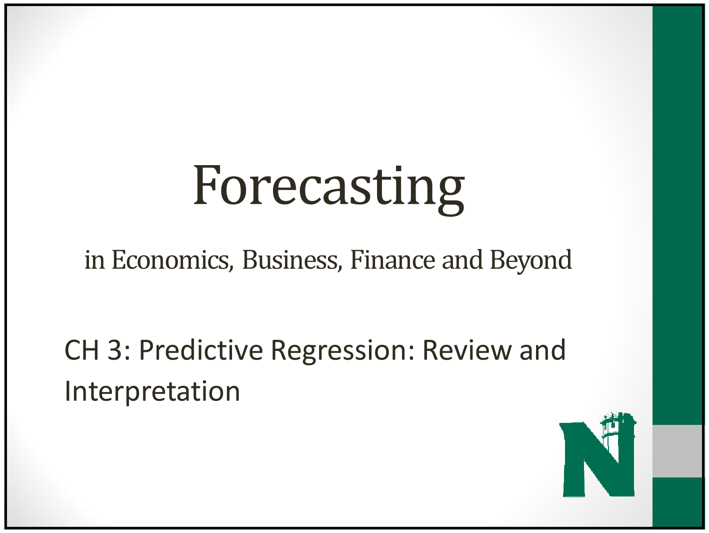
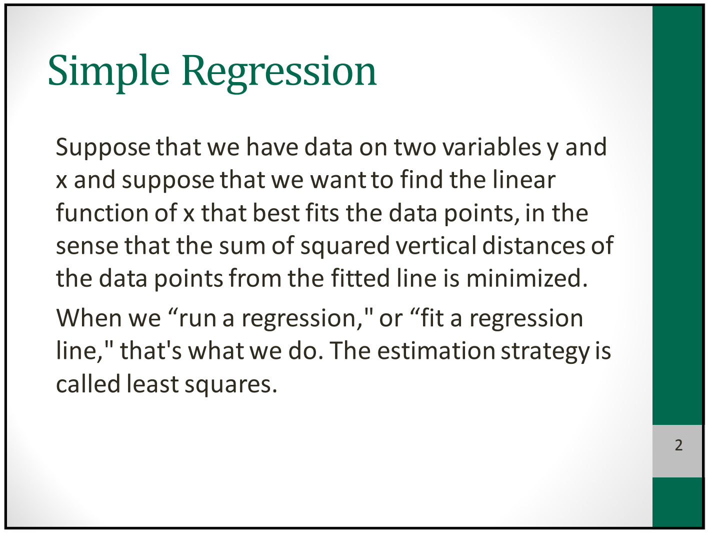


Forecasting

in Economics, Business, Finance and Beyond

CH 3: Predictive Regression: Review and Interpretation

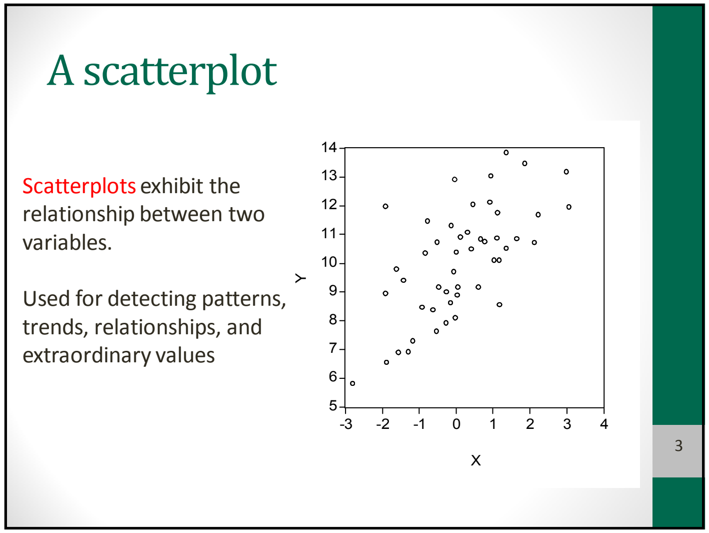


Simple Regression

Suppose that we have data on two variables y and x and suppose that we want to find the linear function of x that best fits the data points, in the sense that the sum of squared vertical distances of the data points from the fitted line is minimized.

When we "run a regression," or "fit a regression line," that's what we do. The estimation strategy is called least squares.

2

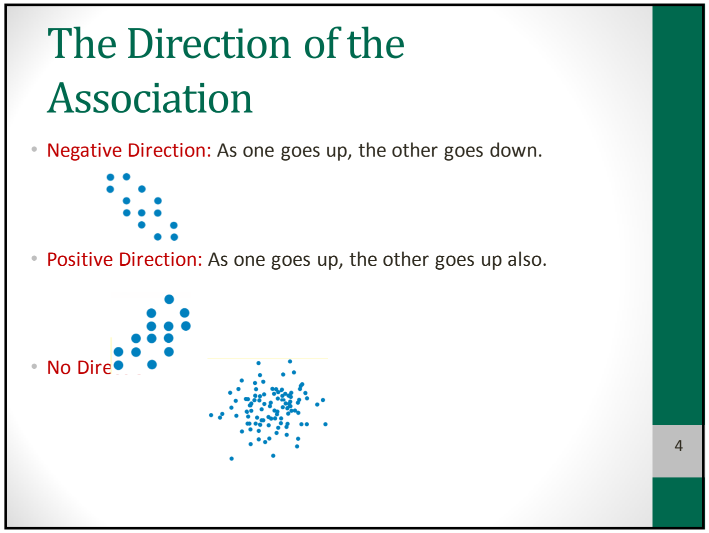


A scatterplot

Scatterplots exhibit the relationship between two variables.

