1 Statistical Assumptions for Linear Models

1.1 Statistical Assumptions

The assumptions of multiple linear regression analysis are similar to those of the simple case involving only one independent variable. For point estimation, the principal assumptions are that

(1) the dependent variable is a continuous random variable ,

(2) the relationship between the several independent variables and the one dependent variable is *linear* (as opposed to quadratic or cubic - this is something we will explore more later).

Additional assumptions for statistical inference (estimation or hypothesis testing) are that

(3) the variances of the conditional distributions of the dependent variable, given various combinations of values of the independent variables, are all equal (*homoscedascity - something we will look at in a forthcoming lecture*),

(4) the conditional distributions of the dependent variable are normally distributed (*i.e. Residuals are nomally distributed*),

(5) the observed values of the dependent variable are independent of each other. (*Violation of this assumption is called autocorrelation.*

2 Model Validation

- Model validation is possibly the most important step in the model building sequence. It is also one of the most overlooked. Often the validation of a model seems to consist of nothing more than quoting the R^2 statistic from the fit (which measures the fraction of the total variability in the response that is accounted for by the model).
- Unfortunately, a high R^2 value does not guarantee that the model fits the data well. Use of a model that does not fit the data well cannot provide good answers to the underlying engineering or scientific questions under investigation.
- Model diagnostic techniques determine whether or not the distributional assumptions are satisfied, and to assess the influence of unusual observations.

2.1 Why Use Residuals?

If the model fit to the data were correct, the residuals would approximate the random errors that make the relationship between the explanatory variables and the response variable a statistical relationship. Therefore, if the residuals appear to behave randomly, it suggests that the model fits the data well. On the other hand, if non-random structure is evident in the residuals, it is a clear sign that the model fits the data poorly.

The subsections listed below detail the types of plots to use to test different aspects of a model and give guidance on the correct interpretations of different results that could be observed for each type of plot.

3 Introduction to Residuals

The difference between the observed value of the dependent variable (y) and the predicted value ($\hat{y}$) is called the **residual** (e). Each data point has one residual.

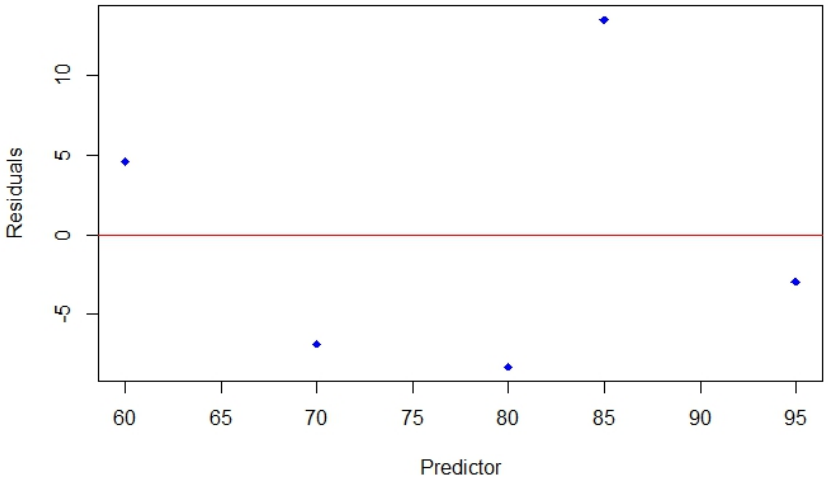
$$\text{Residual} = \text{Observed value} - \text{Predicted value}$$
$$e = y - \hat{y}$$

Both the sum and the mean of the residuals are equal to zero.

