Lecture Outline

- **Objective:** Regression criticism allows evaluating whether the results of an OLS model can be *generalized* towards an *unknown underlying population* or if they appear to be *just specific* to the observed sample.

  $\Rightarrow$ If all model assumptions are satisfied, then the sample based regression model applies to the underlying population. That is, the estimated model is ==representative of the *data generating process*== in the population.

- Chapter Overview:

  - OLS Assumptions (Hamilton Chapter 4)
  - New Terms: Efficient estimator, **BLUE** – ==Gauss-Markov Theorem==, homoscedasticity, large sample properties, consistency
  - Scatterplot matrix (relevant but trivial)
  - Basic time-series analysis
  - Residual versus Predicted Plot => ==heteroscedasticity==   unequal error term
  - Measures identifying temporal autocorrelation.

- o <mark>Non-normality</mark>

- o Multicollinearity   high collinear among independent variables

- o Influence Analysis: DFBetas

- o Other Case Statistics: Studentized Residuals, Leverage and Cook distance

OLS Assumptions

- **Role of assumptions:**   but always restrict our model (only focus in the study area)

  - o <mark>Simplify the complexity</mark> of the real world by imposing ***constraints*** onto the potential variability of reality (basically suppressing variability).

  - o This simplification ***accelerates* <mark>our capabilities to analyze the data</mark>**.

  - o Whenever possible, the ***plausibility*** of <mark>assumptions needs to be check.</mark>

  - o Ultimately, each <mark>model is a filter</mark> (an assumed structure) that just captures specific facets of variability.

- OLS has several outstanding properties if its assumptions are met.

  $\Rightarrow$ These advantages even hold under mild violations.

- Some violations can be addressed while remaining within the OLS modeling framework.

- The advantages of OLS are:

    o ***Simplicity*** of the calculation of the estimator **b** and its associated statistics.

    o ***Forced independence*** of residuals $Corr(\mathbf{x}_k, \mathbf{e}) = 0$ and $Corr(\hat{\mathbf{y}}, \mathbf{e}) = 0$.

    o ==***Ease of interpretation as partial effects for <u>purely</u> linear models***.==

    o Remember:

      [a] If a dependent variable is ***transformed*** the interpretation is only linear in the

      transformed regression system, but non-linear in the original scale of the variables.

      [b] The partial effects interpretation is no longer possible if we deal with ==interaction==

      ==effects==.

    o The ==disturbances== ε_i in $y_i = E(y_i) + \varepsilon_i$ represent the ==sum of several small errors==, which

      due to the ==central limit theorem== are ==expected to be normally distributed==.

    o Under general conditions, regression parameters are

        ▪ ***unbiased*** $(E[b_k] = \beta_k)$ and

means high precision

- **efficient**, that is, $Var(b_k) < Var(a_k) \Leftrightarrow E(b_k - \beta_k)^2 < E(a_k - \beta_k)^2$ for any alternative estimation rule a_k.

  Remember: Efficiency is only defined for **unbiased** estimators and

- **Consistent**, that is, $\lim\limits_{n\to\infty} \Pr(|b_k - \beta_k| < \delta) = 1$ with δ being a small positive constant. Therefore, for increasing sample sizes any potential bias vanishes and the standard error approaches zero.

○ Under the condition of i.i.d (**i**ndependent **i**dentically **d**istributed) and normal distributed disturbances,

  exact small sample tests for the estimates are available (e.g., *F*-test, *t*-test etc.).

  The term *small sample* means: we know the exact distribution of a test statistic for any given number *n* of observations.

- ***Regression criticism provides a mechanism to identify violations of the assumptions and point at potential cures of the problems that are induced by underlying model violations.***

Assumptions about the model structure

- Assumptions related to the expected values:

  - **[A1]** Regressors $\mathbf{X}$ are free of random effects or at least the randomness in an exogenous variable is <mark>uncorrelated with the model's disturbances</mark>.

  - **[A2]** <mark>$E(\varepsilon_i) = 0$ in the population</mark>.

    - These two assumptions lead to unbiased estimators b_k <u>if</u>

      [a] <mark>all **relevant** variables $\mathbf{X}$ are included</mark> in the model, or if

      [b] missing relevant variables are **uncorrelated** with those variables already included in the model.

      [c] In case that <mark>$E(\varepsilon_i) \neq 0$</mark> only the <mark>intercept</mark> b_0 <mark>will be biased</mark>. The remaining regression parameters stay unbiased.