Used for detecting patterns, trends, relationships, and extraordinary values

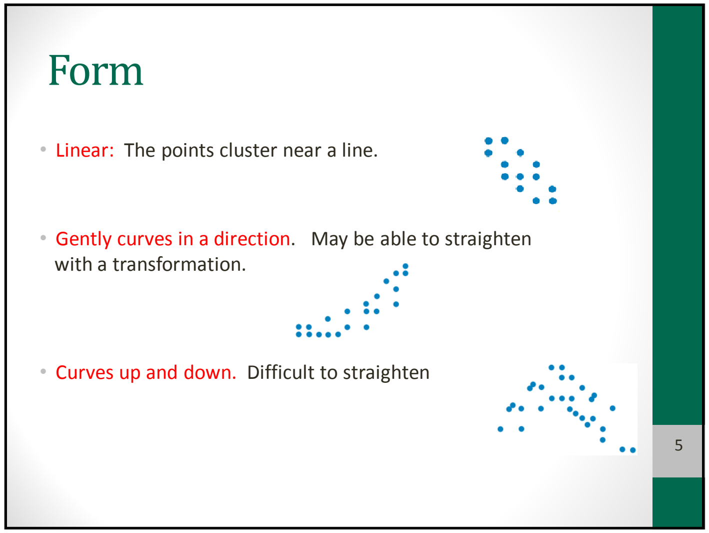
3



The Direction of the Association

- Negative Direction: As one goes up, the other goes down.
- Positive Direction: As one goes up, the other goes up also.
- No Dire

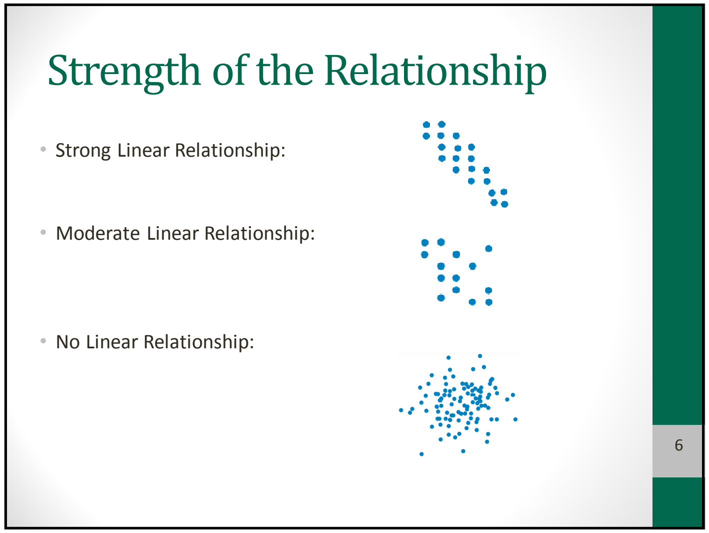
4



Form

- Linear: The points cluster near a line.
- Gently curves in a direction. May be able to straighten with a transformation.
- Curves up and down. Difficult to straighten

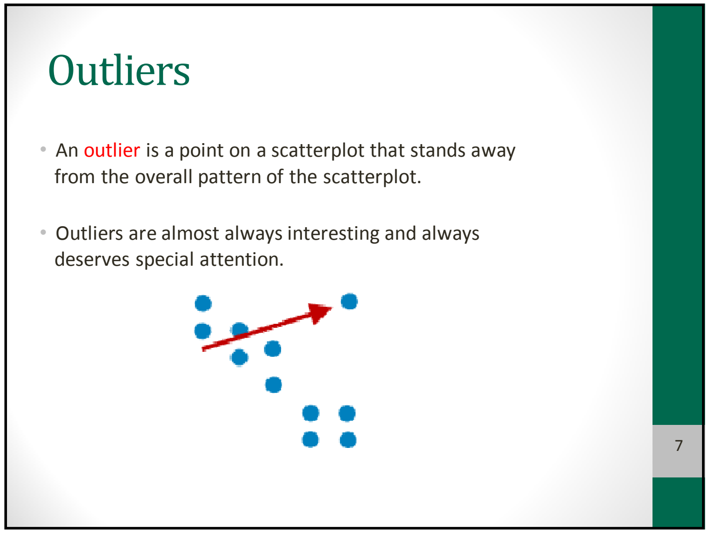
5



Strength of the Relationship

- Strong Linear Relationship:
- Moderate Linear Relationship:
- No Linear Relationship:

