NOVA SCHOOL OF
SCIENCE & TECHNOLOGY

**Data Analysis and Mining**

***Master in Analysis and Engineering of Big Data***

# Linear Regression

2023/2024

(2nd semester)

Susana Nascimento
(snt@fct.unl.pt)

Susana Nascimento (snt@fct.unl.pt)

# Summary

1. The Least Squares Estimates
2. The Correlation & Determination Coefficients: properties
3. The Regression Model
4. Inference in Regression
5. Verifying Regression Assumptions

Susana Nascimento (snt@fct.unl.pt)

# Example of Simple Linear Regression

- *Cereals* data set contains nutritional information for 77 cereals
- Includes *sugars* and *rating* variables

| Cereal Name | Manuf | Sugars | Calories | Protein | Fat | Sodium | Rating |
|---|---|---|---|---|---|---|---|
| 100% Bran | N | 6 | 70 | 4 | 1 | 130 | 68.4030 |
| 100% Natural Bran | Q | 8 | 120 | 3 | 5 | 15 | 33.9837 |
| All-Bran | K | 5 | 70 | 4 | 1 | 260 | 59.4255 |
| All-Bran Extra Fiber | K | 0 | 50 | 4 | 0 | 140 | 93.7049 |
| Almond Delight | R | 8 | 110 | 2 | 2 | 200 | 34.3848 |

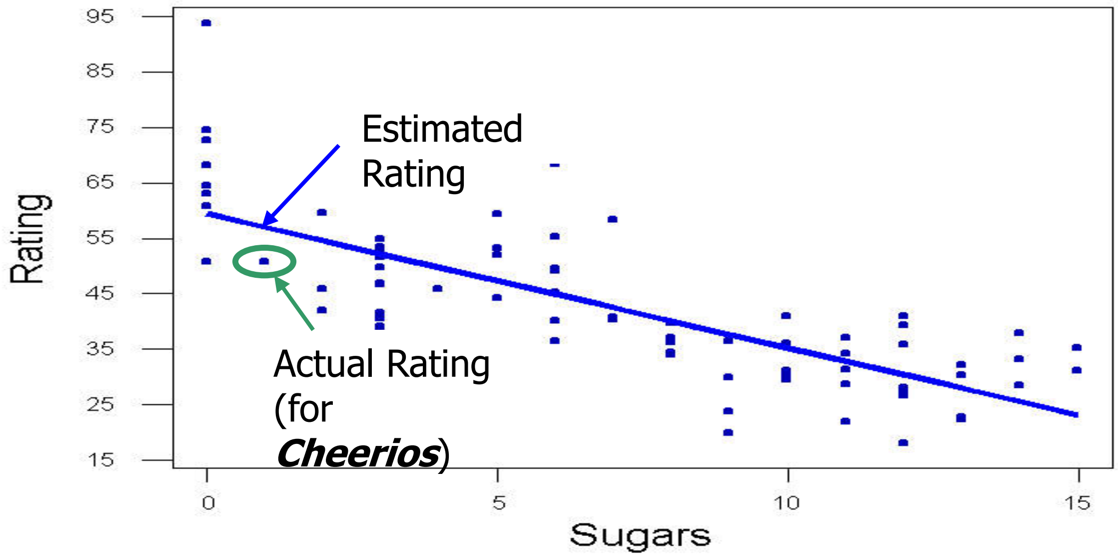
- Let's estimate *nutritional rating* of a cereal, given its sugar content

  Use '*sugars*' (***predictor***) to estimate '*rating*' (***response***)

Susana Nascimento (snt@fct.unl.pt)

# Example of Simple Linear Regression

- Scatter plot of *rating vs. sugars* and least squares regression line




- *Estimated Regression Equation* (ERE)

$$\hat{y} = b_0 + b_1 x$$

Susana Nascimento (snt@fct.unl.pt)

# Estimated Regression Equation

$$\hat{y} = b_0 + b_1 x$$

- $\hat{y}$ estimated value of response variable
- $b_0, b_1$ *regression coefficients*
- $b_0$ *y-intercept* of regression line
- $b_1$ *slope* of regression line

- ERE for *Cereals* data set

$$\hat{y} = 59.4 - 2.42(Sugars)$$

*Estimated cereal rating equals 59.4 minus 2.42 times the sugar content in grams*

- **ERE is used to make estimates or predictions**

Susana Nascimento (snt@fct.unl.pt)

# Making predictions

$$\boxed{\hat{y} = 59.4 - 2.42(Sugars)}$$

- Use ERE to estimate rating for ***new*** *cereal* containing 1 gram of sugar

$$\hat{y} = 59.4 - 2.42(1) = 56.98.$$

- Estimated value lies directly on ERE at ($x$ = 1, $\hat{y}$ = 56.98)
- Data set contains cereal with 1 gram of sugar (Cheerios)
- Cheerios rating = 50.765 at ($x$ = 1, $y$ = 50.765)

Susana Nascimento (snt@fct.unl.pt)

# Example of Simple Linear Regression *(cont'd)*

ERE prediction too high by 56.98 – 50.765 = 6.215 rating points?

- Difference (y - ŷ) = 56.98 – 50.765 known as *prediction error* or *residual*
- One seeks ERE that minimizes the size of residuals
- ***Least Squares Regression*** calculates unique ERE

Susana Nascimento (snt@fct.unl.pt)

# The Least Squares Estimates

- Consider a second data sample of 77 cereals
- Cannot assume the Cereals' ERE $\hat{y} = 59.4 - 2.42(Sugars)$
- $b_0$ and $b_1$ statistics whose values differ from sample to sample
- Requires population parameters $\beta_0$ and $\beta_1$
- ***Regression Equation*** represents true linear relationship between *rating* and *sugars* for ***all cereals***

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Susana Nascimento (snt@fct.unl.pt)

# The Least Squares Estimates *(cont'd)*

$$\boxed{y = \beta_0 + \beta_1 x + \varepsilon}$$

- Suppose two or more cereals have the same nutrition rating but different sugar content
  - Error term $\varepsilon$ accounts for the *indeterminacy* in model
  - Residuals ($y_i$ - $\hat{y}$) are estimates of error terms $\varepsilon_i$ , $i$ = 1, ..., $n$
- Use Least Squares method to derive values for $\beta_0$ and $\beta_1$

Susana Nascimento (snt@fct.unl.pt)

# The Least Squares Estimates *(cont'd)*

$$\boxed{y = \beta_0 + \beta_1 x + \varepsilon}$$

- Consider *n* observations from the model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n$$

- Least squares line minimizes population sum of squares

$$SSE_p = \sum_{i=1}^{n} \varepsilon_i^2$$

$$SSE_p = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

- Determine optimal $\beta_0$ and $\beta_1$ minimizing

$$\sum_{i=1}^{n} \varepsilon_i^2$$

Susana Nascimento (snt@fct.unl.pt)

# The Least Squares Estimates *(cont'd)*

- Partial derivatives with respect to $\beta_0$ and $\beta_1$

$$\frac{\partial\, SSE_p}{\partial\, \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{\partial\, SSE_p}{\partial\, \beta_1} = -2\sum_{i=1}^{n}x_i(y_i - \beta_0 - \beta_1 x_i)$$

- Estimates for $b_0$ and $b_1$ parameters

$$\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) = 0$$

$$\sum_{i=1}^{n}x_i(y_i - b_0 - b_1 x_i) = 0$$

Susana Nascimento (snt@fct.unl.pt)

# The Least Squares Estimates *(cont'd)*

$$\sum_{i=1}^{n}\left(y_i - b_0 - b_1 x_i\right) = 0$$
$$\sum_{i=1}^{n} x_i\left(y_i - b_0 - b_1 x_i\right) = 0$$

- Equations re-expressed

$$b_0 n + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$
$$b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

- Solving for $b_0$ and $b_1$ yields

$$b_0 = \bar{y} - b_1 \bar{x} \qquad b_1 = \frac{\sum x_i y_i - \left[\left(\sum x_i\right)\left(\sum y_i\right)\right] / n}{\sum x_i^2 - \left(\sum x_i\right)^2 / n}$$

with $n$ = total number of observations

Susana Nascimento (snt@fct.unl.pt)

# The Least Squares Estimates

- Estimations to 77 observations in *Cereals* data set

$$b_0 = \bar{y} - b_1\bar{x} = 42.6657 - 2.42(6.935) = 59.4$$

$$b_1 = \frac{\sum x_i y_i - \left[\left(\sum x_i\right)\left(\sum y_i\right)\right]/n}{\sum x_i^2 - \left(\sum x_i\right)^2/n} = \frac{19{,}186.7 - (534)(3285.26)/77}{5190 - (534)^2/77}$$

$$= \frac{-3596.791429}{1486.675325} = -2.42$$

$$\boxed{\hat{y} = 59.4 - 2.42(Sugars)}$$

- $b_0$ is $y$-intercept
- $b_1$ is slope for ***Estimated Regression Equation*** (ERE)