  - Assumptions about the **variance** and **covariance** of the disturbances ε_i in the population:

    - **[A3]** <mark>Homoscedasticity</mark>: $Var(\varepsilon_i) = \text{constant} \; \forall i$ (The symbol $\forall$ means "for all")

▪ **[A4]** Independence among the disturbances:

$$Cov\left(\varepsilon_i,\varepsilon_j\right)=\begin{cases} 0 & \forall i \neq j \\ \sigma^2 & \forall i = j \end{cases} \quad \Leftrightarrow \quad Cov(\boldsymbol{\varepsilon})=\sigma^2 \cdot \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

- If above assumptions are satisfied then the **Gauss-Markov Theorem** holds, which states that OLS is **BLUE** (a **b**est **l**inear **u**nbiased **e**stimator).

- Discussion of alternative structures of the covariance matrix under:

  [a] Independence and Homoscedasticity => $Cov(\boldsymbol{\varepsilon})=\sigma^2 \cdot \mathbf{I}$

  [b] Heteroscedasticity $\Rightarrow$ diagonal would no longer be constant σ^2

  $$Cov(\boldsymbol{\varepsilon}) = \sigma^2 \cdot \begin{pmatrix} \omega_{11} & 0 & \cdots & 0 \\ 0 & \omega_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \omega_{nn} \end{pmatrix}$$

  [c] Autocorrelation $\Rightarrow$ off-diagonal elements would no longer be zero

$$Cov(\boldsymbol{\varepsilon}) = \sigma^2 \cdot \begin{pmatrix} \omega_{11} & \omega_{12} & \cdots & \omega_{1n} \\ \omega_{21} & \omega_{22} & \cdots & \omega_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{n1} & \omega_{n2} & \cdots & \omega_{nn} \end{pmatrix}$$

and the terms ω_{ij} measure the relationship (co-variance) between the ith and j^{th} observations

- **Problems:** we can only observe the estimated **sample** residuals $\mathbf{e} = \left(e_1, e_2, \ldots, e_n\right)^T$

  but not the underlying **population** disturbances $\boldsymbol{\varepsilon} = \left(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n\right)^T$.

  This complicates testing of hypotheses. For instance,

  (a) Assumptions [A1] and [A2] cannot be tested because **OLS estimation procedure guarantees** the properties of [A1] and [A2],

  (b) There is always some **baseline correlation** among the estimated regression residuals $\mathbf{e} = \mathbf{M} \cdot \mathbf{y}$ due to the exogenous projection matrix:

  $$Cov(\mathbf{e}) = \hat{\sigma}^2 \cdot \underbrace{\left(\mathbf{I} - \mathbf{X} \cdot \left(\mathbf{X}^T \cdot \mathbf{X}\right)^{-1} \cdot \mathbf{X}^T\right)}_{\mathbf{M}}.$$

- ▪ Assumptions about the ***distribution*** of the disturbances ε_i:

  - • The assumptions [A2], [A3], [A4] lead to the i.i.d. property.
    The ***central limit theorem*** allows for testing in large samples by using the resulting normal distribution. That is, we can use the normal distribution even if we do not make assumptions about the specific distributional form of the individual disturbances ε_i.

  - ▪ **[A5]** The disturbances are normal distributed $\varepsilon_i \sim N(0, \sigma^2)$

    - • Assumption **[A5]** allows using statistical tests for ***small samples*** such as the *t*- and *F*-tests, which are based on the normal distribution.

- • Problems arise when $Cov(\mathbf{X}, \boldsymbol{\varepsilon}) \neq \mathbf{0}$, because the estimates become ==***inconsistent***==, i.e., biased even for large sample sizes.
  The ==***instrumental variable***== approach potentially can recover the consistency in the estimates.

  assumption : instrumental variable are not correlated with error term

Problems associated with violations

- Discuss **HAM** Table 4.1 of problems based on violations of the OLS assumptions.

  Note:

  [a] Only if b_k is **unbiased** then b_k can be **efficient**.

  [b] *SE* refers to the estimate of the standard error of the disturbances, that is,

$$SE = \sqrt{\frac{\mathbf{e}^T \cdot \mathbf{e}}{n - K}} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n - K}}$$, which is used in

  many regression statistics.

Table 4.1 Some common statistical problems and their consequences for OLS

Problem	Biased b	Biased SE	Invalid t & F Tests	High Var[b] (Inefficient)[1]
Nonlinear relationship	yes	yes	yes	—
Omit relevant X	yes	yes	yes	—[2]
Include irrelevant X	no	no	no	yes
X measured with error	yes	yes	yes	—
Heteroscedasticity	no	yes	yes	yes
Autocorrelation	no	yes	yes	yes
X correlated with ε	yes	yes	yes	—
Nonnormal ε distribution	no	no	yes	yes
Multicollinearity	no	no	no	yes

[1] Inefficiency (Var[b]) noted only for unbiased estimators. For biased estimators a more general criterion, mean squared error (MSE = Var[b] + bias^2), reflects variation around the true parameter.
[2] Omission of relevant variables sometimes improves MSE, because Var[b] shrinks more than bias^2 grows.

