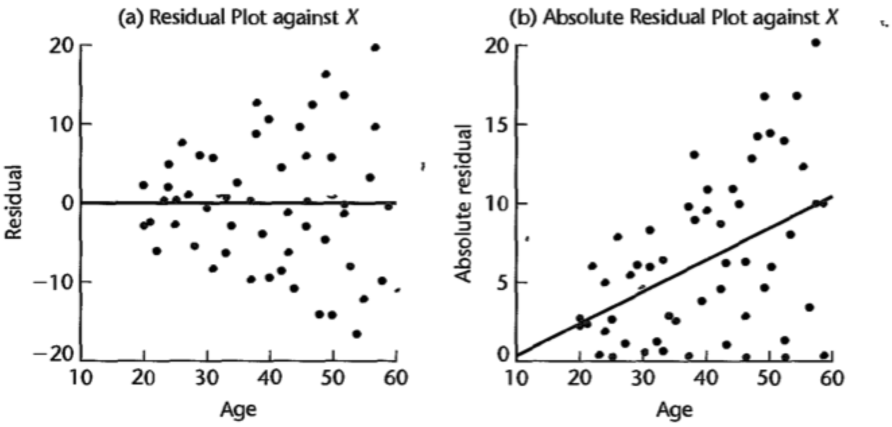
Definition. ***Nonlinearity of regression function***
Deciding if a linear regression function is appropriate can be evaluated from a a ***residual plot against*** X ***or*** $\hat{Y}$ *(preferred) or from a scatter plot. We look for if data points depart from zero in a systematic fashion*

FIGURE 3.3 **Scatter Plot and Residual Plot Illustrating Nonlinear Regression Function—Transit Example.**



Definition. ***Nonconstancy of error variance***

FIGURE 3.5 **Residual Plots Illustrating Nonconstant Error Variance.**

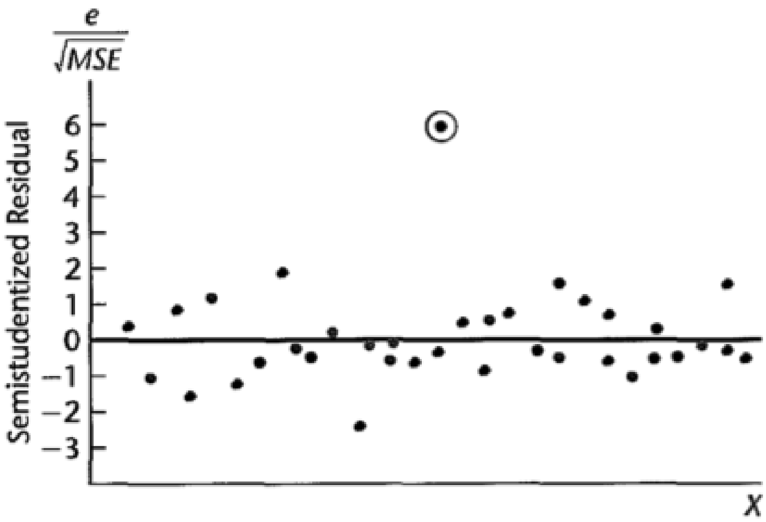


Definition. ***Presence of outliers***
*Can be examined from a **residual plot against** X **or** $\hat{Y}$ as well as*

plot of semistudentized residual

The latter is useful in that its easy to identify residuals that lie many standard deviations from zero (≥ 4 standard deviation may be considered outlier)