Susana Nascimento (snt@fct.unl.pt)

# ERE: interpretation

$$\boxed{\hat{y} = 59.4 - 2.42(Sugars)}$$

- ERE estimates a cereal with zero grams of sugar to have 59.4 rating points ($b_0$ = 59.4)
- Interpretation for *y*-intercept when predictor value is zero may be meaningless
- **Example**: *predicting person's weight based on height?*
  - Height = 0 unclear
  - ***y* - intercept cannot be interpreted**

Susana Nascimento (snt@fct.unl.pt)

# ERE: interpretation

$$\boxed{\hat{y} = 59.4 - 2.42(Sugars)}$$

- Slope of ERE measures change in *y* per unit increase *x*

"*For each increase in one gram in sugar content, the estimated nutritional rating decreases by 2.42 rating points*"

Susana Nascimento (snt@fct.unl.pt)

# Summary

1. The Least Squares Estimates
2. Correlation & Determination Coefficients: properties
3. The Regression Model
4. Inference in Regression
5. Verifying Regression Assumptions

Susana Nascimento (snt@fct.unl.pt)

# The Coefficient of Determination $r^2$

- Least Squares Regression method can approximate linear relationship between any two variables
- ***How useful is the regression line?***
- ***Useful for making predictions?***
- **Coefficient of Determination, *$r^2$ statistic*,**

  **measures ERE's *goodness of fit to data***

Susana Nascimento (snt@fct.unl.pt)

# Coefficient of Determination $r^2$

- Orienteering data set (#10) with ERE ŷ = 6 + 2x
- Measures elapsed time and distance traveled by hikers
- Residual ($y$ - $\hat{y}$) and residual squared $(y - \hat{y})^2$ shown

| Subject | $X$ = Time | $Y$ = Distance | Predicted Score $\hat{y} = 6 + 2x$ | Error in Prediction $(y - \hat{y})$ | (Error in Prediction)$^2$ $(y - \hat{y})^2$ |
|---|---|---|---|---|---|
| 1 | 2 | 10 | 10 | 0 | 0 |
| 2 | 2 | 11 | 10 | 1 | 1 |
| 3 | 3 | 12 | 12 | 0 | 0 |
| 4 | 4 | 13 | 14 | -1 | 1 |
| 5 | 4 | 14 | 14 | 0 | 0 |
| 6 | 5 | 15 | 16 | -1 | 1 |
| 7 | 6 | 20 | 18 | 2 | 4 |
| 8 | 7 | 18 | 20 | -2 | 4 |
| 9 | 8 | 22 | 22 | 0 | 0 |
| 10 | 9 | 25 | 24 | 1 | 1 |
| | | | | | $SSE = \sum_{i=1}^{10} (y_i - \hat{y}_i)^2 = 12$ |

Susana Nascimento (snt@fct.unl.pt)

# The Coefficient of Determination $r^2$

- *Sum of Squares Error* $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = 12$
- Represents overall measure of error in ERE's prediction

***Is SSE = 12 large or small?***

***What can we compare it to?***

- Assume we estimate *distance* traveled ($y$) without knowledge of *time* ($x$)

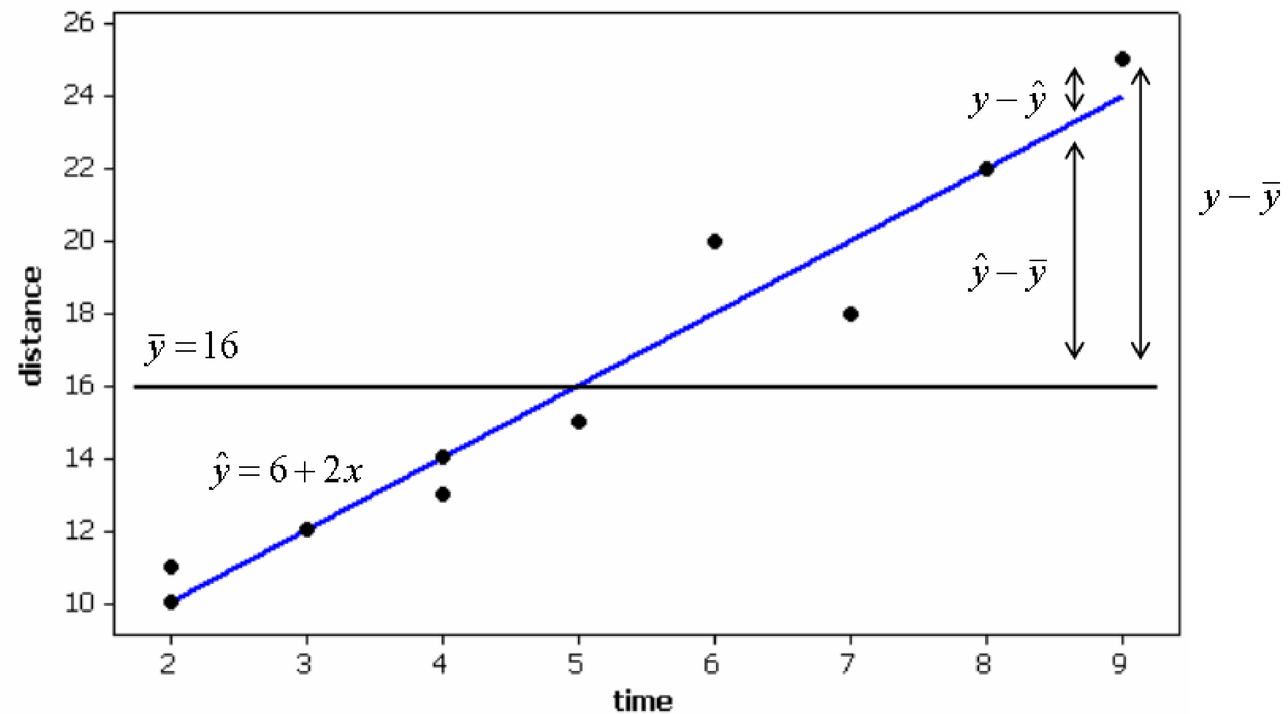
  we lack access to predictor information

- Less information available usually results in less accurate estimates

Susana Nascimento (snt@fct.unl.pt)

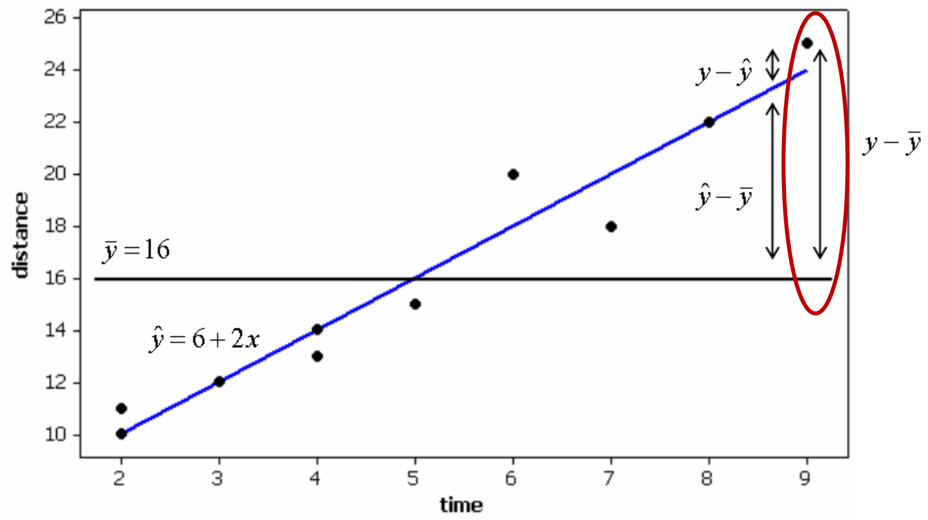
# The Coefficient of Determination $r^2$

- Now, $\bar{y}$= 16 best estimate for ***all*** competitors, regardless of time hiking
- Not an optimal prediction



Susana Nascimento (snt@fct.unl.pt)

# The Coefficient of Determination $r^2$




- Data points cluster more tightly along ERE, as compared to line $\bar{y}$ = 16
  - Residuals smaller when using *time* (x-information)
- Consider competitor #10 where ignoring *x*-value leads to estimation error (25 - 16) = 9 km
- Indicated by distance between $\bar{y}$ = 16 and data point (9, 25)
- Suppose estimation error for all data points are calculated similarly