Means we could not tell if the regression coefficient are significant

Identifying and Overcoming some Assumption Violations

- Perform bivariate scatterplots (or scatterplot matrix). See **HAM** Figure 4.1.

- First step of this exploratory analysis:

o Correlation coefficients **conceal** <u>non-linear</u> relationships and outliers. In contrast, visualization of the <mark>pairwise relationships</mark> in scatterplots is more informative.

o <mark>Make the dependent variable first</mark> in the list of metric variables.

o Exclude factors and investigate them separately in <mark>side-by-side</mark> box-plots with the dependent variable. Do not include factor(categorial variable)

o Evaluate the univariate distribution of each metric variable on the diagonal of the scatterplot matrix. Deal with non-linearity first (e.g., Box-Cox transformation, quadratic terms, interaction terms etc.) before investigating other model violations.

o Y-X combinations: potentially poor impact variables (*Change in # People*) potential heteroscedastisity, non-linearity and outliers

o <mark>X-X combinations: potential multicollinearity.</mark>
   X means all the independent variables here

Residuals versus Predicted Y Plots

- Recall: The OLS estimator is designed in a way that the ***residuals and the predicted values are linearly uncorrelated***. Thus, we would not expect to see any systematic pattern in their scatterplot.

- Uncovers problems: see **HAM Fig. 2.11** [a] for influential cases, [b] curvilinear relations, [c] non-normal residual distribution and [d] heteroscedasticity.
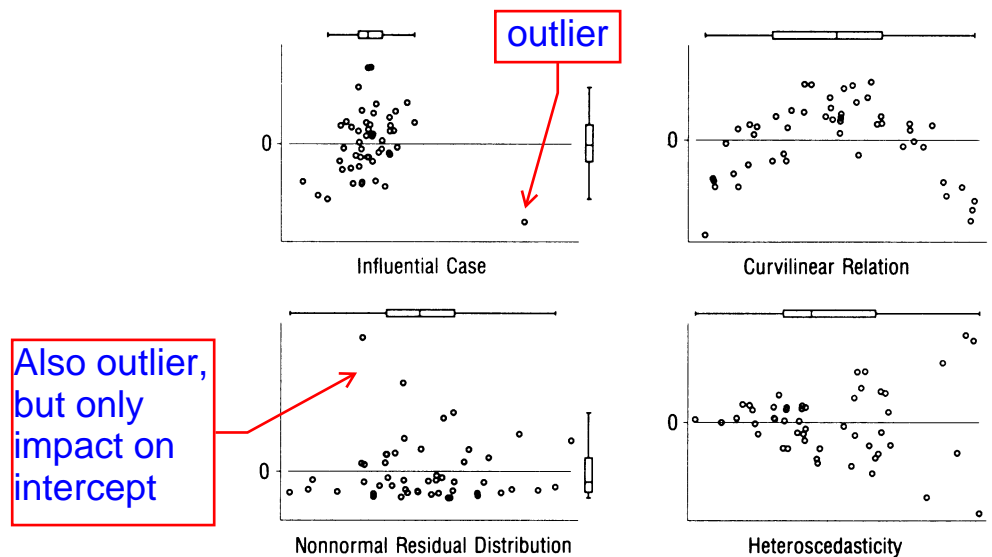

Figure 2.11   Examples of trouble seen in e-versus-$\hat{Y}$ plots (artificial data).

- ***Heteroscedasticity*** is associated with non-constant variance of the disturbances (as represented by the residuals). This non-constant variance may be systematically associated with [a] the level of an independent variable in the model X_j and, therefore, the predicted dependent variable $\hat{Y}$.

[b] ***any other variable*** not included in the regression model.

Example for heteroscedasticity: ***rates*** calculated in areas with <mark>a large population base usually have a lower variance than rates of areas with a smaller population base</mark>. This applies in particular to geographic units like census data:

- o A rate for observation i is defined as:

$$r_i = \frac{\text{\# of cases in } i \text{ satisfying a condition}}{\text{total population at risk in } i}$$

- o From the *binomial* distribution we know that its <mark>variance</mark> is:

$$Var(r_i) = \frac{\pi_i \cdot (1 - \pi_i)}{\text{total population at risk in } i}$$

- o Here the variance of an observed rate r_i depends on both the usually unknown π_i and the population at risk n_i.

- Heteroscedasticity leads to ***inefficient estimators*** (=> invalid *t*-tests and *F*-test). However, the estimated regression coefficients ***remain unbiased***.

