# Building the Regression Model
# Diagnostics and Remedial Measures

Paweł Polak

November 20, 2017

Linear Regression Models - Lecture 14

# Content:

- Diagnostics
    - Marginal effect of adding a new explanatory variable
    - Outliers and Influential Observations
    - Deleted Residuals - Identifying Outlying $Y$ Observations
    - Leverage - Identifying Outlying $X$ Observations
    - DFFITS measure, Cook's distance, and DFBETAS measure
    - Multicollinearity & Variance Inflating Factors (see also Lecture 12)
- Remedial Measures
    - Heteroskedasticity - Weighted Least Squares
    - Dealing with outliers - Robust Regression

# General Linear Model

- *Independent responses* of the form $Y_i \sim N(\mu_i, \sigma^2)$, where

$$\mu_i = \mathbf{X}_i^\top \boldsymbol{\beta}$$

for some known vector of *explanatory* variables $\mathbf{X}_i^\top = (X_{i1}, \ldots, X_{ip})$.

- Unknown *parameter* vector $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{P-1})^\top$, where $P < N$.

- This is the *linear model* and is usually written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

(in vector notation) where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_1^\top \\ \vdots \\ x_N^\top \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{P-1} \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{pmatrix},$$

where $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$, for $i = 1, 2, \ldots, N$.

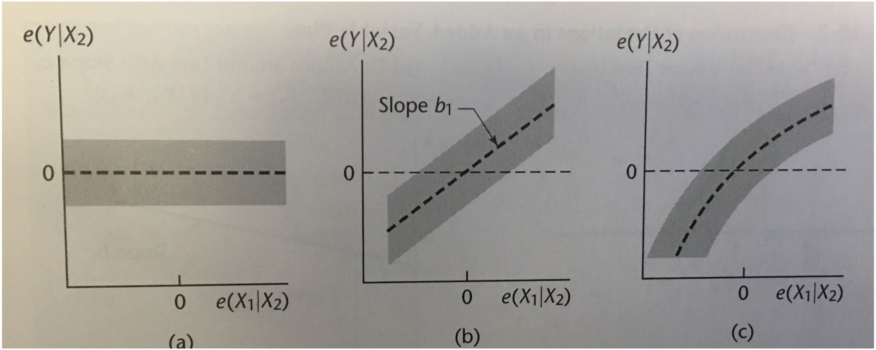
# Advanced Diagnostics

- There are a number of advanced diagnostic tools for checking the adequacy of a regression model.
- These include methods for investigating the *appropriate* functional form for an explanatory variable, *outliers*, *influential* observations and *multi-collinearity*.

# Marginal Effect of a Variable: Added-variable plots

- Added-variable plots are refined residual plots that show the *marginal effect* of an explanatory variable, *given* the *other variables* in the model.
- These plots can be used to show the marginal importance of $X_k$ in reducing the residual variability and provide suggestions about the nature of the functional relation for $X_k$ in the regression model.
- Procedure for making an added-variable plot of $Y$ against $X_k$:
  - Regress *both* the response variable $Y$ and the explanatory variable $X_k$ against all the other explanatory variables in the regression model.
  - Obtain *residuals* for both.
  - Plot the residuals against each other.

# Marginal Effect of a Variable: Added-variable plots




(a) $X_1$ contains no additional information useful for predicting $Y$ beyond that contained in $X_2$.

(b) A linear term in $X_1$ might be a helpful to addition to the regression model already containing $X_2$. (Recall from Lecture 12 that the slope is equal to the estimate of $b_1$ in a model

$$\mathbf{Y} = b_0 + b_1\mathbf{X}_1 + b_2\mathbf{X}_2 + \mathbf{e}$$

(c) Indicates that the addition of $X_1$ to the regression model may be helpful and suggesting the possible nature of the curvature effect by the pattern shown.