Susana Nascimento (snt@fct.unl.pt)

# The Coefficient of Determination $r^2$

- Leads to ***Sum of Squares Total***

$$SST = \sum_{i=1}^{n} (y - \bar{y})^2$$

- Measures total variability in response variable, without reference to predictor variable

- SST univariate measure of *y*

$$SST = \sum_{i=1}^{n} (y - \bar{y})^2 = (n-1)Var(y) = (n-1)(SD(y))^2$$

# The Coefficient of Determination $r^2$ *(cont'd)*

- **Is SST larger or smaller than SSE?**

  SST = 228 much larger than SSE = 12
- Smaller values for SSE better
  - So, including predictor improves estimates
- **How much does SSE improve estimates?**
  - For competitor #10, by ignoring *x*-value leads to estimation error $(y - \bar{y}) = (25 - 16) = 9$ km
  - Including *x*-value in regression produces $(y - \hat{y}) = (25 - 24) = 1$ *km* estimation error
  - Improvement: $(\hat{y} - \bar{y}) = 24 - 16 = 8$

Susana Nascimento (snt@fct.unl.pt)

# The Coefficient of Determination $r^2$

- *Sum of Squares Regression*, SSR measures overall improvement in prediction accuracy

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

- Consider SST = SSR + SSE:

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

***SSR*** measures the amount of variability in the response explained by Estimated Regression Equation (ERE)

***SSE*** measures variability in *y* from all other sources, after linear relationship between *x* and *y* accounted for

Susana Nascimento (snt@fct.unl.pt)

# The Coefficient of Determination $r^2$

- Coefficient of Determination

$$r^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

- $r^2$ is proportion of variability in response variable explained by ERE

- **Maximum** of $r^2$ occurs when all data points lie exactly on ERE (perfect fit)

$r^2$ = SSR/SST = 1 (with SSE= 0)

- **Minimum** of $r^2$ occurs when ERE shows no improvement
  - ERE explains no variability in response variable
  - SSR = 0, resulting in $r^2$ = 0/SST = 0

Susana Nascimento (snt@fct.unl.pt)

# The Coefficient of Determination $r^2$

$$r^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- Interpreting $r^2$
  - Score above 90% -- "*very good*"
  - Score above 80% -- "*good*""
  - Score above 70% -- "*somewhat satisfied*"
  - Score below 50% -- "*bad*"

- **Heuristic general guideline only**

  - it requires judgment

Susana Nascimento (snt@fct.unl.pt)

# The Standard Error of the Estimate

- Mean Square Error (*MSE*)

$$MSE = SSE/(n - m - 1)$$

$m$ = number predictors, $n$ = number observations

- Standard Error of the Estimate *s* measures accuracy of estimates produced by *ERE*

$$s = \sqrt{MSE} = \sqrt{SSE/(n - m - 1)}$$

- *s* 'typical' **residual or error in estimation**
  - From Orienteering data $s$ = 1.2 km
  - Using ERE estimate of hiking distance typically differs from actual distance by about 1.2 km

Susana Nascimento (snt@fct.unl.pt)

# The Standard Error of the Estimate

- Consider typical estimation error when ignoring predictor variable

$$SD_y = \frac{\sum_{i=1}^{n} (y - \bar{y})^2}{n-1} = 5.0$$

- Using ERE reduces "typical" estimation error from 5 *km* to 1.2 *km*
- Calculation of ***SST*** and ***SSR***

$$SST = \sum y^2 - \left(\sum y\right)^2 / n$$

$$SSR = \frac{\left[\sum xy - \left(\sum x\right)\left(\sum y\right)/n\right]^2}{\sum x^2 - \left(\sum x\right)^2 / n}$$

***Sum of Squares Regression***

Susana Nascimento (snt@fct.unl.pt)

# The Standard Error of the Estimate

- *SST* and *SSR* calculated from Orienteering data set

$$SST = \sum y \;\; - \left(\sum y\right)^2 / n = 2788 - (160)^2 / 10 = 2478 - 2560 = 228$$

$$SSR = \frac{\left[\sum xy - \left(\sum x\right)\left(\sum y\right) / n\right]^2}{\sum x^2 - \left(\sum x\right)^2 / n} = \frac{[908 - (50)(160)/10]^2}{304 - (50)^2 / 10} = \frac{108^2}{54} = 216$$

- *ERE* accounting of variability in distance traveled

$$r^2 = \frac{SSR}{SST} = \frac{216}{228} = 0.9474$$

Susana Nascimento (snt@fct.unl.pt)

# The Correlation Coefficient

- *Measures* strength of linear relationship between two quantitative variables

$$r = \frac{\sum \left(x - \bar{x}\right)\left(y - \bar{y}\right)}{(n-1)\ s_x s_y}$$

  - $-1.0 \le r \le 1.0$,
  - $s_x$ and $s_y$ are sample standard deviations for *x* and *y*, respectively

- *r* close to 1: variables are ***positively*** correlated
- *r* close to -1: variables are ***negatively*** correlated
- Other values: ***uncorrelated*** variables

Susana Nascimento (snt@fct.unl.pt)

# The Correlation Coefficient *(cont'd)*

- *Rough guidelines* for interpreting correlation between two variables
  - $r > 0.7$ positively correlated
  - $0.33 < r \leq 0.7$ mildly positively correlated
  - $-0.33 < r \leq 0.33$ not correlated
  - $-0.7 < r \leq -0.33$ mildly negatively correlated
  - $r \leq -0.7$ negatively correlated

- **Obs.**: presence/absence of correlation requires more rigorous tests!
  - depending on the field of study

Susana Nascimento (snt@fct.unl.pt)

# The Correlation Coefficient *(cont'd)*

- Definition of *r*

$$r = \frac{\sum xy - (\sum x)(\sum y)/n}{\sqrt{\sum x^2 - (\sum x)^2/n}\ \sqrt{\sum y^2 - (\sum y)^2/n}}$$

- *Calculate and interpret r for* Orienteering data set

$$r = \frac{908 - (50)(160)/10}{\sqrt{304 - (50)^2/10}\ \sqrt{2788 - (160)^2/10}} = \frac{108}{\sqrt{54}\sqrt{228}} = 0.9733$$

*r* indicates that time and distance hiking are strongly positively correlated

- *r* conveniently expressed as $r = \pm\sqrt{r^2}$
  - *r* is positive when $b_1$ is positive
  - *r* is negative when $b_1$ is negative
- From Orienteering example: $r = \sqrt{r^2} = \sqrt{0.9474} = 0.9733$

Susana Nascimento (snt@fct.unl.pt)

# ANOVA : Simple Linear Regression

- Regression statistics summarized in *Analysis of Variability* (ANOVA) table

| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Square | F |
|---|---|---|---|---|
| Regression | $SSR$ | $m$ | $MSR = \frac{SSR}{m}$ | $F = \frac{MSR}{MSE}$ |
| Error (or Residual) | $SSE$ | $n-m-1$ | $MSE = \frac{SSE}{n-m-1}$ | |
| Total | $SST = SSR + SSE$ | $n-1$ | | |

  - *m* = total predictors, *n* = total observations
  - *F* statistic used for inferential purposes

Susana Nascimento (snt@fct.unl.pt)

# Summary

1. The Least Squares Estimates
2. Correlation & Determination Coefficients: properties
3. The Regression Model
4. Inference in Regression
5. Verifying Regression Assumptions

Susana Nascimento (snt@fct.unl.pt)

# The Regression Model

- ***For model building and inference purposes, regression model assumptions require validation***
- Deploying model with unverified assumptions can result in failure!

Susana Nascimento (snt@fct.unl.pt)

# The Regression Model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- $\beta_0$ and $\beta_1$ model parameters are $y$-intercept and slope, respectively.
- True values unknown, estimated by Estimation of Regression Equation (ERE).
- $\varepsilon$ error term required because linear approximation to actual predictor-response relationship not deterministic.
- $\varepsilon$ is a random variable.