Heteroscedasticity can be accommodated by weighted regression in a maximum likelihood

estimation context (a topic discussed later when working with the **F**easible **G**eneral **L**east **S**quares estimator)

- See **HAM** Figs 4.3 and 4.4 for heteroscedasticity. Large predicted water consumption is associated with a larger prediction error.
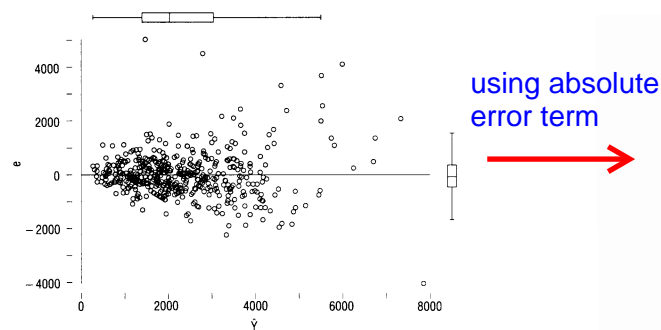

Figure 4.3   Residuals (e) versus predicted values ($\hat{Y}$) from regression of 1981 household water use on seven predictors.
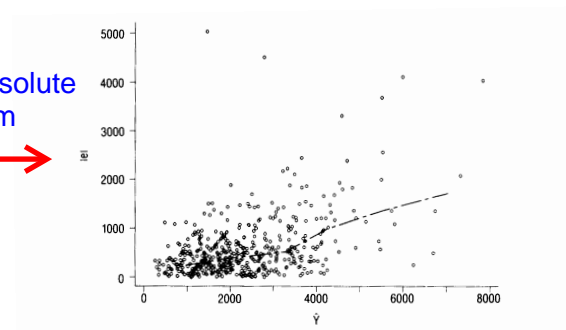
using absolute error term

Figure 4.4   Absolute residuals $|e|$ versus $\hat{Y}$, with band regression line indicating heteroscedasticity (household water-use regression).

- Heteroscedasticity usually leads to an error distribution with higher kurtosis (distribution with heavier tails).

Non-normality
- Problem: t and F-test become unreliable for small sample sizes.

- Problem: Observations in heavy-tails (outliers) have a substantially stronger impact on the estimates because their residuals enter **squared** into the OLS estimation equation: Let $e_i^* = 2 \cdot e_i$ Thus observation i has <u>4-times the weight</u> in OLS than it would if this residual just were e_i .   Dealing with outliers

- There are several ways to deal with these problems:

  [a] **transformations** to pull outlying observations in, then use OLS,

  [b] **down weighting** extreme observations and use of OLS,

  [c] **robust** non-parametric estimation methods (see HAM Chapter 6),

  [d] generate **Bootstrap simulations** to obtain the distribution of the estimated parameters and, therefore, their associated standard errors (see HAM Appendix 2).

Identifying Autocorrelation in Residuals

- In order to test for autocorrelation the observations must be **internal ordered**.

  o In the temporal context this order relates to the past influencing the presence which in turn will influence the future.

  o In a spatial context, neighboring observations may mutually influence each other.

- For serial autocorrelation see the appendix to these lecture notes.

Residuals

- How to standardize residuals?

  o A first choice would be to use $s_e^2 = RSS / (n - K)$. This <u>ignores</u> the increase in the variance of the residuals the further an observation move away from $\bar{X}$. Remember: the confidence interval around a regression line gets wider as we move away from $(\bar{X}, \bar{Y})$.

  o The variance of the residuals $e_i = y_i - \hat{y}_i$ is $Var(e_i) = s_e^2 \cdot (1 - h_{ii})$ where h_{ii} is the i-th element on the diagonal of the hat matrix $\mathbf{H} = \mathbf{X} \cdot (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T$

  o Therefore, the ***standardized residuals*** become: $z_i = \dfrac{e_i}{s_e \cdot \sqrt{1 - h_{ii}}}$. However, their exact distribution is unknown.

o The **studentized residuals** are $t_i = \dfrac{e_i}{s_{e(i)} \cdot \sqrt{1 - h_{ii}}}$. These residuals follow a t-distribution

  with $df = n - K - 1$ because numerator and denominator of t_i are independent.

  $s_{e(i)}$ is the estimated standard deviation of the residuals with the ith case being deleted.

o Studentized residuals are based on an augmented regression model:

  Sensitive for outliers
  e.g. Spatial filtering
  $$E(y_i) = \beta_0 + \beta_1 \cdot X_{i1} + \cdots + \beta_{K-1} \cdot X_{i,K-1} + \boxed{\delta_i \cdot I_i}$$ could use T-test on it

  where I_i is a dummy variable that is only 1 for the ith case and otherwise zero.

  This means that we treat the ith case as a special observation from a different

  regression regime. The inclusion of this dummy variable forces the residual of that

  particular case to be zero.