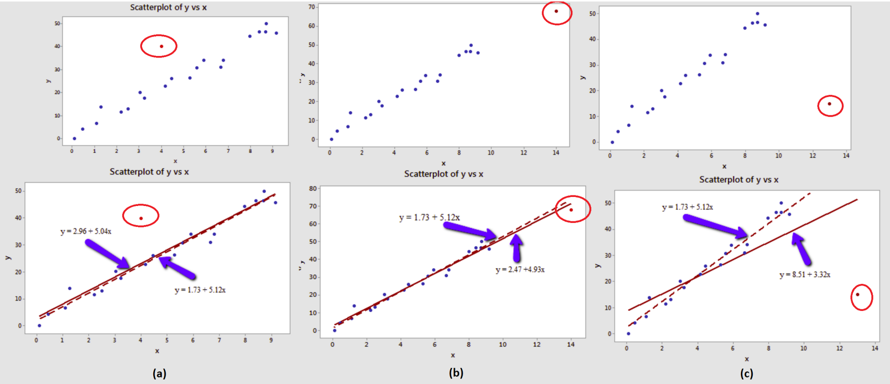
## Outliers

- *Outliers* are observations that are separated from the remainder of the data in some way.
  - They can be *extreme* in the $x$ or $y$-direction or both.
  - Certain types of outliers have dramatic effects on the fitted regression function, while others do not.

## Influential observations

- An *influential point* is an observation that, if removed, would considerably *change* the position of the regression line.
- A key step in any regression analysis is to determine if the model is heavily influenced by one or few of the observations.
- We have used residual plots to detect extreme observations.
- However, residual plots are often not useful for identifying influential points since such points tend to have *small* residuals.

# Outliers vs. Influential Points



# Residuals

- Recall that the vector or *residuals* can be expressed as:

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y},$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ (hat matrix).

- Properties: $\mathbb{E}(\mathbf{e}) = 0,\ Var(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$
- The *variance* of the $i$-th residual is $\sigma^2(1 - h_{ii})$
- The *covariance* between $e_i$ and $e_j$ is $Cov(e_i, e_j) = -\sigma^2 h_{ij}$, for $i \neq j$.
- *Studentized* residuals:

$$r_i = \frac{e_i}{\sqrt{MSE(1 - h_{ii})}}$$

# Deleted Residuals - Identifying Outlying $Y$ Observations

- Often it is more efficient to compute the $i$-th residual using a fitted regression equation based on all the data *except* the $i$-th observation.
- In the event that the $i$-th observation is an influential point the fitted value $\hat{Y}_{i(i)}$ will not be influenced by this observation and will tend to give a *larger* residual making it easier to detect.
- The *deleted residual* for the $i$-th case is given by

$$d_i = Y_i - \hat{Y}_{i(i)},$$

  where $\hat{Y}_{i(i)}$ denotes the fitted value, computed without the $i$-th observation, at $X$ levels corresponding to the $i$th observation.
- Recall the prediction sum of squares $PRESS = \sum \left( Y_i - \widehat{Y}_{i(i)} \right)^2$ criterion.
- One can show that $d_i = e_i/(1 - h_{ii})$.

# Deleted Residuals: Proof of $d_i = e_i/(1-h_{ii})$

**Proof.**

$$\text{Note that } d_i = \mathbf{Y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}_{(i)}, \quad \text{where} \quad \widehat{\boldsymbol{\beta}}_{(i)} = \left(\mathbf{X}_{(i)}^T\mathbf{X}_{(i)}\right)^{-1}\mathbf{X}_{(i)}^T\mathbf{Y}_{(i)}.$$

$$\text{Moreover, } \mathbf{X}_{(i)}^T\mathbf{X}_{(i)} = \mathbf{X}^T\mathbf{X} - \mathbf{X}_i\mathbf{X}_i^T, \quad \mathbf{X}_{(i)}^T\mathbf{Y}_{(i)} = \mathbf{X}^T\mathbf{Y} - \mathbf{X}_iY_i, \quad \mathbf{X}_i^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_i = h_{ii},$$

and, by *Sherman-Morrison formula*, $\left(\mathbf{X}_{(i)}^T\mathbf{X}_{(i)}\right)^{-1} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1} + \dfrac{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_i\mathbf{X}_i^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}}{1-h_{ii}}$.