Susana Nascimento (snt@fct.unl.pt)

# Error Term $\varepsilon$: Assumptions

1. **Zero Mean Assumption**

   - Error term ε random variable with mean, $E(\varepsilon) = 0$

2. **Constant Variance Assumption**

   - Variance of ε constant, regardless of $x$-value

3. **Independence Assumption**

   - Values of ε independent

4. **Normality Assumption**

   – Error term ε normally distributed random variable

- So, $\varepsilon_i$ are independent normal random variables, with *mean* = 0 and constant variance

Susana Nascimento (snt@fct.unl.pt)

# The Regression Model *(cont'd)*

- Implied Behavior of Response Variable *y*

1. Based on: **Zero Mean Assumption**

$$E(y) = E(\beta_0 + \beta_1 x + \varepsilon) = E(\beta_0) + E(\beta_1 x) + E(\varepsilon) = \beta_0 + \beta_1 x$$

   For each *x*, mean of *y*'s lie on regression line

2. Based on: **Constant Variance Assumption**

$$Var(y) = Var(\beta_0 + \beta_1 x + \varepsilon) = Var(\varepsilon) = \sigma^2$$

   Regardless of x-value, variance of y's constant

3. Based on: **Independence Assumption**

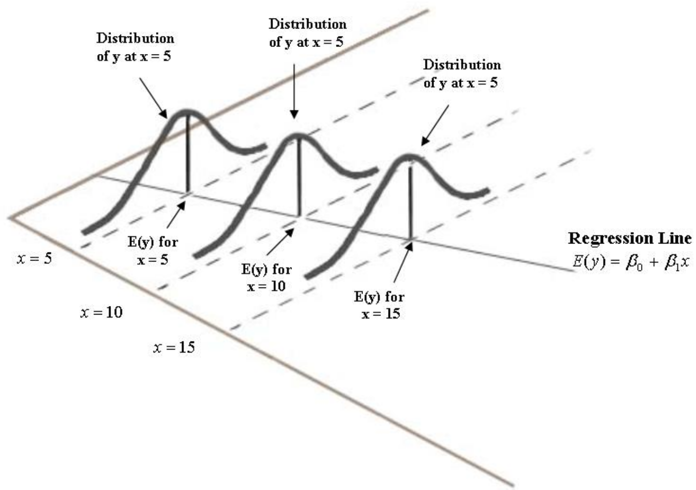
   For any *x*, values of *y* are independent

4. Based on: **Normality Assumption**
   - *y* normally distributed random variable

Susana Nascimento (snt@fct.unl.pt)

# The Regression Model *(cont'd)*

- Normality of $y_i$, with mean $\beta_0 + \beta_1 x$ and constant variance $\sigma^2$ shown
- Observed *y*-values corresponding to predictor values x = 5, 10, and 15 are samples from normal distribution with mean $\beta_0 + \beta_1 x$
- Normal curves have exactly same shape




- For each *x* value, the corresponding *y* are normally distributed.

Susana Nascimento (snt@fct.unl.pt)

# Validating Regression Assumptions

- Validating the regression assumptions **is not important** when inference or model building is *not* performed
  - Regression analysis can be applied in a descriptive manner

- **Regression assumptions *must be validated* when inference or model building is performed**

# Summary

1. The Least Squares Estimates
2. Correlation & Determination Coefficients: properties
3. The Regression Model
4. Inference in Regression
5. Verifying Regression Assumptions

# Inference in Regression

- Suppose the data set is unfamiliar with *x* and *y* values in range – 4.0 to 4.0
- Predicting y with *x* from ERE: $r^2$ = 0.3%
- $r^2$ value indicates linear relationship not useful

- ***Are we sure***?
  - Can a **linear relationship** between *x* and *y* exist when $r^2$ is small?
  - **Inference offers a systematic framework to assess significance of linear association between *x* and *y*.**

Susana Nascimento (snt@fct.unl.pt)

# Inferential Methods

1. *T*-test for relationship between response and predictor
2. Confidence interval for slope ($\beta_1$)
3. Confidence interval for mean of response, given an *x*-value
4. Prediction interval for random response value, given an *x*-value

Susana Nascimento (snt@fct.unl.pt)

# Inference in Regression *(cont'd)*

- Regression model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

$\beta_1$ model parameter, whose *true* value unknown

- What value of $\beta_1$ indicates **non-existent** linear relationship between *x* and *y*?

When $\beta_1 = 0$

$$y = \beta_0 + \varepsilon$$

- Linear relationship
  - exists when $\beta_1 \neq 0$; no longer exists $\beta_1 = 0$
- **Inference is based on this key idea**

Susana Nascimento (snt@fct.unl.pt)

# *T*-test for Relationship *x* and *y*

- Least squares estimate of slope $b_1$ is a *statistic*
- Sampling distribution of $b_1$ has mean = $\beta_1$ , and standard error $\sigma_{b_1}$:

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{\sum x^2 - \left(\sum x\right)^2 / n}}$$

- Regression inference about $\beta_1$ based on sampling distribution of $b_1$
- $s_{b_1}$ is a point estimate of $\sigma_{b_1}$, with $s$ = standard error of the estimate

$$s_{b_1} = \frac{s}{\sqrt{\sum x^2 - \left(\sum x\right)^2 / n}}$$

Susana Nascimento (snt@fct.unl.pt)

# *T*-test for Relationship bt *x* and *y* *(cont'd)*

- $s_{b1}$ interpreted to measure variability of slope
  - Large $s_{b1}$ values indicate estimate of slope $b_1$ *unstable*
  - Small $s_{b1}$ values indicate estimate of slope $b_1$ *precise*

- *t*-test based on *t*-distribution with *n* - 2 degrees of freedom

$$t = \frac{\left(b_1 - \beta_1\right)}{s_{b_1}}$$

- **null hypothesis true**
  - $t = b_1 / s_{b1}$ follows *t*-distribution with *n* - 2 degrees of freedom

Susana Nascimento (snt@fct.unl.pt)

# *T*-test for Relationship bt *x* and *y*

- **Example**: applying *t*-test to regression results of nutritional rating on sugar content

- Minitab results

```
The regression equation is
Rating = 59.4 - 2.42 Sugars
Predictor      Coef   SE Coef       T      P
Constant     59.444     1.951   30.47  0.000
Sugars      -2.4193    0.2376  -10.18  0.000
S = 9.16160   R-Sq = 58.0%   R-Sq(adj) = 57.5%
Analysis of Variance
Source          DF        SS       MS       F      P
Regression       1    8701.7   8701.7  103.67  0.000
Residual Error  75    6295.1     83.9
Total           76   14996.8
```

Susana Nascimento (snt@fct.unl.pt)

# *T*-test for Relationship bt *x* and *y*

- $b_1$ = -2.4193 (under "Coef")
- $s_{b_1}$ = 0.2376 (under "SE Coef")
- T-statistic value, $t = b_1/s_{b_1}$ = -2.4193/0.2376 = -10.18 found under "T"
- *p*-value for t-statistic found under "*p*" represented by:

$$p - value = P(|t| > t_{obs}) = P(|t| > -10.18) \approx 0.000,$$

$t_{obs}$ observed value of *t*-statistic from regression

- Actual *p*-value less than 0.000

Susana Nascimento (snt@fct.unl.pt)

# *T*-test for Relationship bt *x* and *y*

- $H_0$: asserts $\beta_1 = 0$ (no linear relationship exists)
- $H_a$: asserts $\beta_1 \neq 0$ (linear relationship exists)

- *t*-test rejects $H_0$ when *p*-value small
- Routinely, 0.05 used as rejection threshold
- Because *p*-value ~ 0.000, $H_0$ rejected

- ***Indicates a linear relationship exists between nutritional rating and sugar content***

Susana Nascimento (snt@fct.unl.pt)

# Confidence Interval for Slope of Regression Line

- Can estimate slope of regression line $\beta_1$, using *confidence interval*
- The *t*-interval is based on sampling distribution for $b_1$
- 100(1-$\alpha$)% confidence interval for slope $\beta_1$

$$b_1 \pm (t_{n-2})(s_{b_1})$$

$t_{n-2}$ based on 2 degrees of freedom

- **One get's 100(1- $\alpha$)% confident true slope $\beta_1$ lies within interval**

Susana Nascimento (snt@fct.unl.pt)

# Confidence Interval for Slope of Regression Line

- **Example (*Nutritional Data*)** Construct 95% confidence interval for true slope $\beta_1$

Regression results of nutritional rating on sugar content

$b_1 = -2.4193$, $S_{b1} = 0.2376$

*T*-critical value for 95% confidence, $n - 2 = 75$ deg. freedom

$t_{75,95\%} = 2.0$

Confidence interval equals

$-2.4193 \pm 2.0 \times 0.2376 = (-2.8945, -1.9441)$

Susana Nascimento (snt@fct.unl.pt)

# Confidence Interval for Slope of Regression Line *(cont'd)*

Confidence interval

–2.4193 ± 2.0 *0.2376= (–2.8945, –1.9441)