FIGURE 3.6 **Residual Plot with Outlier.**

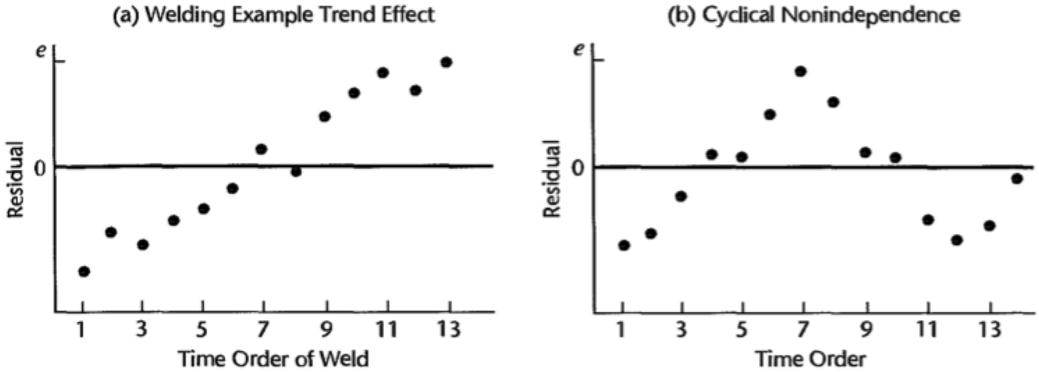


Outliers create difficulty for modeling. Since for least squared estimators, the fitted line may be pulled disproportionately toward an outlying observation, causing misleading fit if outlier is from a mistake or other cause

Definition. ***Nonindependence of error terms***
Whenever data obtained in a time sequence or some other type of sequence (geographical

area), its good idea to use a ***sequence plot of residuals*** *to see if there is correlation between error terms near each other in sequence. If error terms are independent, we would expect residuals in a sequence plot to fluctuate in a random pattern around base line 0*

FIGURE 3.8 Residual Time Sequence Plots Illustrating Nonindependence of Error Terms.



Definition. ***Nonnormality of error terms***
Test with distribution plot, like **box plot** *or* ***histogram****. Another possibility is to prepare a*

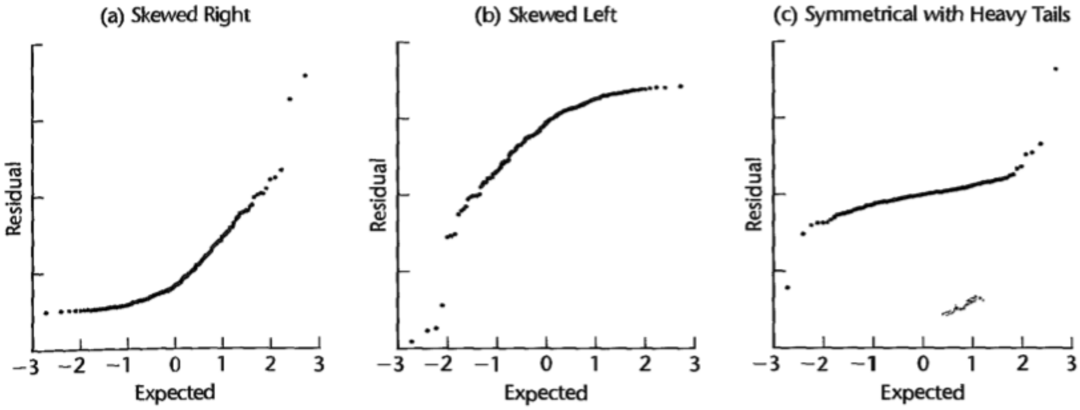
Normal probability plot

where each residual is plotted against its expected value under normality. A plot that is linear suggest agreement with normality.

TABLE 3.2 Residuals and Expected Values under Normality—Toluca Company Example.

Run i	(1) Residual e_i	(2) Rank k	(3) Expected Value under Normality
1	51.02	22	51.95
2	−48.47	5	−44.10
3	−19.88	10	−14.76
...	...	...	...
23	38.83	19	31.05
24	−5.98	13	0
25	10.72	17	19.93