  The null hypothesis is $H_0 : \delta_i = 0$, which leads to the t-test mentioned above.

o We run into the problem performing multiple testing on the same dataset, which

  inflates the α-error.

  Therefore, one should use an adjusted α-error for the individual tests. The Bonferroni

  adjustment $\alpha^* = \alpha / (\# \ of \ tests)$ is one potential choice.

These adjusted significance levels are reported by the function

`car::influenceIndexPlot( )`   implement Borferroni adjustment

o Outlying observations are best investigated using studentized residuals.

independent means covariance equal to 0.

Leverage: Extreme observations with regards to the X variables

- The diagonal elements h_{ii} on the hat matrix $\mathbf{H} = \mathbf{X} \cdot (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T$ measure, in **relative terms**,

  the distance of the *i*-th observations from the center $(\overline{x}_1, \ldots \overline{x}_{K-1})^T$ of all independent variables.

- The potential range of the leverage is $h_{ii} = [1/n, 1]$ with a mean of $\overline{h}_{ii} = K / n$

- A leverage of $h_{ii} > 0.2$ is deemed risky. There is no test relevant, So using
  Cutting points

- An alternative critical value is $h_{ii} > 2 \cdot K / n$ with no more than 5% of the hat-values exceeding

  this value   Or, we could use histogram to explore outliers
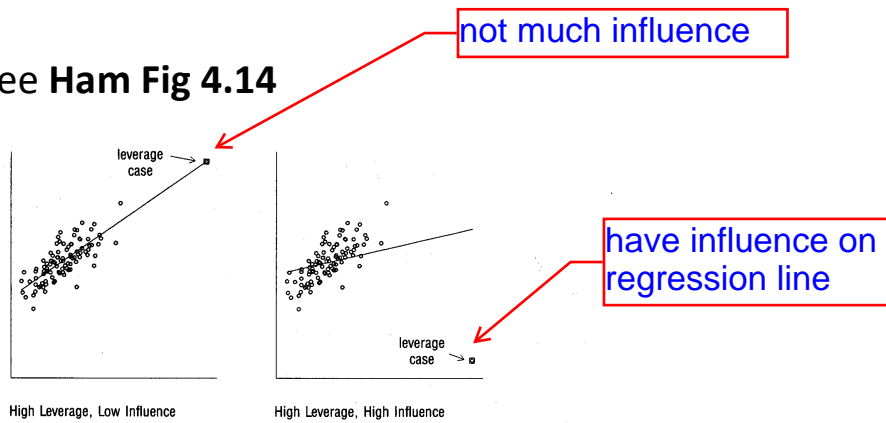
- See **Ham Fig 4.14**

Figure 4.14   "Good" (left) and "bad" (right) outliers: "bad" outliers influence the slope (artificial data).

Influential combinations of Y-X

- Deletion of an influential case changes the parameter estimates substantially.

- Influential case analysis takes all X variables and the Y variable simultaneously into account.

- Discuss **FOX Fig 11.1** and **Fig 11.4** for the impact of influential cases
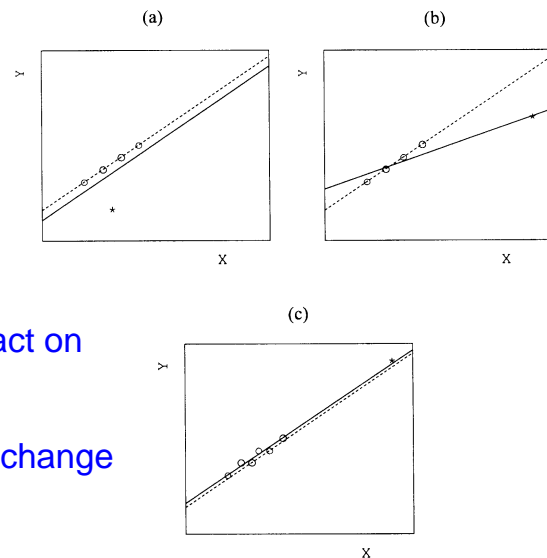
a: only impact on intercept
b: on both
c: does not change anything