- 95% confident true slope of regression line lies between

  –2.8945 and –1.9441

- For each additional gram of sugar, nutritional rating ***decreases between*** 1.94 and 2.89 points

- Since $\beta_1 = 0$ *not contained* in (–2.8945, –1.9441)

  95% confident of significance in **linear relationship** between *nutritional rating* and *sugar content*

Susana Nascimento (snt@fct.unl.pt)

# Confidence Interval for Mean Value of *y* given *x*

- The point estimate obtained using ERE ***do not* provide** probability statement regarding their accuracy

- Two intervals provide probability statement for estimate

(1) **Confidence interval** for mean value of *y*, given *x*

(2) **Prediction interval** for value of randomly chosen *y* given *x*

Susana Nascimento (snt@fct.unl.pt)

# Confidence Interval for Mean Value of y Given x *(cont'd)*

- Confidence Interval for Mean Value of *y* given *x*

$$\hat{y}_p \pm t_{n-2}\,(s)\sqrt{\frac{1}{n}+\frac{\left(x_p-\bar{x}\right)^2}{\sum\left(x_i-\bar{x}\right)^2}}$$

- $x_p$ given value of *x*, for which prediction being made
- $y_p$ point estimate of y, for given value of *x*
- $t_{n-2}$ multiplier associated with sample size and confidence level
- $s$ standard error of estimate

# Prediction Interval for Randomly Chosen Value of *y* given *x* *(cont'd)*

- **Example**: Probably not too unusual for randomly chosen student's score to exceed 98% on exam
  - However, class mean on exam extremely likely not to exceed 98%
  - Variability associated with variable's mean smaller than variability of individual observation for same variable

- **Example**: random variable $x$ has standard deviation $\sigma$, while sample mean $\bar{x}$ has standard deviation $\sigma/n$

**Predicting class average for exam easier than predicting exam score for randomly chosen student**

Susana Nascimento (snt@fct.unl.pt)

# Prediction Interval for Randomly Chosen Value of *y* Given *x* *(cont'd)*

- Often, analysts interested in predicting individual value, rather than mean of all values, for given $x$
  - **Example**: predict credit score for *single* applicant, rather than mean credit score for all applicants, given $x$

- **Prediction Interval for Randomly Chosen Value of $y$, given $x$:**

$$\hat{y}_p \pm t_{n-2}\,(s)\sqrt{1+\frac{1}{n}+\frac{\left(x_p-\bar{x}\right)^2}{\sum\left(x_i-\bar{x}\right)^2}}$$

$s$ is the standard error of the estimate

- Obs.: similar to confidence interval for mean value of $y$, given $x$

Susana Nascimento (snt@fct.unl.pt)

# Prediction Interval for Randomly Chosen Value of *y* Given *x* *(cont'd)*

$$\hat{y}_p \pm t_{n-2}(s)\sqrt{1+\frac{1}{n}+\frac{(x_p-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$$

- Formula reflects greater variability associated with estimating single value of $y$
  (includes additional constant "1+" under square root)

- Prediction interval ***always* wider** than confidence interval, given same confidence level

Susana Nascimento (snt@fct.unl.pt)

# Prediction Interval for Randomly Chosen Value of y Given x *(cont'd)*

**Example:** ERE for Orienteering data set (10 hikers)

- Estimated distance traveled for hiker $x = 5$ hours: $\hat{y} = 6 + 2(5) = 16$ km

**How accurate is point estimate = 16?**

- Point estimate = 16 ***does not provide*** probability statement regarding its accuracy
- *For each x-value, regression model assumes observed y-values are samples from normal population with mean located on regression line*
    - regression model assumes existence of normal population of hikers with $x = 5$
    - Of all hikers in *this* population, 95% will travel within bounded interval, where distance = 16 km

Susana Nascimento (snt@fct.unl.pt)

# Prediction Interval for Randomly Chosen Value of *y* Given *x* *(cont'd)*

- Calculate 95% confidence interval for mean distance traveled, for all hikers with $x = 5$
- **Confidence Interval (for the *mean*)**
  - $x_p = 5$, x-bar = 5, $y_p = 16$, $t_{8,95\%} = 2.306$
  - $s = 1.22474$ (from regression results), and $n = 10$

$$\hat{y}_p \pm t_{n-2}(s)\sqrt{\frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{\sum\left(x_i - \bar{x}\right)^2}}$$

$$= 16 \pm (2.306)\ (1.22474)\sqrt{\frac{1}{10} + \frac{(5-5)^2}{54}}$$

$$= 16 \pm 0.893$$

$$= (15.107, 16.893)$$

  - 95% confident that **mean distance** traveled by all hikers traveling 5 hours, lies between 15.11 and 16.89 km

Susana Nascimento (snt@fct.unl.pt)

# Prediction Interval for Randomly Chosen Value of y Given x *(cont'd)*

- Estimate distance traveled by *randomly selected hiker* who hiked $x = 5$ hours?
- **Prediction Interval (for single hiker)**

$$\hat{y}_p \pm t_{n-2}(s)\sqrt{1+\frac{1}{n}+\frac{\left(x_p-\bar{x}\right)^2}{\sum\left(x_i-\bar{x}\right)^2}}$$

$$= 16 \pm (2.306)\ (1.22474)\sqrt{1+\frac{1}{10}+\frac{\left(5-5\right)^2}{54}}$$

$$= 16 \pm 2.962$$

$$= (13.038,\ 18.962)$$

  - 95% confident that **distance traveled by randomly select hiker** traveling 5 hours, lies between 13.04 and 18.96 *km*

Susana Nascimento (snt@fct.unl.pt)

# Prediction Interval for Randomly Chosen Value of *y* Given *x* *(cont'd)*

- Prediction interval is *wider* than confidence interval
- Estimating **single response** is *more difficult* than estimating **mean response**
- **In general, interpretation of prediction interval more useful to the analyst**

Susana Nascimento (snt@fct.unl.pt)

# Confidence Interval for Correlation Coefficient $\rho$

- Assume *x* and *y* normally distributed

**THE $100(1-\alpha)\%$ CONFIDENCE INTERVAL FOR THE POPULATION CORRELATION COEFFICIENT $\rho$**

We can be $100(1-\alpha)\%$ confident that the population correlation coefficient $\rho$ lies between:

$$r \pm t_{\alpha/2,\, n-2} \cdot \sqrt{\frac{1-r^2}{n-2}}$$

where $t_{\alpha/2,n-2}$ is based on $n-2$ degrees of freedom.

Susana Nascimento (snt@fct.unl.pt)

# Example

- Regression of *ln rating* on *carbo-hydrates*

$$r = +\sqrt{r^2} = +\sqrt{0.025} = 0.1581$$

$$t_{\alpha/2,\, n-2} = t_{0.025,\, 74} = 1.99$$

$$r \pm t_{\alpha/2,\, n-2} \cdot \sqrt{\frac{1-r^2}{n-2}}$$
$$= 0.1581 \pm 1.99 \cdot \sqrt{\frac{1-0.025}{74}}$$
$$= (-0.0703,\ 0.3865)$$

*ln rating* and *carbohydrates* are not linearly correlated

Susana Nascimento (snt@fct.unl.pt)

# Using a Confidence Interval to Assess Correlation

- If both endpoints of the confidence interval are **positive**, then

  ***x* and *y* are positively correlated**, with confidence level $100(1-\alpha)\%$

- If both endpoints of the confidence interval are **negative**, then

  ***x* and *y* are negatively correlated**, with confidence level $100(1-\alpha)\%$

- If one endpoint is negative and one endpoint is positive, then

  ***x* and *y* are not linearly correlated**, with confidence level $100(1-\alpha)\%$

Susana Nascimento (snt@fct.unl.pt)

# Verifying Regression Assumptions

- Four inferential methods described *require* adherence to regression assumptions