$$\text{Therefore, } \widehat{\boldsymbol{\beta}}_{(i)} = \left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1} + \frac{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_i\mathbf{X}_i^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}}{1-h_{ii}}\right]\left(\mathbf{X}^T\mathbf{Y} - \mathbf{X}_iY_i\right)$$

$$= \widehat{\boldsymbol{\beta}} - \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_iY_i + \frac{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_i\mathbf{X}_i^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}}{1-h_{ii}}\mathbf{X}^T\mathbf{Y} - \frac{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_i\mathbf{X}_i^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}}{1-h_{ii}}\mathbf{X}_iY_i$$

$$= \widehat{\boldsymbol{\beta}} - \left[\frac{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_i}{1-h_{ii}}\right]\left[Y_i(1-h_{ii}) - \mathbf{X}_i^T\widehat{\boldsymbol{\beta}} + h_iY_i\right] = \widehat{\boldsymbol{\beta}} - \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\frac{\mathbf{X}_ie_i}{1-h_{ii}}.$$

$$\text{Hence, } d_i = Y_i - \mathbf{X}_i^T\widehat{\boldsymbol{\beta}}_{(i)} = Y_i - \mathbf{X}_i^T\left(\widehat{\boldsymbol{\beta}} - \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\frac{\mathbf{X}_ie_i}{1-h_{ii}}\right) = e_i + h_i\frac{e_i}{1-h_i} = \frac{e_i}{1-h_{ii}}.$$

□

## Deleted residuals

- The estimated variance of $d_i$ is given by

$$s^2(d_i) = MSE_{(i)} \left(1 + \mathbf{X}_i^T \left(\mathbf{X}_{(i)}^T \mathbf{X}_{(i)}\right)^{-1} \mathbf{X}_i\right) = \frac{MSE_{(i)}}{1 - h_{ii}}$$

- Here $MSE_{(i)}$ is the mean square error when the $i$-th case is omitted from the regression model.

## Studentized deleted residuals

- The *studentized deleted residual* is given by

$$t_i = \frac{d_i}{se(d_i)} = \frac{e_i/(1 - h_{ii})}{\sqrt{MSE_{(i)}/(1 - h_{ii})}} = \frac{e_i}{\sqrt{MSE_{(i)}(1 - h_{ii})}}$$

- It can be shown that

$$t_i \sim t_{N-P-1}$$

- Note that there are $(N - 1) - P$ d.f. associated with $MSE_{(i)}$ since we only use $N - 1$ observations when estimating its value.

# Relationship

- Ideally, we would like to avoid computing $MSE_{(i)}$ for each observation.

- Fortunately, the following relationship holds (proof in the next slide):

$$(N-P)MSE = (N-P-1)MSE_{(i)} + \frac{e_i^2}{1-h_{ii}}$$

- Hence, we can write $t_i$ as

$$t_i = e_i \left[ \frac{N-P-1}{SSE(1-h_{ii}) - e_i^2} \right]^{1/2}$$

- Thus the deleted residuals can be computed without refitting the data.

# Deleted Residuals: Proof of the *MSE*s Relationship

Proof.

$$
\begin{aligned}
SSE_{(i)} &= \sum_{j \neq i} \left(Y_j - \mathbf{x}_j^T \widehat{\boldsymbol{\beta}}_{(i)}\right)^2 = \sum_{j \neq i} \left(Y_j - \mathbf{x}_j^T \left(\widehat{\boldsymbol{\beta}} - \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \frac{\mathbf{x}_i e_i}{1-h_{ii}}\right)\right)^2 \\
&= \sum_{j=1}^N \left(Y_j - \mathbf{x}_j^T \left(\widehat{\boldsymbol{\beta}} - \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \frac{\mathbf{x}_i e_i}{1-h_{ii}}\right)\right)^2 - \left(Y_i - Y_{i(i)}\right)^2 \\
&= \sum_{j=1}^N \left(\left(Y_j - \mathbf{x}_j^T \widehat{\boldsymbol{\beta}}\right) + \mathbf{x}_j^T \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \frac{\mathbf{x}_i e_i}{1-h_{ii}}\right)^2 - \left(Y_i - Y_{i(i)}\right)^2 \\
&= \sum_{j=1}^N \left(Y_j - \mathbf{x}_j^T \widehat{\boldsymbol{\beta}}\right)^2 + \frac{2}{1-h_{ii}} \sum_{j=1}^N \left(Y_j - \mathbf{x}_j^T \widehat{\boldsymbol{\beta}}\right) \mathbf{x}_j^T \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{x}_i e_i \\
&+ \frac{1}{(1-h_{ii})^2} \mathbf{x}_i^T \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \sum_{j=1}^N \mathbf{x}_j \mathbf{x}_j^T \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{x}_i e_i^2 - \frac{e_i^2}{(1-h_{ii})^2} \\
&= SSE + 0 + \frac{1}{(1-h_{ii})^2} \mathbf{x}_i^T \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{x}_i e_i^2 - \frac{e_i^2}{(1-h_{ii})^2} = SSE - \frac{e_i^2}{1-h_{ii}}
\end{aligned}
$$