Figure 11.1. Leverage and influence in simple regression. In each graph, the solid line gives the least-squares regression for all of the data, while the broken line gives the least-squares regression with the unusual data point (the asterisk) omitted. (*a*) An outlier near the mean of *X* has low leverage and little influence on the regression coefficients. (*b*) An outlier far from the mean of *X* has high leverage and substantial influence on the regression coefficients. (*c*) A high-leverage observation in line with the rest of the data does not influence the regression coefficients. In panel (c), the two regression lines are separated slightly for visual effect, but are, in fact, coincident.
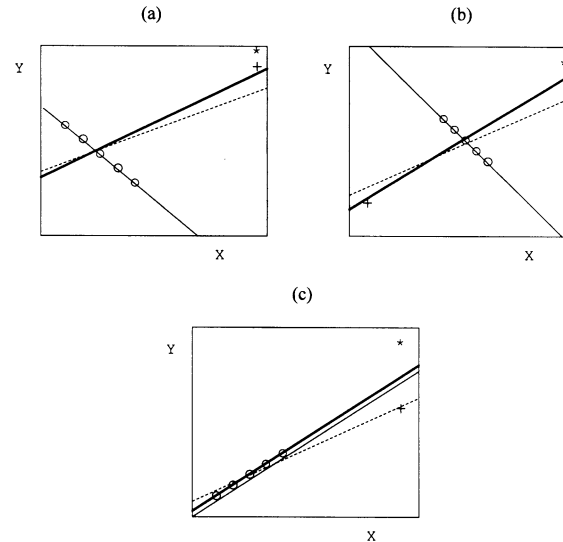
Figure 11.4. Jointly influential data in simple regression. In each graph, the heavy solid line gives the least-squares regression for all of the data; the broken line gives the regression with the asterisk deleted; and the light solid line gives the regression with both the asterisk and the plus deleted. (*a*) Jointly influential observations located close to one another: Deletion of both observations has a much greater impact than deletion of only one. (*b*) Jointly influential observations located on opposite sides of the data. (*c*) Observations that offset one another: The regression with both observations deleted is the same as for the whole dataset (the two lines are separated slightly for visual effect).

- ***DFBETAS***

  o Underlying idea: re-estimated the regression coefficients $b_{k(i)}$ with the *i*-th case deleted and standardize the difference $b_k - b_{k(i)}$ between regression coefficients.

- An individual variable *k* for each case *i* has a $DFBETAS_{ik} = \dfrac{b_k - b_{k(i)}}{s_{e(i)} / \sqrt{RSS_k}}$

- $DFBETAS_{ik}$ asks the question "By how many standard errors $s_{e(i)}/\sqrt{RSS_k}$ does b_k change, if we drop i^{th} case"

- Potential critical values:

  [a] External scaling: $|DFBETAS_{ik}| > 2/\sqrt{n}$

  [b] Internal scaling by Box-plots, gaps in the distribution etc.

- Proportional leverage plot with points weighted by $|DFBETAS_{ik}|$.

  Remember: leverage plots show the residuals of the dependent and independent variables after controlling for all other variables in the model.

- o Influential cases may cluster in the plot (**HAM Fig 4.11**).

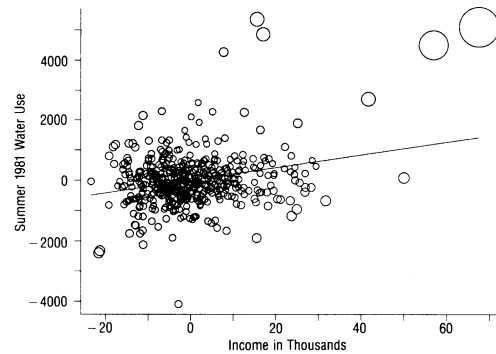Could plot on 3d, (DFBETA could be the Z value)

Figure 4.11 Proportional leverage plot (symbols proportional to DFBETAS) of 1981 household water use versus income, adjusting for five other predictors.
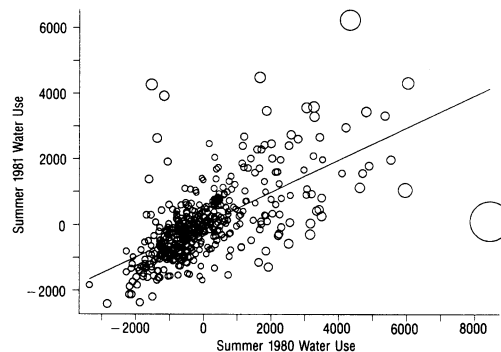
- o Counterbalancing cases. See **HAM Fig 4.12**.

Figure 4.12 Proportional leverage plot of 1981 household water use versus 1980 water use, adjusting for five other predictors.

- Cook's Distance (**HAM p 132**)

  - It measures the influence of the i-th case on the ***model as a whole*** rather than on individual partial regression coefficients.

    It combines the size of the ***standardized residuals*** and the ***leverage***:

    $$D_i = \frac{z_i^2 \cdot h_{ii}}{K \cdot (1 - h_{ii})}$$

    A critical value is $D_i > 4 / n$ Using the distance first (dealing with the whole dataset)

  - Technically the Cook's distance is equivalent to:

    $$D_i = \frac{\left(\mathbf{b} - \mathbf{b}_{(-i)}\right)^T \cdot (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \left(\mathbf{b} - \mathbf{b}_{(-i)}\right)}{K \cdot \hat{\sigma}^2}$$

Tukey Test

- The ***Tukey*** test evaluates whether by adding a ***squared term*** of an independent variable, which is already in the model, improves the model fit.