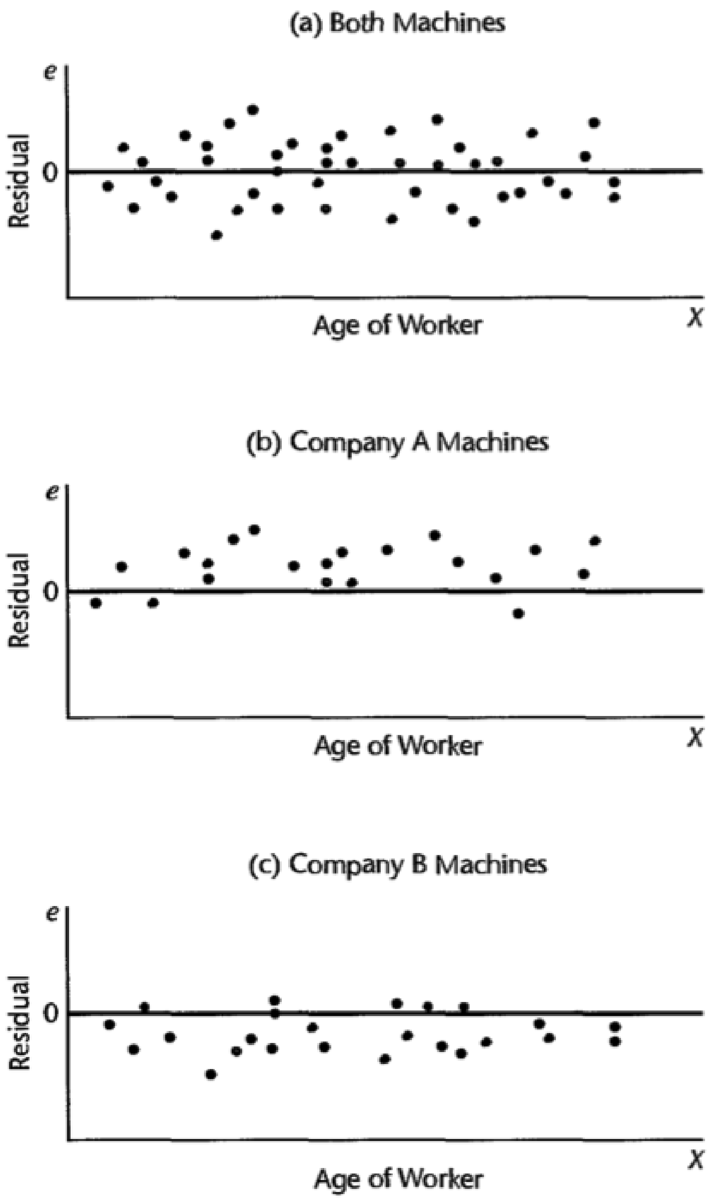
FIGURE 3.9 Normal Probability Plots when Error Term Distribution Is Not Normal.



Heavy tail means that distribution has higher probabilities in the tails than a normal distribution. Note left skew distribution has a long tail to the left or graph (mean less than median)

Definition. ***Omission of important predictor variables***
Residuals should also be plotted agianst variables omitted from model that might have important effect on the model. Goal is to identify if there is other key variables in providing important additional predictive power to the model

FIGURE 3.10 **Residual Plots for Possible Omission of Important Predictor Variable—Productivity Example.**



3.4 Overview of Tests Involving residuals

3.4 Overview of Remedial Measures

10.4 Identifying Influential cases - DFFITS, Cooks distance, DFBETAS

Definition. ***Intro*** *After identifying cases that are outlying w.r.t. Y and/or X values, we have to ascertain if they are influential. A data point is considered* ***Influential*** *if its exclusion causes major changes in the fitted regression function*

Definition. ***Influence on single fitted value DFFITS***
$DFFITS_i$ represents the difference between ***fitted value*** *$\hat{y}_i$ for ith case when all n cases are used in fitting regression function and the the* ***predicted values*** *$\hat{y}_{i(i)}$ for which the ith case obtained when the ith case is omitted in fiotting the regression function.*

$$(DFFITS)_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{se(\hat{y}_{i(i)})} \quad where \quad se(\hat{y}_{i(i)}) = \sqrt{MSE_{(i)}h_{ii}}$$

Note it uses $MSE_{(i)}$, the error mean square when i the case is omitted in fitting the regression function for estiamting error variance σ^2. By standardization, $DFFITS_i$ represents number of estimated standard deviation of $\hat{y}_i$ that the fitted value $\hat{y}_i$ increases/decreases with the inclusion of ith case in fitting the regression model

Definition. ***Influence on all fitted values - Cook's distance***
Cook's distance considers influence of ith case on all n fitted values, i.e. an aggregate influence measure,

$$D_i = \frac{\sum_j (\hat{y}_j - \hat{y}_{j(i)})^2}{pMSE}$$

The larger the $\hat{e}_i$ and h_{ii}, the larger D_i. So ith case can be influential if