- **Two graphical methods used to verify assumptions**

  (1) Normal probability plot of residuals

  (2) Plot of standardized residuals against predicted values

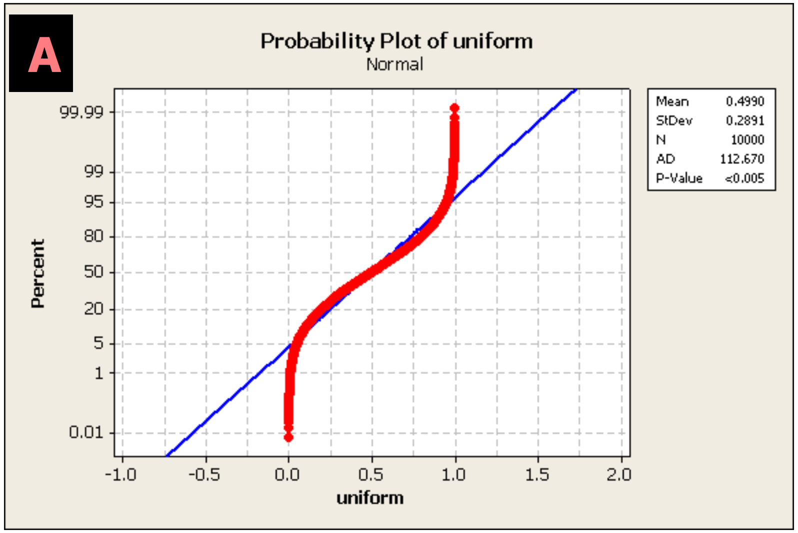
- **Method 1: Normal Probability Plot**

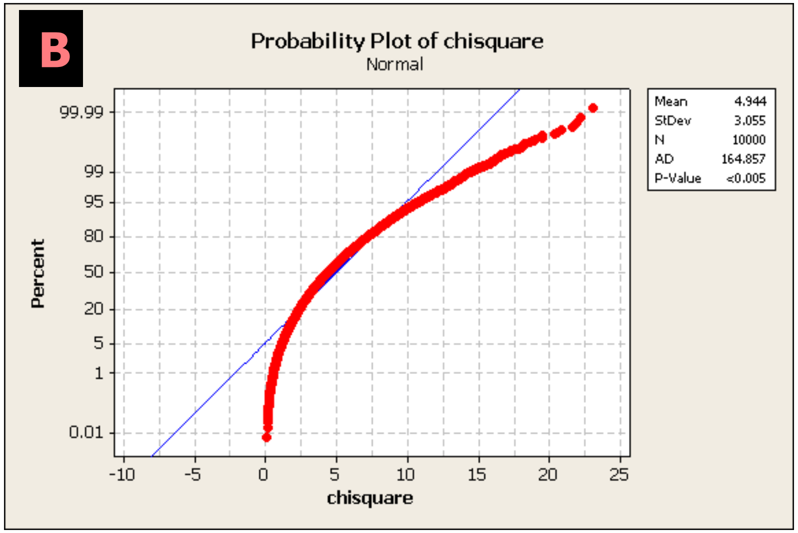
  Quantile-Quantile plot of quantiles of particular distribution against quantiles of standard normal distribution

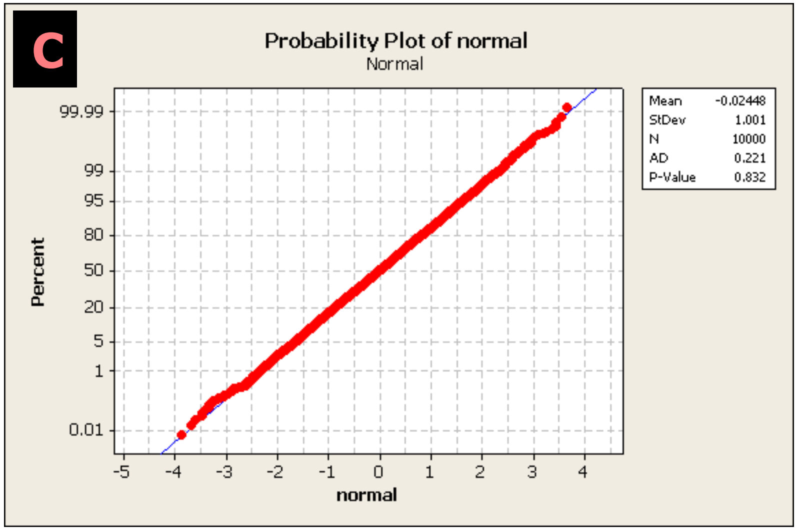
Susana Nascimento (snt@fct.unl.pt)

# Verifying Regression Assumptions I

- Three examples show normal probability plots for:
  - (A) Uniform(0,1)
  - (B) Chi-square(5)
  - (C) Normal(0,1)

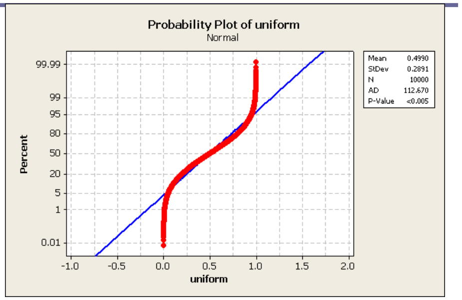


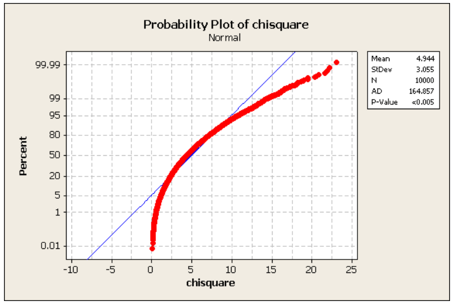



Susana Nascimento (snt@fct.unl.pt)

# Verifying Regression Assumptions I



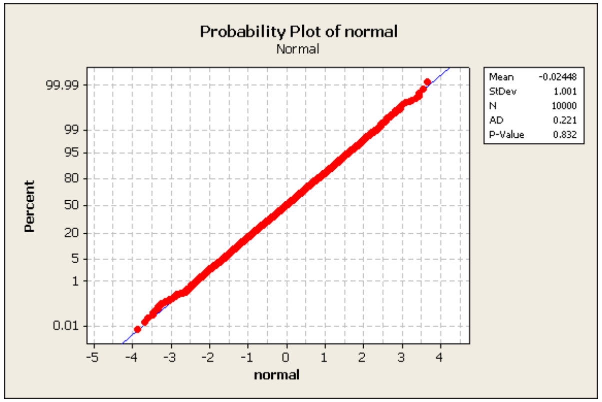

- (A) Uniform(0,1)
  - Clear pattern (reverse S curve) exists indicating systematic deviation from normality
  - Uniform distribution is rectangular-shaped distribution with heavy tails




- (B) Chi-Square(5)
  - Again, clear pattern indicates systematic deviation from normality
  - Chi-Square(5) distribution is right-skewed
  - Plot appearance typical of right-skewed distributions

Susana Nascimento (snt@fct.unl.pt)

# Verifying Regression Assumptions I




- (C) Normal(0,1)
  - All points line up on straight line indicating normality
  - Expected behavior for data drawn from Normal distribution
  - However, sampling error and noise found in real-world data make decisions about normality less certain
  - Note, each probability plot generated by Minitab shows AD statistic and corresponding *p*-value

Susana Nascimento (snt@fct.unl.pt)

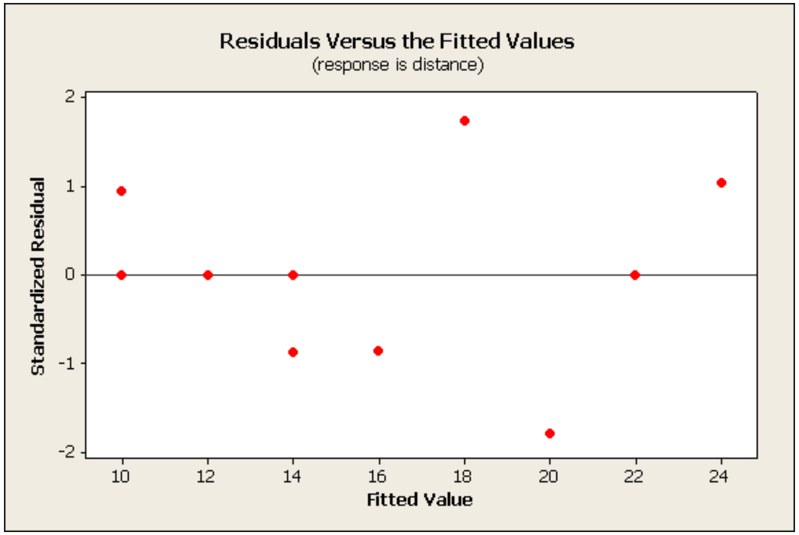
# Verifying Regression Assumptions I

- Quantile of distribution $x_p$ measures $p$% of distribution less than or equal to $x_p$
- Determines whether specified distribution deviates from normality
- Observed values from distribution compared against same number observations expected from normal distribution
- Where normal, bulk of points should lie on straight line
- **Otherwise, systematic deviations from linearity denote non-normality**

Susana Nascimento (snt@fct.unl.pt)