- It is implemented in the function `car::residualPlots( )`, which

  - Provides an plot of the quadratic function against the residuals

- A t-test, that evaluates whether the quadratic term is significant.

- The global Tukey test adds the squared predicted values (remember that the predicted value is a combination of all independent variables) to the model and performs a t-test, whether this term is significant.

- Note: A model, which was already adjusted according to the Tukey test, should **not be tested again**. Any significant outcome would mean that now a fourth power term should be added to the model by skipping any third order term. Powers, higher than the square, are difficult to interpret.

  Do only once (only using on the original variables)

Dealing with influential cases and outliers

- Hamilton example of using a collection of extreme value statistics:

Table 4.4 Case statistics for three influential households (see Figures 4.11 and 4.12)

Household i	Residual e_i	Leverage h_i	Standardized Residual z_i	Studentized Residual t_i	Cook's D_i	Income DFBETAS$_{i1}$	Water DFBETAS$_{i2}$
101	−4037	.08	−4.96[3]	−5.08[3]	.31[1]	0.06	−1.39[1]
127	3316	.06	4.03	4.09	.15[3]	0.98[2]	−0.18
134	3687	.09[3]	4.55	4.65	.29[2]	1.34[1]	0.25

[1] Absolute value of this statistic is largest in sample.
[2] Absolute value of this statistic is second-largest in sample.
[3] Absolute value of this statistic is third-largest in sample.

most important

In general, one should look for cases that are constistently "extreme".

- Overcoming "extreme" observations:

  o Search for additional explanatory variables, which may explain why particular cases appear to be extreme.

  o If a case appears to be affected by **measurement errors**,

    [a] "polish" the measurement,

    [b] weight the case down (Hamilton Chapter 6) or

    [c] drop the case.

o If a case appears to come from a ***different population***, drop the case or incorporate the structural difference through other means, e.g., dummy variables and interaction effects.

o Apply a data transformation.

Multicollinearity

- The ***tolerance*** of variable $\mathbf{x}_k$, relative to the other variables in the model, is the variance of $\mathbf{x}_k$ ***not shared*** by the other variables: $tol = 1 - R_k^2$.

  $\Rightarrow$ Large tolerances are better than smaller ones. The R_k^2 is the explained sum of squares of the kth independent variable $\mathbf{x}_k$ when regressed on all the remaining independent variables $\mathbf{X}_{[-k]}$.

- Strong bivariate correlation among the independent variables is only a first ***indicator for multicollinearity***.

- However, the ***correlation among the estimated regression parameters*** (see **HAM eq 4.16**) is a better indicator, because it takes simultaneously all pairwise correlations among the independent variables into consideration (because of the simultaneous structure of the inverse $\left(\mathbf{X}^T \cdot \mathbf{X} \right)^{-1}$).

- Effects of multicollinearity:

    o ***Inflation of standard error*** of parameter estimate $\Rightarrow t$-test may not reject $H_0 : \beta_k = 0$, even though it is different from zero

    o The estimated regression parameters may exhibit ***counterintuitive signs***.
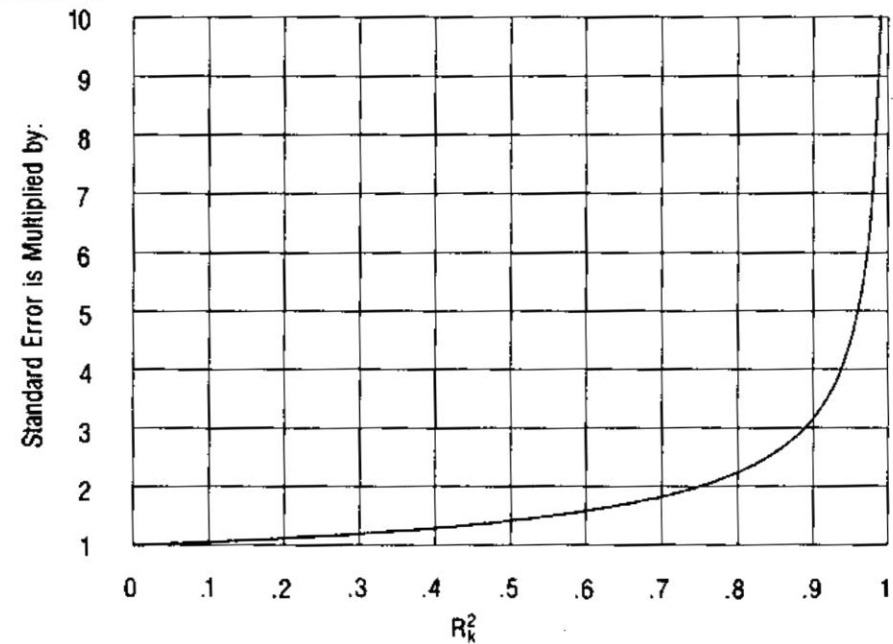

Figure 4.15   Effect of multicollinearity on standard errors (simplified).