1. *have a large residual $\hat{e}_i$ and only a moderate leverage value h_{ii} or*
2. *have a larger leverage value h_{ii} but a moderately sized residual $\hat{e}_i$*
3. *have both a large residual and a large leverage*

Definition. ***Influence on the regression coefficient DFBETAS***
$DFBETAS_i$ is a measure of the influence of the ith case on each regression coefficient $\hat{\beta}_k$ ($k = 0, 1, \cdots$), specifically it is the difference between estimated regression coefficient $\hat{\beta}_k$ based on all n cases and the regression coefficient obtained when ith case is omitted, denoted as $\hat{\beta}_{k(i)}$

$$(DFBETAS)_{k(i)} = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{se(\hat{\beta}_{k(i)})} \quad k = 0, 1, \cdots, p-1 \quad where \quad se(\hat{\beta}_{k(i)}) = \sqrt{MSE_{(i)}c_{kk}}$$

where c_{kk} is kth diagonal element of $(X'X)^{-1}$, so variance of $\hat{\beta}_k$ is given by

$$\sigma^2(\hat{\beta}_k) = \sigma^2 c_{kk}$$

Note

1. *sign of DFBETAS indicates if inclusion of a case leads to an increase or a decrease in the estimated regression coefficient*

2. *absolute magnitude of DFBETAS indicate size of difference relative to estiamted standard deviation of regression coefficient. So large $DFBETAS_{k(i)}$ is indicative of large impact of ith case on kth regression coefficient*

3. *A case is **influential** if $|DFBETAS|$ exceeds 1 for small to medium datasets and $2/\sqrt{n}$ for large datasets*

modern regression with r ch3

3.1.2 Use residual plot to determine if proposed regression is a valid model

3.2 Regression diagnostics tools for checking validity of a model

Definition. ***Steps***

1. *determine if proposed regression model is valid*

   $$\text{See if there is pattern in standardized residual plot}$$

2. *Identify **leverage points***

   $$\text{See if leverage } h_{ii} \text{ satisfies } h_{ii} > \frac{4}{n}$$

3. *Identify **outliers***

   $$\text{See if studentized residual } r_i \text{ satisfies } |r_i| > 2$$

4. *Identify **bad leverage points or influential points***

   $$\text{A influential point is a leverage point that is also an outlier}$$

5. *Assess **error homoscedasticity***

6. *For time series, examine if data correlated over time*

7. *Assess assumption of **normally distributed error***

3.2.1 leverage points

Definition. ***characerizing leverage points***

1. ***Leverage point*** *is a point whose x-value is distance from other x-values.*
2. ***good leverage point*** *is a leverage point which is not an outlier*
3. ***bad leverage point or influential point*** *is a leverage point which is also an outlier. It has a large effect on fitted regression line, whose inclusion/exclusion from the model changes the fitted model ($\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{y}_i$) dramatically*
4. ***outlier that is not a leverage point***

Definition. ***Express $\hat{y}_i$ in terms of y_i***
Conceptually this relation implies the extend to which the fitted regression line is attracted by a given point. We can express $\hat{y}_i$ as a linear combination of y_i

$$\hat{y}_i = \sum_j h_{ij} y_j \quad \textit{where} \quad h_{ij} = \left[\frac{1}{n} + \frac{(x_i - \overline{x})(x_j - \overline{x})}{S_{XX}}\right]$$

or equivalently

$$\hat{y}_i = h_{ii} y_i + \sum_{j, j \neq i} h_{ij} y_j \quad \textit{where} \quad h_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_j (x_j - \overline{x})^2}$$

with property

$$\sum_j h_{ij} = 1 \qquad h_{ij} = h_{ji} \qquad \sum_j h_{ij}^2 = h_{ii}$$

*h_{ii} is the **leverage** of (x_i, y_i). Note*

1. *$(x_i - \overline{x})$ measures distance x_i is away from $\overline{x}$.*
2. *h_{ii} shows how y_i affects $\hat{y}_i$. The idea is that if $h_{ii} \sim 1$, then the other h_{ij} terms will be zero (by $\sum_j h_{ij} = 1$), so $\hat{y}_i \sim y_i$ (the actual value) regardless of what values of rest of data take*
3. *h_{ii} is purely depedent on x_i, so*