# Verifying Regression Assumptions II

- **Method 2: Plot Standardized Residuals Against Fits (Predicted Values)**
  - (**qq-plot in MatLab**)
- Example shows regression of distance *vs* time for Orienteering data set
- Regression line (ERE) shown as horizontal blue line
- Discernable patterns in residuals *vs*. fits indicates regression assumptions violated

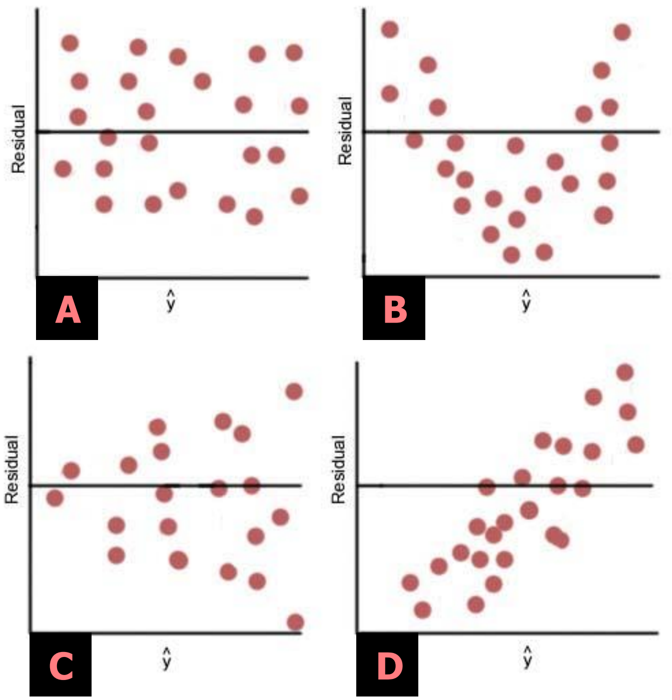



- Too few data points exist to make determination

Susana Nascimento (snt@fct.unl.pt)

# Verifying Regression Assumptions II

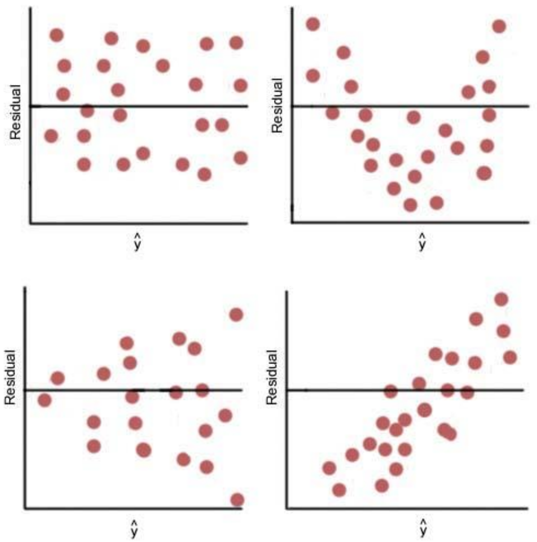
- Four commonly found patterns in residual-fit plots shown
- Plot (A): **“healthy” plot where no discernible patterns exist**
- Data points form overall rectangular shape
- Regression assumptions remain intact



Susana Nascimento (snt@fct.unl.pt)

# Verifying Regression Assumptions II

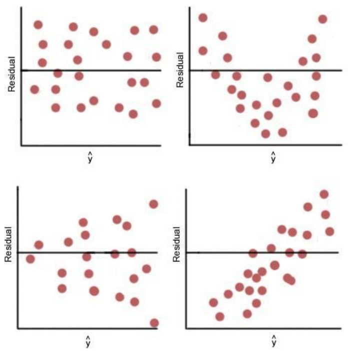
- **Plot (B): exhibits curvature, which violates independence assumption**




- Why (plot B)?
  - Residuals, which estimate errors, assumed to exhibit independence
  - Residuals form curved pattern
  - Therefore, for given residual, we may predict where neighboring residuals fall
  - If residuals independent, no prediction is possible

Susana Nascimento (snt@fct.unl.pt)

# Verifying Regression Assumptions II




- Plot (C): **displays "*funnel*" pattern, which violates constant variance assumption**
- Why (plot C)?
  - Residual variance lower for smaller values of *x*, where variance increases as *x*-values increase
  - Constant variance assumption violated
- Plot (D): **shows pattern increasing from left to right, which violates zero-mean assumption**
- Why (plot D)?
  - Zero-mean states mean of error term zero, regardless of x-value
  - Plot shows small x-values have mean *less* than 0
  - Large x-values have mean *greater* than 0
  - Therefore, zero-mean assumption violated

Susana Nascimento (snt@fct.unl.pt)

# Verifying Regression Assumptions

- **Diagnostic Tests** ***(just informative)***
  - Several diagnostic hypothesis tests exist to validate regression assumptions
  - Anderson-Darling test validates residual fit to normal distribution
  - Bartlett’s test or Levene’s test indicates whether constant variance assumption violated
  - Durban-Watson test or runs test assesses whether independence assumption violated

Susana Nascimento (snt@fct.unl.pt)

# Regression Assumptions: Outlook

- Suppose normality plot shows no systematic deviation from linearity, and
- Residuals-fits plot shows no discernible patterns
- Therefore, graphical evidence shows no evidence regression assumptions violated

**What if graphical tests indicate regression assumption(s) violated?**

For example, constant variance assumption violated

- **Transforming response using *ln*(y) may help**

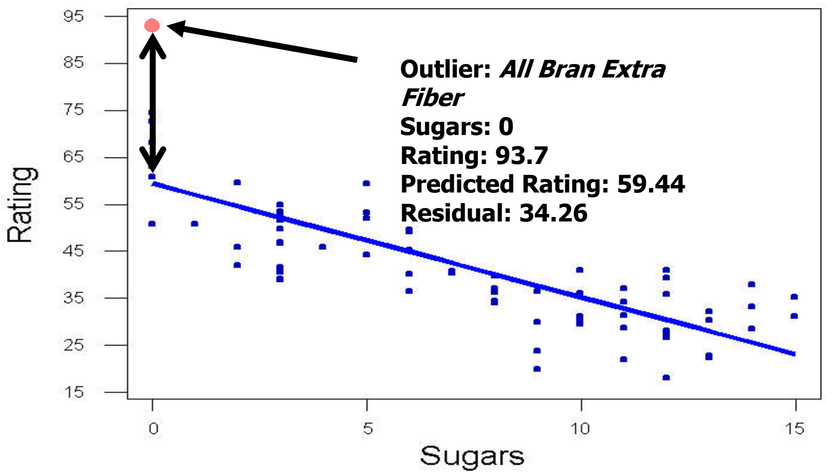
# Supplementary

## Outliers, High Leverage Points, and Influential Observations

Susana Nascimento (snt@fct.unl.pt)

# Outliers, High Leverage Points, and Influential Observations

- ***Outliers*** are observations with very large standardized residual, in absolute value
- Recall *Cereals* scatter plot of *rating* against *sugars*




  - Vertical distance from *All Bran Extra Fiber* to ERE has largest residual in data set
  - residual = (y - ŷ) = 93.7 – 59.44 = 34.26

Susana Nascimento (snt@fct.unl.pt)

# Outliers, High Leverage Points, and Influential Observations *(cont'd)*

- Residuals may have different variances =>

  **standardize residuals to same the scale**
- Let $s_{i,resid}$ denote standard error of $i_{th}$ residual, with leverage $h_i$
- Standardized Residual: $s_{i,resid} = s\sqrt{1-h_i}$

$$residual_{\text{i,standardized}} = \frac{y_i - \hat{y}_i}{s_{i,resid}}$$

- Generally, observations with standardized residual > 2 are flagged as outliers

Susana Nascimento (snt@fct.unl.pt)

# Outliers, High Leverage Points, and Influential Observations *(cont'd)*

- Minitab detects *All Bran Extra Fiber* with standardized residual = 3.83
- **Residual positive** observed *y*-value **is *higher*** than estimated *y*-value
- **Residual negative** observed y-value **is *lower*** than estimated y-value

- **High Leverage Points** extreme observations in predictor space
- Very large values of *x* variable without reference to *y* variable

Susana Nascimento (snt@fct.unl.pt)

# Outliers, High Leverage Points, and Influential Observations *(cont'd)*

- *Leverage* $h_i$ for $i_{th}$ observation:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}$$