Hence, $(N-P)MSE = (N-P-1)MSE_{(i)} + \frac{e_i^2}{1-h_{ii}}$. □

# Leverage - Identifying Outlying $X$ Observations

- Recall that the fitted values can be written as $\hat{\boldsymbol{Y}} = \boldsymbol{HY}$, i.e.,

$$\hat{Y}_i = h_{i1}Y_1 + h_{i2}Y_2 + \ldots + h_{ii}Y_i + \ldots + h_{iN}Y_N \text{ for } i = 1, \ldots, N.$$

- $h_{ii}$ is called the *leverage* for the $i$-th observation.
- The leverage is always between 0 and 1, and $\sum_{i=1}^{N} h_{ii} = P$
- Since the leverage is a function only of $X$ it measures the role of the $X$ *values* in determining how $Y_i$ affects the fitted value.
- Outliers in the $X$-direction tend to have higher leverage values and thus a larger effect on the fitted regression function.

# Leverage - Identifying Outlying $X$ Observations

- Recall that $Var(e_i) = \sigma^2(1 - h_{ii})$
- Hence, the larger $h_{ii}$ the smaller the variance of the residuals.
- When $h_{ii}$ equals 1 the variance of $e_i$ is 0 and the fitted value is equal to the observed value $Y_i$.
- On the other hand, a point with zero leverage has *no effect* on the regression model.

# Identifying influential cases

- We can identify outliers in the $Y$-direction using *studentized deleted residuals* and outliers in the $X$-directions using *leverage* values.
- However, not all outliers will have a large effect on the fitted regression function and therefore do not require remedial measures.
- After identifying outliers the next step is to determine whether or not they are influential.

# DFFITS - Influence on Single Fitted Value

- A measure of the *influence* that the $i$-th observation has on the fitted value $\hat{Y}_i$ is given by

$$(DFFITS)_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}}$$

- It represents the number of *std. dev.* of $\hat{Y}_i$ that the fitted value *increases* or *decreases* with the inclusion of the $i$-th observation.

- Note: $Var(\hat{Y}_i) = \sigma^2 h_{ii}$

- It can be re-expressed as

$$(DFFITS)_i = t_i \left( \frac{h_{ii}}{1 - h_{ii}} \right)^{1/2}$$

- Hence, *DFFITS* is a *studentized deleted residual* scaled by a factor that is a function of the *leverage* of the observation.

- Absolute values above 1 considered influential.

# Cook's distance - Influence on All Fitted Values

- *Cook's distance* measures the *aggregate* influence of the $i$-th value on all $N$ fitted values.
- It is defined as $D_i = \frac{\sum_{j=1}^{N}(\hat{Y}_j - \hat{Y}_{j(i)})^2}{P \;\; MSE}$
- In contrast to DFFITS each of the $N$ fitted values is compared with the fitted value when the $i$-th observation is omitted.
- Cook's distance can be re-expressed as $D_i = \frac{e_i^2}{P \;\; MSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$
- The value of $D_i$ depends on two functions, the *size* of the residual $e_i$ and the *leverage* value $h_{ii}$.
- Hence, an observation can be influential by having a large residual and/or a large leverage.
- Typically, points with $D_i$ greater than 1 are classified as influential.

# DFBETAS - Influence on the Regression Coefficients

- A measure of the influence of the $i$-th observation on *each regression coefficient* $b_k$ in the model is given by

$$(DFBETAS)_{k(i)} = \frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)} c_{kk}}}.$$

- Here $c_{kk}$ is the $k$-th diagonal element of $(\mathbf{X}^\top \mathbf{X})^{-1}$.

- Recall: $Var(b) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$.