a point of high leverage can be found by looking at x-values only

For simple linear regression

$$average(h_{ii}) = \frac{2}{n} \quad i = 1, 2, \cdots, n$$

*A **leverage point** is a point with high leverage. Practically, instead of constraining $x_i \to 1$ we classify x_i as a point of high leverage if*

$$h_{ii} > 2 \times average(h_{ii}) = \frac{4}{n}$$

Proof.
Proving $\hat{y}_i$ is a linear combination of y_i

$$\begin{aligned}
\hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i = \overline{y} - \hat{\beta}_1 \overline{x} + \hat{\beta}_1 x_i \\
&= \overline{y} - \hat{\beta}_1 (x_i - \overline{x}) \\
&= \frac{1}{n} \sum_j y_j + \sum_j \left(\frac{(x_j - \overline{x})}{S_{XX}} \right) y_j (x_i - \overline{x}) \\
&= \sum_j \left(\frac{1}{n} + \frac{(x_i - \overline{x})(x_j - \overline{x})}{S_{XX}} \right) y_j
\end{aligned}$$

Proving some properties

$$\sum_j h_{ij} = \sum_j \left[\frac{1}{n} + \frac{(x_i - \overline{x})(x_j - \overline{x})}{S_{XX}} \right] = \frac{n}{n} + \frac{(x_i - \overline{x})}{S_{XX}} \sum_j (x_j - \overline{x}) = 1$$

$$\begin{aligned}
\sum_j h_{ij}^2 &= \sum_j \left(\frac{1}{n} + \frac{(x_i - \overline{x})(x_j - \overline{x})}{S_{XX}} \right)^2 \\
&= \frac{1}{n} + \left(\frac{x_1 - \overline{x}}{S_{XX}} \right)^2 \sum_j (x_j - \overline{x})^2 + \frac{2}{n} \frac{x_i - \overline{x}}{S_{XX}} \sum_j (x_j - \overline{x}) \\
&= \frac{1}{n} + \left(\frac{x_1 - \overline{x}}{S_{XX}} \right)^2 S_{XX} + 0 \\
&= \frac{1}{n} + \frac{(x_1 - \overline{x})^2}{S_{XX}} \\
&= h_{ii}
\end{aligned}$$

□

3.2.2 Standardized residual

Definition. *standardized residuals*
Residuals $\hat{e}_i = y_i - \hat{y}_i$ do not have same variance. This is apparent given,

$$Var(\hat{e}_i) = \sigma^2 (1 - h_{ii}) \qquad \textit{where} \qquad h_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})(x_j - \overline{x})}{S_{XX}}$$

$$Var(\hat{y}_i) = \sigma^2 h_{ii}$$

Proof. To find $Var(\hat{e}_i)$, idea is to use the formula representing $\hat{y}_i$ as a linear combination of y_i and then take the variance

$$
\begin{aligned}
Var(\hat{e}_i) &= Var\left(y_i - \hat{y}_i\right) \\
&= Var\left(y_i - h_{ii}y_i - \sum_{j\neq i} h_{ij}y_j\right) && (\hat{y}_i = h_{ii}y_i + \sum_{j\neq i} h_{ij}y_j) \\
&= Var\left((1-h_{ii})y_i - \sum_{j\neq i} h_{ij}y_j\right) \\
&= (1-h_{ii})^2\sigma^2 + \sum_{j\neq i} h_{ij}^2\sigma^2 && \text{(covariance=0 by ind. of } y) \\
&= \sigma^2\left(1 - 2h_{ii} + h_{ii}^2 + \sum_{j\neq i} h_{ij}^2\right) && (h_{ii}^2 + \sum_{j\neq i} h_{ij}^2 = \sum_j h_{ij}^2) \\
&= \sigma^2(1 - 2h_{ii} + h_{ii}) && (\sum_j h_{ij}^2 = h_{ii}) \\
&= \sigma^2(1-h_{ii})
\end{aligned}
$$

So that

$$Var(\hat{y}_i) = Var(\sum_{j\neq i} h_{ij}y_j) = \sum_{j\neq i} h_{ij}^2 Var(y_j) = \sigma^2 h_{ii}$$