$h_i$ only dependents on $(x_i - \bar{x})^2$

- So, as distance from *x*-value to $\bar{x}$ increases, leverage increases

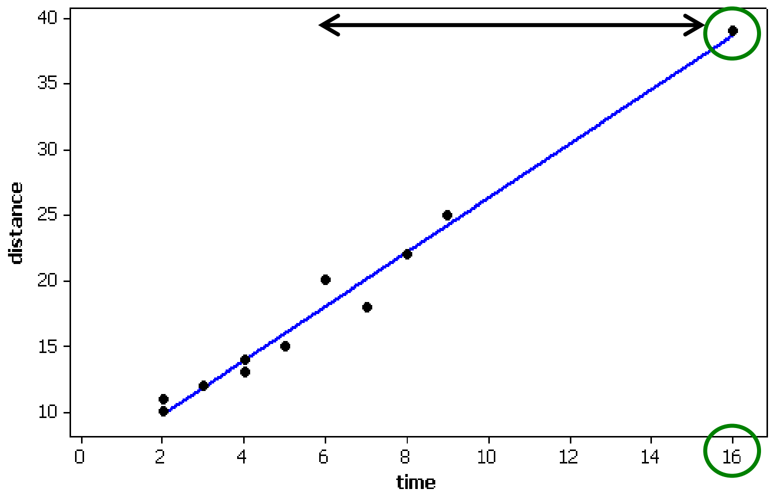
$$1/n \leq Leverage \leq 1.0$$

- **High leverage**
  - When leverage > 2($m$ + 1) or 3($m$ + 1)

Susana Nascimento (snt@fct.unl.pt)

# Outliers, High Leverage Points, and Influential Observations *(cont'd)*

- Suppose 11th hiker from Orienteering data set hiked $x$ = 16 hours




- Identified as high leverage point by extreme number of hours, without reference to distance traveled

- Minitab flags "*unusual observation*" with "*large influence*"

Susana Nascimento (snt@fct.unl.pt)

# Outliers, High Leverage Points, and Influential Observations *(cont'd)*

- *Influential* observation significantly alters regression parameters based on absence/presence in data set
- Outlier ***may or may not*** be influential
- High leverage point ***may or may not*** be influential

- **Example**: suppose 11th hiker traveled 20 km in 5 hours

  Although Minitab flags observation as outlier, is it?

  Including/removing the observation results in ER'Es with:

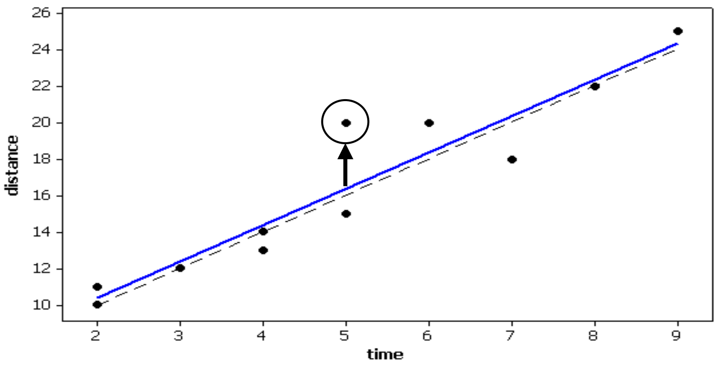
  $b_0 = 6.00$, $b_1 = 2.00$

  $b_0 = 6.36$, $b_1 = 2.00$

Susana Nascimento (snt@fct.unl.pt)

# Outliers, High Leverage Points, and Influential Observations *(cont'd)*

- Diagram shows mild effect on ERE by including/not including outlier




- In fact, *x* = 5 same as *x*-bar leading to leverage:

$$h_{(5,20)} = \frac{1}{11} + \frac{(5-5)^2}{54} = 0.0909$$

- So, point has low leverage and is not influential

Susana Nascimento (snt@fct.unl.pt)

# Outliers, High Leverage Points, and Influential Observations *(cont'd)*

- Standard Error of the Residual and Standardized Residual are derived:

$$s_{(5,20),\,resid} = 1.71741\sqrt{1-0.0909} = 1.6375$$

$$residual_{(5,20),\text{standardized}} = \frac{y_i - \hat{y}_i}{s_{(5,20),\,resid}} = \frac{20-16.364}{1.6375} = 2.22,$$

- Cook's Distance measures an observation's level of influence
- Considers both size of residual and leverage for observation

Susana Nascimento (snt@fct.unl.pt)

# Outliers, High Leverage Points, and Influential Observations *(cont'd)*

- Cook's Distance:

$$D_i = \frac{(y_i - \hat{y}_i)^2}{(m+1)s^2}\left[\frac{h_i}{(1-h_i)^2}\right]$$

  - $(y_i - \hat{y}_i)$ $i_{th}$ residual
  - s standard error of the estimate
  - $h_i$ leverage of $i_{th}$ observation
  - $m$ number of predictors
- Combines elements representing outlier and leverage
- Cook's Distance for 11th (5, 20) hiker:

$$D_i = \frac{(20-16.364)^2}{(1+1)1.71741^2}\left[\frac{0.0909}{(1-0.0909)^2}\right] = 0.2465$$

Susana Nascimento (snt@fct.unl.pt)

# Outliers, High Leverage Points, and Influential Observations *(cont'd)*

- In general, influential observations have Cook's Distance > 1.0
- Cook's Distance also compared against $F_{m,n-m}$ distribution
- Measure greater than median percentile of $F_{m,n-m}$ influential
- **Example**: 11th hiker (5, 20) with Cook's Distance = 0.2465 not influential, lies within 37th percentile of $F_{1,10}$
- **Example**: hard-core hiker (16, 39) has high leverage = $h_i$ = 0.7007, and standardized residual = 0.46801
- However, Cook's Distance = 0.2564 shows it's not influential

Susana Nascimento (snt@fct.unl.pt)

# Outliers, High Leverage Points, and Influential Observations *(cont'd)*

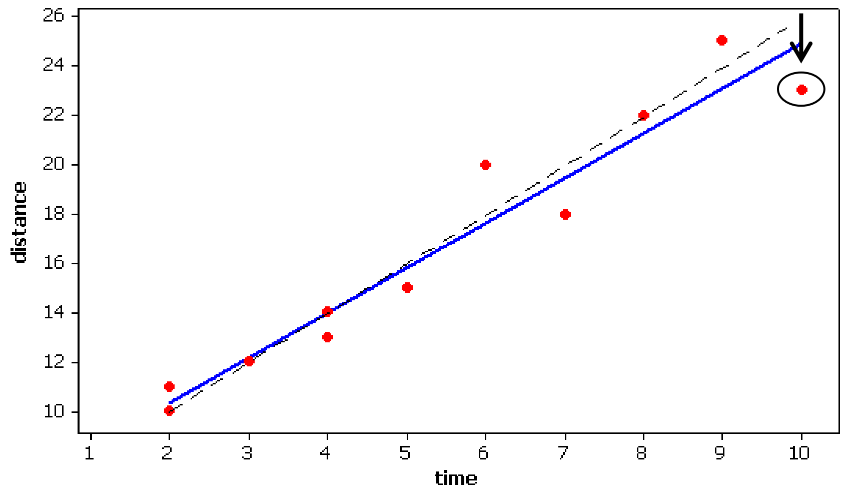
- So, outlier with **low influence *or* high leverage** point with **small residual** is not necessarily influential

- **What about an observation with moderately high leverage and residual?**
- **Example**: 11th hiker (10, 23) has leverage $h_i = 0.36019$, and standardized residual = -1.70831
  - Observation influential with Cook's Distance = 0.821457, lies within 62nd percentile of $F_{1,10}$
- Influence results from moderately high leverage ***combined* with** moderately high residual

Susana Nascimento (snt@fct.unl.pt)

# Outliers, High Leverage Points, and Influential Observations *(cont'd)*

- Influence of 11th hiker "pulls down" regression line where slope $b_1$ decreases from 2.00 to 1.82



Susana Nascimento (snt@fct.unl.pt)

# Homework

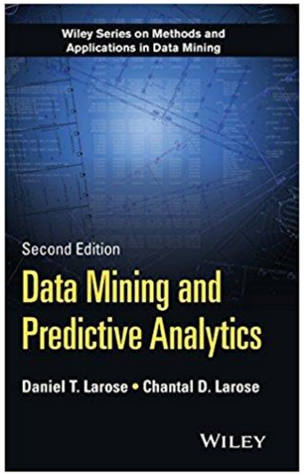
Do the tutorials available on CLIP

- *Exploratory Data Analysis*

- *Linear Regression*
  - Case Study: Baseball data set

Susana Nascimento (snt@fct.unl.pt)

# References




Larose, T. & Larose, C. (2015). *Data Mining and Predictive Analytics*, Wiley Series on Methods and Applications in Data Mining, Wiley (2nd edition), **Chapter 8**