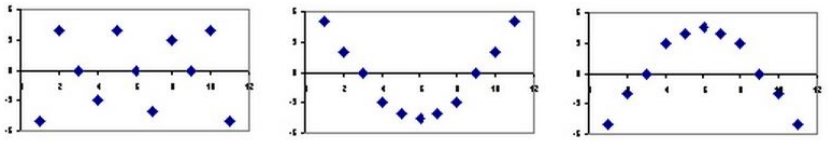
3.1 Residual Plots

- A residual plot is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis.
- If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate.
- Below the table on the left shows inputs and outputs from a simple linear regression analysis, and the chart on the right displays the residual (e) and independent variable (X) as a residual plot.

x	60	70	80	85	95
y	70	65	70	95	85
$\hat{y}$	65.411	71.849	78.288	81.507	87.945
e	4.589	-6.849	-8.288	13.493	-2.945



The residual plot shows a fairly random pattern - the first residual is positive, the next two are negative, the fourth is positive, and the last residual is negative. This random pattern indicates that a linear model provides a decent fit to the data.



Random pattern | Non-random: U-shaped | Non-random: Inverted U

- Below, the residual plots show three typical patterns.
- The first plot shows a random pattern, indicating a good fit for a linear model.
- The other plot patterns are non-random (U-shaped and inverted U), suggesting a better fit for a non-linear model.

Assumption of Constant Variance

Homoscedasticity

- ***Homoscedascity*** is the technical term to describe the variance of the residuals being constant across the range of predicted values.
- ***Heteroscedascity*** is the converse scenario : the variance differs along the range of values.

Heteroscedascity can be detected by inspecting the scatterplots.

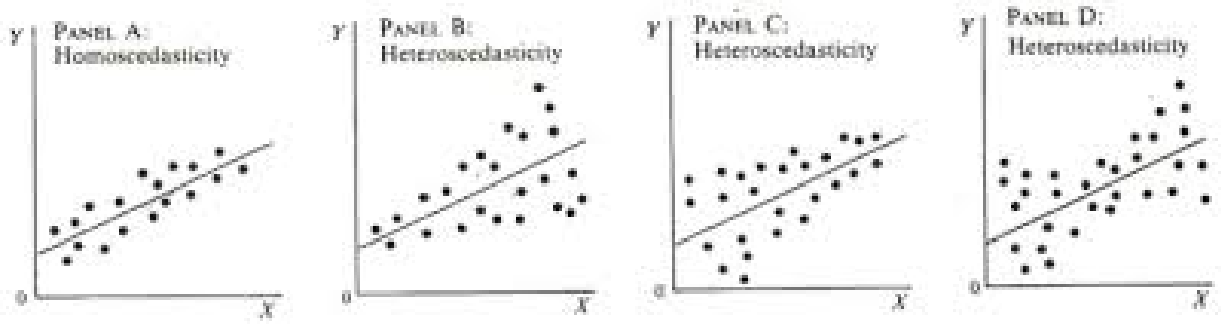


Fig. 9-1

You can also detect heteroscedasciity by inspecting the residual plots.

- Suppose you plot the individual residuals against the predicted value, the variance of the residuals predicted value should be constant.
- Consider the red arrows in the picture below, intended to indicate the variance of the residuals at that part of the number line. For the OLS summption to be valid , the length of the red lines should be more or less the same.

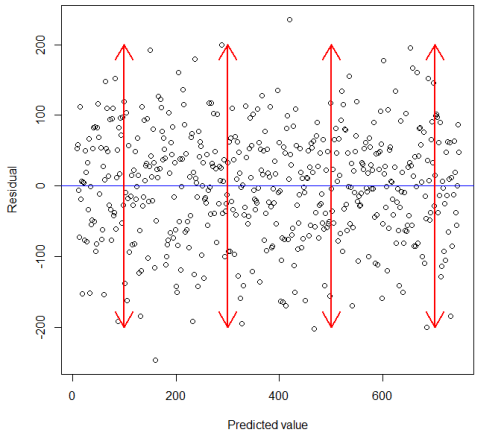


Figure 1:

```
>#  Evaluate homoscedasticity
>#  non-constant error variance test
> FitAll

Call:
lm(formula = Taste ~ Acetic + H2S + Lactic)

Coefficients:
(Intercept)       Acetic           H2S       Lactic
-28.8768       0.3277       3.9118      19.6705

> ncvTest(FitAll)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1.157465    Df = 1     p = 0.2819919
```