□

Comments

1. *if $h_{ii} \approx 1$, then ith point is a leverage point, the corresponding residual $\hat{e}_i$ has a small variance.*
2. *Note expanding h_{ii} gives the familiar formula for variance of $\hat{y}_i$.*
3. *if $h_{ii} \approx 1$, then $\hat{y}_i \approx y_i$, with $Var(\hat{y}_i) \approx \sigma^2 = Var(y_i)$.*

We can overcome different variances of $\hat{e}_i$ by standardizing each residual by dividing it by estimate of its standard deviation. The ith ***standardized/studentized residual****, r_i is given by*

$$r_i = \frac{\hat{e}_i}{s\sqrt{1-h_{ii}}} \qquad \textit{where} \qquad s = \sqrt{MSE} = \sqrt{\frac{\sum_j \hat{e}_j^2}{n-2}}$$

Residual plot vs studentized residual plot

1. *When point of high leverage does not exist, $h_{ii} \not\approx 1$, and so $Var(\hat{e}_i) = \sigma^2$ for all $i = 1, \cdots, n$. residual plot and studentized residual plot are similar if not identical*

2. *However, when point of high leverage does exist, $h_{ii} \approx 1$, studentized residual plot is more informative because residual plot will have nonconstant variance (because $Var(\hat{e}_i) = \sigma^2(1 - h_{ii})$ varies depending on if ith data has high leverage or not) even if the errors have constant variance.*

3. *standardized residuals immediately tell us how many estimated standard deviation any point is away from the fitted regression model*

Use standardized residual to identify outliers
*A rule of thumb for identifying **outliers** is when the point is > 2 standard deviations from the fitted regression model (i.e. $|r_i| > 2$ on a studentized residual plot)*

3.2.4 Assess influence of certain cases

Definition. *Cooks distance describes how far, on average, $\hat{y}$ would move if observation in question is dropped*

$$D_i = \frac{\sum_j (\hat{y}_{j(i)} - \hat{y}_j)^2}{2S^2} = \frac{r_i^2 h_{ii}}{2(1 - h_{ii})}$$

where subscript (i) means ith case has been deleted from fit. So the fit is based on the other $n - 1$ cases, i.e. $1, 2, \cdots, i - 1, i + 1, \cdots, n$. So $\hat{y}_{j(i)}$ denote jth fitted value based on the fit obtained when ith case has been deleted from the fit. r_i is the ith standard residual and h_{ii} is the ith leverage value.

1. $\frac{r_i^2}{2}$ *measures extent to which ith case is outlying*

2. $\frac{h_{ii}}{1 - h_{ii}}$ *measures leverage of ith case*

3. *So either/both large r_i or/and h_{ii} yields alrge value of D_i*

4. *D_i identifies **influential point** if $D_i > \frac{4}{n-2}$ or if D_i is separated by a large gap from the other D_is*

3.2.5 Normality of Error

Definition. *Assumption of normal error needed in **small samples** for validity of t-distribution based hypothesis tests and confidence interval and for **all sample sizes** for prediction interval. There are two types of deviation from normality*

1. ***asymmetry***

2. ***heavy tails***

*Usually checked by looking at distribution of residuals or standardized residual with **box plot** or **scatter plot**. A common way to assess normality of error is to look at **normal probability plot (normal Q-Q plot)** of standardized residuals.*

Definition. ***Q-Q plot***
Q-Q plot tests if a sequence of numbers follow a certain distribution. In case of residual analysis, it can be obtained from plotting ***ordered standardized residuals*** *on vertical axis against* ***expected order statistics*** *from a standard normal distribution on the horizontal axes. Plots with points close to a straight line support error normality.*

Definition. *What does* ***residual*** *informs us*

1. ***linear of model****, check for systematic pattern*
2. ***missing predictor variable*** *check for systematic pattern when plotted against potential predictor variable*
3. ***outliers*** *with Cooks distance plot (standardized residual* r_i *vs leverage* h_{ii}*) check for outliers*
4. ***error homoscedasticity*** *with scale-location plot, check if residuals are spread evenly over range of predicted variable*
5. ***error normality*** *with Q-Q plot (normal probability plot), check for deviation of points from straight line*
6. ***error independence*** *with a residual sequence plot (plot against time/space), check for temporal or spatial dependence*