- DFBETAS measures the difference between the estimated regression coefficients $b_k$ based on all $N$ observations and the equivalent coefficient when the $i$-th observation is removed.

- A large absolute value ($> 1$ for small to medium data sets and $> 2/\sqrt{N}$ for large data sets) is indicative of a large impact of the $i$-th observation on the $k$-th regression coefficient.

- The *sign* indicates whether the inclusion of the observation leads to an increase or decrease in the estimated regression coefficient.

# Multicollinearity

- In multiple regression, the hope is that the explanatory variables are highly correlated with the response variable.
- However, it is not desirable for the explanatory variables to be correlated with one another.
- Multicollinearity exists when *two or more* of the *explanatory variables* used in the regression model are *highly correlated* and provide redundant information about the response.

# Variance Inflating Factors

- Recall from Lecture 12, that the *variance inflation factor* (*VIF*) can be used to *detect* the presence of *multicollinearity*

- These factors measure how much the variances of the estimated regression coefficients are inflated as compared to when the explanatory variables are not linearly related.

- The *VIF* for $b_k$ is given by

$$VIF_k = (1 - R_k^2)^{-1}$$

where $R_k^2$ is the multiple correlation coefficient when $X_k$ is regressed on the $(P - 1)$ other explanatory variables.

- A *large VIF* ($> 10$) is taken as an indication that multicollinearity may be influencing the estimates.

# Remedial Measures

- After fitting a regression model it is necessary to check the model assumptions by analyzing the *residuals* and studying *diagnostic plots*
- When the diagnostics indicate that the model assumptions are violated, *remedial measures* may be needed
- Some possible problems and their solutions:
  - Non-constant variance: *Weighted least squares*
  - Multicollinearity: *Ridge regression*
  - Outliers: *Robust* regression

# Heteroskedasticity

- In our regression model we have assumed that $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$

- That is, that the errors are *independent* and have the same variance (*homoskedasticity*)

- We now want to extend our models to allow for non-constant variance (*heteroskedasticity*).

- The *generalized linear regression* model is given by

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\epsilon}, \qquad \text{where}$$

$$Var(\boldsymbol{\epsilon}) = \Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_N^2 \end{pmatrix}$$

# Estimation – known variance

- Let us first consider the case where the error variances are *known*
- Could fit this model using ordinary least squares.
  - Estimators still *unbiased* and consistent, but no longer *minimum variance*
- To obtain estimators with minimum variance we must take into consideration that different points no longer have the same reliability.
- We can use the method of *maximum likelihood* to obtain estimators of the regression coefficient.
- The likelihood function is given by:

$$L(\beta) = \prod_{i=1}^{N} \frac{1}{(2\pi\sigma_i^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma_i^2}(Y_i - \beta_1 X_{i1} - \ldots - \beta_P X_{i,P})^2\right\}$$

- Define the *weight*: $w_i = \frac{1}{\sigma_i^2}$

- We can express the likelihood function as

$$L(\beta) = \left[ \prod_{i=1}^{N} \left( \frac{w_i}{2\pi} \right)^{1/2} \right] \exp\left\{ -\frac{1}{2} \sum_{i=1}^{N} w_i (Y_i - \beta_1 X_{i1} - \ldots - \beta_P X_{i,P})^2 \right\}$$

- We find the maximum likelihood estimators by minimizing the sum in the exponential term.

## Weighted least squares

- Estimates of $\beta_1, \ldots, \beta_P$ are obtained by minimizing the *weighted least squares* criterion:

$$Q = \sum_{i=1}^{N} w_i \left[ Y_i - (\beta_1 X_{i1} + \ldots + \beta_P X_{i,P}) \right]^2$$

where $w_i$ are weights *inversely proportional* to the *variances* (i.e., $w_i = 1/\sigma_i^2$ )

# Comments

- Ordinary least squares minimizes the sum of the squared residuals
- *WLS* minimizes the sum of the squared residuals multiplied by the *inverse of their variances*
- This allows us to give observations with low variability a higher weight than observations with high variability.
- In matrix notation we write the weighted least squares criterion as:

$$Q = (\mathbf{Y} - \mathbf{X}\beta)^\top \mathbf{W}(\mathbf{Y} - \mathbf{X}\beta)$$

  where $W$ is a *diagonal matrix* with elements $w_i$

- Taking the derivative with respect to $\beta$ and setting it to 0 allows us to derive the *normal equations*.