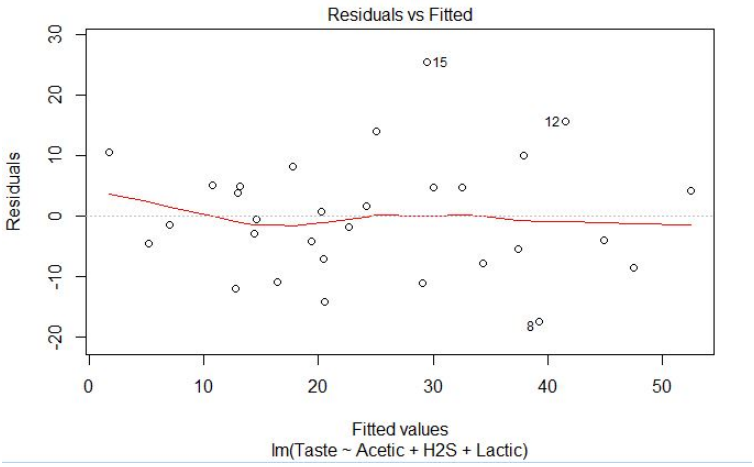


Figure 2:

4 Outliers and Influential Observations

> "Outliers are sample values that cause surprise in relation to the majority of the sample" (W.N. Venables and B.D. Ripley. 2002. Modern applied statistics with S. New York: Springer, p.119).

- Crucially, surprise is in the mind of the beholder and is dependent on some explicit model of the data.
- Importantly, Normality is only an assumption : There may be another model under which the outlier is not surprising at all, say if the data really are lognormal or gamma rather than normally distributed.

4.1 Regression Outliers

Data points that diverge in a big way from the overall pattern are referred to as "outliers". In the case of Simple Linear Regression, there are four ways that a data point might be considered an outlier.

- It could have an extreme X value compared to other data points.
- It could have an extreme Y value compared to other data points.
- It could have extreme X and Y values.
- It might be distant from the rest of the data, even without extreme X or Y values.

- After a regression line has been computed for a group of data, a point which lies far from the line (and thus has a large residual value) is known as an outlier. Such points may represent erroneous data, or may indicate a poorly fitting regression line.
- If a point lies far from the other data in the horizontal direction, it is known as an ***influential observation***. The reason for this distinction is that these points have may have a significant impact on the slope of the regression line.

Testing for Regression Outliers with `R`

- Suppose we have two unseen fitted models and we would like to see if there are any outliers.
- For this purpose, we can use `outlierTest()` from `library(car)` in `R`.
- In the first example, two outliers are detected.
- For the second fitted model, no outliers are detected. The most unusual observation, relative to the rest of the data set, is reported instead.

```
library(car)
outlierTest(fit1)

**Result:**
   rstudent unadjusted p-value  Bonferonni p
21    -4.12            4.39e-05        0.0209
15    -4.08            5.39e-05        0.0257

outlierTest(fit2)

**Result:**
No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
       rstudent unadjusted p-value      Bonferonni p
177    -2.52             0.0119                NA
```

The row numbers (here : 21, 15 and 177) indicate the outlier points in the data.

Autocorrelation

- Adjacent residuals should not be correlated with each other (**autocorrelation**).
- If you can use one residual to predict the next residual, there is some predictive information present that is not captured by the predictors.
- Typically, this situation involves time-ordered observations.

- For example, if a residual is more likely to be followed by another residual that has the same sign, adjacent residuals are positively correlated.
- You can include a variable that captures the relevant time-related information, or use a time series analysis.