High inflation factor represents our data is highly redundancy.(do not according to this to drop variables.)

- The standard error of b_k can be re-written as $SE_{b_k} = \dfrac{1}{\sqrt{(1-R_k^2)}} \cdot \dfrac{s_e}{\sqrt{TSS_k}}$ with the tolerance of the

  k-th independent variable $1-R_k^2$ and $TSS_k = \sum(x_{ik} - \bar{x}_k)^2$ the total sum of squares of the kth

  independent variable.

- Figure 4.15 shows how the standard error is inflated by increasing multicollinearity R_k^2.

- The term $VIF = 1/(1-R_k^2)$ is called the ***variance inflation factor***. Its impact on the standard error

  of b_k is displayed in **HAM Fig 4.15**.

  Note: **R**'s `car::vif( )` function reports the VIF in ***terms of variances*** and not standard

  deviations.

- How can we handle multicollinarity (see **HAM p 136**):

  drop one hard to interpret

  o Keep highly correlated variable in the model but ***be aware of their impact***.

  o ***Drop*** one or more offending variable because they are ***redundant***. But do not drop all.

  In general, keep that variable which you can interpret the most logical.

- o Combine variables into **Factor Scores** through principal component analysis (HAM Chapter 8).

- o **Collect more data**. This decreases the *VIF* due to an increase of $TSS_k = \sum \left(x_{ik} - \bar{x}_k \right)^2$.

Appendix: Serial Autocorrelation

See the online book in UTD's library Cowpertwait & Metcalfe (2009). Introductory Time Series with R. *Springer Verlag*

- Each time-series can be decomposed into two components:
  - o The expected value, such as cyclic behavior or trends, can be modeled with the use of exogenous information.

    This is called the **first order** component.
  - o The covariance in the random error terms leads to a stochastic random process. Here we want to learn from the sample observation about the internal covariance structure of the underlying data generating process.

    This is called the **second order** component.
- If the first order component is mis-specified then the covariance structure of the underlying data generating process cannot be properly identified.

What is autocorrelation?

What is autocorrelation and how does it differ from bivariate correlation?

- Hamilton's definition (page 118 top): "Autocorrelation refers to correlation between values of the same variable across different cases".

- We need to operationalize the concept of internal correlation among observations. This implies that our **observations are internally structured** (i.e., have an arrangement) and that the observations are **stochastically linked together** with respect to this internal structure.

Identifying Autocorrelation in Residuals

- In order to test for autocorrelation an **internal order** of the observations needs to be assumed.

    o In the temporal context this order relates to the past influencing the presence which in turn will influence the future.

    o In a spatial context, neighboring observations may mutually influence each other.

- In a time series the residuals $\mathbf{e} = \mathbf{M} \cdot \mathbf{y}$ are ordered sequentially with $\mathbf{e} = \left(e_{t_0}, e_{t_1}, e_{t_2}, \ldots, e_{t_n} \right)^T$ with $Cov\left(\varepsilon_{t_i}, \varepsilon_{t_{i+1}} \right) \neq 0$ for the underlying population disturbances.

- For strong positive autocorrelation consecutive residuals are similar, that is, $\left|e_{t_i} - e_{t_{i-1}}\right| \gtrsim 0$, and for strong negative autocorrelation $\left|e_{t_i} - e_{t_{i-1}}\right| \lesssim 2 \cdot \left|e_{t_i}\right|$ because consecutive residuals have alternating signs.

- This property of the difference between consecutive residuals gives rise the Durbin-Watson d-statistic:

$$d = \frac{\sum_{i=2}^{n}\left(e_{t_i} - e_{t_{i-1}}\right)^2}{\sum_{i=1}^{n} e_{t_i}^2}$$

- The Durbin-Watson statistic can be written as a ratio of quadratic forms. See the ®-script **DWTestInMatrixForm.R**.

- The expected value of the Durbin-Watson d-statistic under the assumption of **independence** is $E(d) = 2$.

  o An observed value $0 < d < 2$ indicates the presence of positive autocorrelation.

  o An observed value $2 < d < 4$ indicates the presence of negative autocorrelation.

- The distribution of the Durbin-Watson d-statistic under the assumption of independence is difficult to evaluate (see Table A4.4 in Hamilton).