# Least squares estimators

- The normal equations can be expressed in matrix notation as

$$\mathbf{X}^\top \mathbf{W}\mathbf{X}\mathbf{b}_w = \mathbf{X}^\top \mathbf{W}\mathbf{Y}$$

- The least squares estimators are given by

$$\mathbf{b}_w = (\mathbf{X}^\top \mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{W}\mathbf{Y}$$

- The variance-covariance matrix is given by

$$Var(\mathbf{b}_w) = (\mathbf{X}^\top \mathbf{W}\mathbf{X})^{-1}$$

- Note that these estimators are *minimum variance unbiased*, consistent and sufficient.

- They are also maximum likelihood estimators (under normal errors)

# Determining weights

- In practice, the variances maybe *unknown* and need to be estimated
- Methods to determine weights:
  - Find the relationship between the *absolute* (or squared) residual and another variable and use this as a model for the variance.
  - Use *grouped* data or approximately grouped data to estimate the variance
- Recall that $\sigma_i^2 = \mathbb{E}(\epsilon_i^2) - [\mathbb{E}(\epsilon_i)]^2 = \mathbb{E}(\epsilon_i^2)$
- Hence, the squared residuals $e_i$ is an estimator of $\sigma_i^2$ and the absolute value is an estimator of $\sigma_i$

# Iterative re-weighted least squares (IRLS)

1. Fit the regression model using ordinary least squares.
2. Estimate the *variance function* using the residuals.
3. Use the fitted values from the estimated variance to obtain the weights $w_i$.
4. Estimate the regression coefficients using the weights.
5. Repeat Steps 2-4 until convergence.

- When the error variances are unknown the variance-covariance matrix of the estimated regression coefficients is estimated using

$$s^2(\mathbf{b}_w) = (\mathbf{X}^\top \hat{\mathbf{W}} \mathbf{X})^{-1}$$

- This value is used to make *confidence intervals* and perform hypothesis tests.

- If all weights are equal, the WLS estimators reduce to the ordinary least squares estimators.

# Dealing with outliers

- The method of least squares is *susceptible* to outliers, and they can result in incorrect fitted models.
- We discussed methods for detecting outliers, but no specific remedies for dealing with them.
    - One alternative is to *discard* potential outliers. However, this is not always a good idea.
    - Another alternative is to *down weight* their influence.

# Robust Regression

- *Robust regression* is a compromise between dropping outliers and including observations that may seriously violate the assumptions of OLS regression.

- IRLS robust regression is a form of weighted least squares regression where at each step weights are based on the size of the residuals.

- Outliers with large residuals will be down weighted to *decrease* their effect.

# IRLS Robust Regression

1. Choose a weight function.
2. Obtain starting weights for all observations.
3. Use the starting weights in WLS and obtain the residuals.
4. Use the residuals to obtain revised weights.
5. Repeat steps 3-4 until convergence.

# Weight Functions

- Huber weight function

$$w(u) = \begin{cases} 1 & \text{for } |u| \leq 1.345 \\ 1.345/|u| & \text{for } |u| > 1.345 \end{cases}$$

- Bi-square weight function

$$w(u) = \begin{cases} \left[1 - \left(\frac{u}{4.685}\right)^2\right]^2 & \text{for } |u| \leq 4.685 \\ 0 & \text{for } |u| > 4.685 \end{cases}$$

where $u$ is the scaled residual.

# Scaled residuals

- The weight functions are designed to be used with scaled residuals such as
$$e_i^* = \frac{e_i}{\sqrt{MSE}}$$
- However, *MSE* is *not* a resistant estimator of $\sigma$ and will be influenced by outliers.

- A resistant and robust estimator called the *median absolute deviation* (MAD) is often used instead:
$$MAD = \frac{1}{0.6745} median\{|e_i - median(e_i)|\}$$
- The constant 0.6745 is chosen to provide an unbiased estimate of $\sigma$ for independent observations from a normal distribution.
- The scaled residual used in robust regression is given by
$$e_i^* = \frac{e_i}{MAD}$$