Durbin-Watson Test for Autocorrelated Errors

The ***Durbin-Watson*** procedure is commonly used to to test for autocorrelation of residuals. To perform this test, we use the `durbinWatsonTest()` from the car R package. All you have to do is to specify the name of the fitted mode.

```
 FitMod <- lm(mpg~wt+cyl,data=mtcars)

# library(car)
durbinWatsonTest(FitMod)
```

The null hypothesis can simply be stated as "There is no autocorrelation" present. The R code output provides a $p-$value to base a determination on.

```
> durbinWatsonTest(FitMod)
 lag Autocorrelation D-W Statistic p-value
   1       0.1302185      1.671096   0.252
 Alternative hypothesis: rho != 0
```

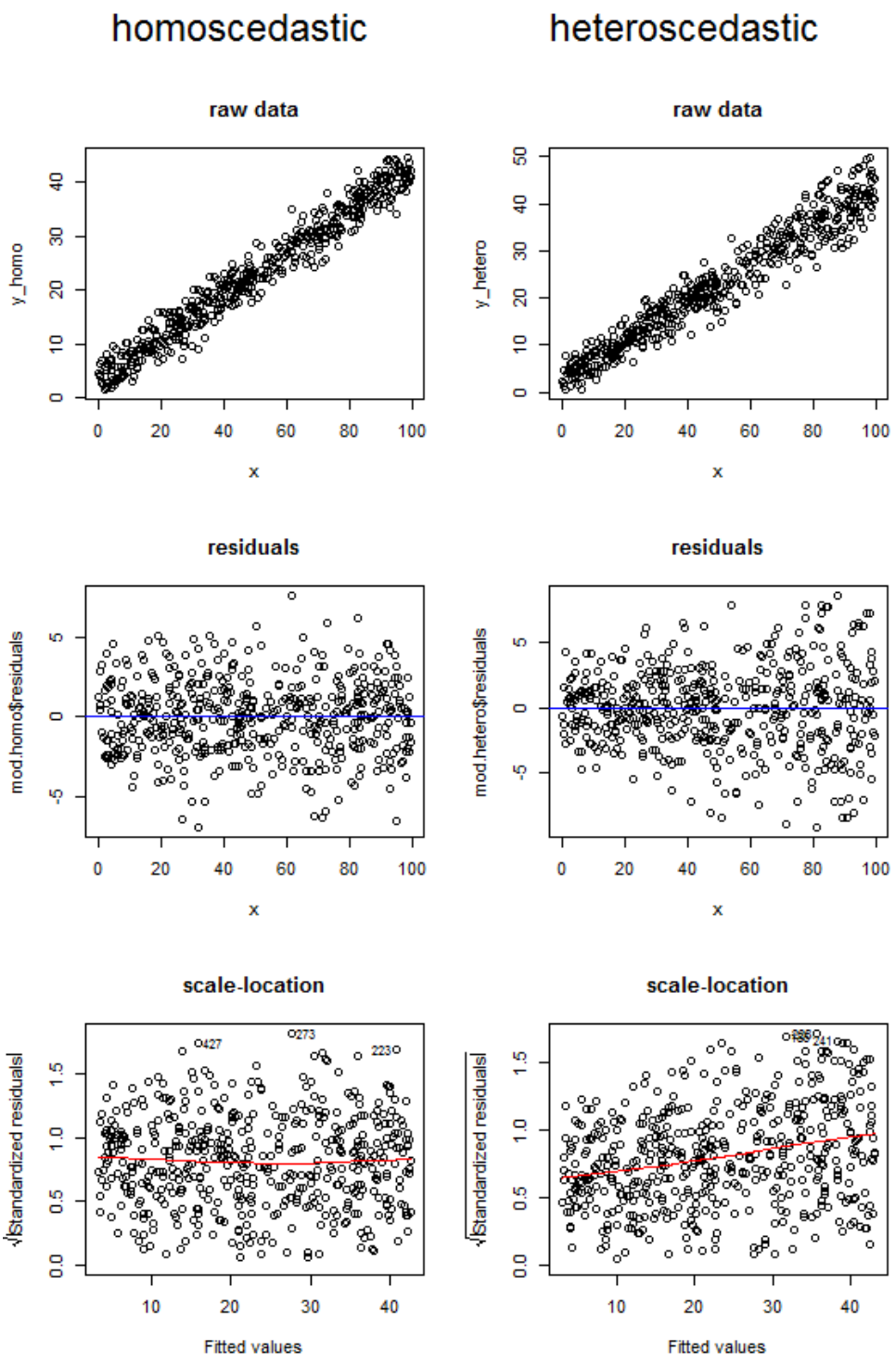
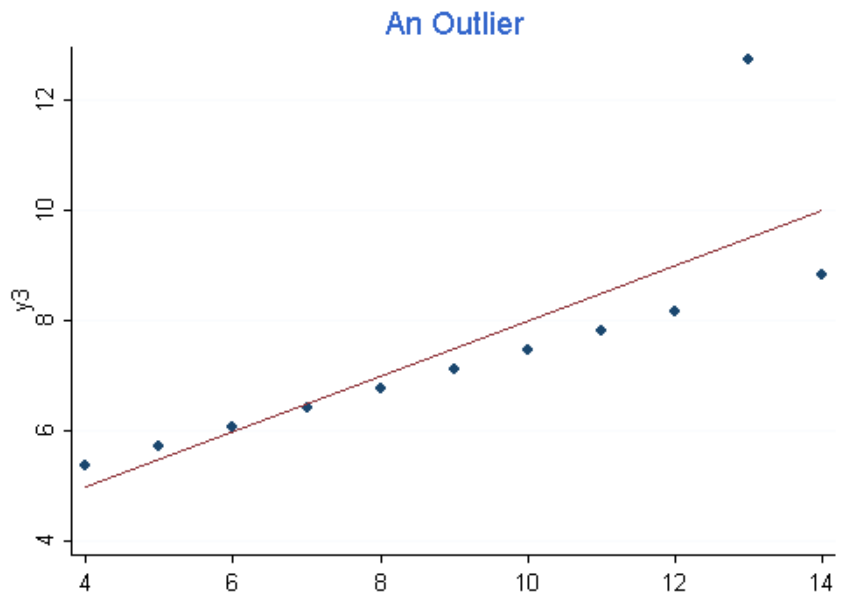

homoscedastic
heteroscedastic
raw data
raw data
y_homo
y_hetero
x
x
residuals
residuals
mod.homo$residuals
mod.hetero$residuals
x
x
scale-location
scale-location
√|Standardized residuals|
√|Standardized residuals|
427
273
223
Fitted values
Fitted values

Figure 3:



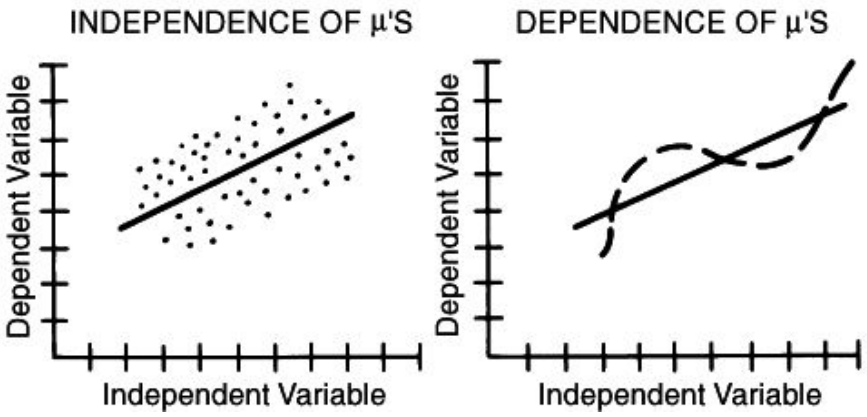


Figure 4: (disregard the titles)