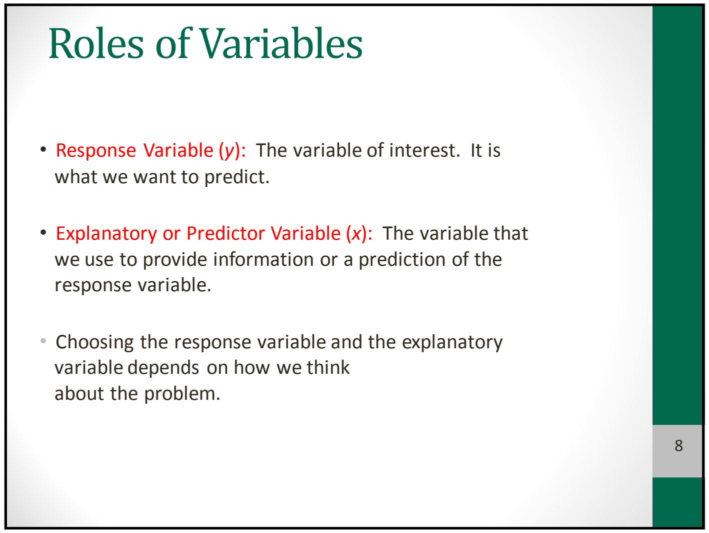
6



Outliers

- An outlier is a point on a scatterplot that stands away from the overall pattern of the scatterplot.
- Outliers are almost always interesting and always deserves special attention.

7



Roles of Variables

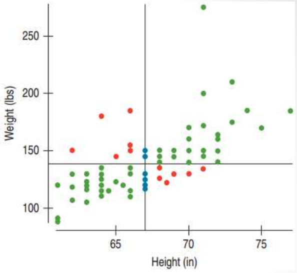
- Response Variable (*y*): The variable of interest. It is what we want to predict.
- Explanatory or Predictor Variable (*x*): The variable that we use to provide information or a prediction of the response variable.
- Choosing the response variable and the explanatory variable depends on how we think about the problem.

8

Correlation

Define the correlation coefficient by almost the average product of the *z*-scores:

$$r = \frac{\sum z_x z_y}{n-1}$$
$$= \frac{1}{(n-1)s_x s_y}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$$



Green → +

Red → -

Blue → No Association

For the green dots: *z*-scores have the same sign, so multiplying the z-scores produces a positive value.

For the red dots: *z*-scores have opposite signs, so multiplying the *z*-scores produces a negative value.

Assumptions and Conditions for Correlation

- To use *r*, there must be a true underlying linear relationship between the two variables.
- The variables must be quantitative.
- The pattern for the points of the scatterplot must be reasonably straight.
- Outliers can strongly affect the correlation. Look at the scatterplot to make sure that there are no strong outliers.

Properties of Correlation

- $r > 0$ → positive association
- $r < 0$ → negative association
- $-1 \leq r \leq 1$, with $r = -1$ only if the points all lie exactly on a negatively sloped line and $r = 1$ only if the points all lie exactly on a positively sloped line.
- Interchanging *x* and *y* does not change the correlation.
- *r* has no units.

Properties of Correlation

- Changing the units of *x* or *y* does not affect *r*.
  - Measuring in dollars, cents, or Euros will all produce the same correlation.
- Correlation measures the strength of the linear association between the two variables.
- Correlation is sensitive to outliers. An extreme outlier can cause a dramatic change in *r*.
- The adjectives *weak*, *moderate*, and *strong* can describe correlation, but there are no agreed upon boundaries.

Reasons for Correlation

- Causation is a possibility, but more must be done to prove causation.
- The causation could be in reverse (*y* causes *x*)
- A lurking variable may cause both.
  - Number of gray hairs and number of wrinkles are strongly correlated, but dyeing hair black does not undo wrinkles. Age is the lurking variable that causes both to increase.

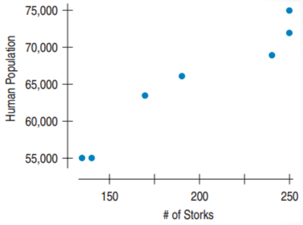
13

Warning: Correlation ≠ Causation

Storks and Babies

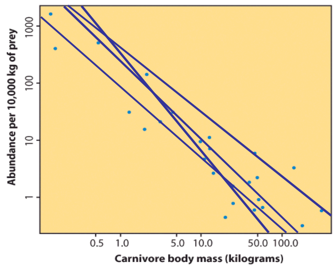
There is a clear positive association between the number of storks and the population.

This does not prove that an increase in storks has caused an increase in babies being born.



Causation is in reverse. Storks nest on house chimneys, so the increased population has increased nesting sites

14



Correlation tells us about *strength* (scatter) and *direction* of the linear relationship between two quantitative variables.

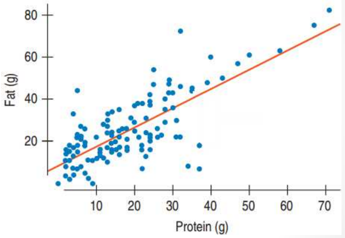
In addition, we would like to have a numerical description of how both variables vary together. For instance, is one variable increasing faster than the other one? And we would like to make predictions based on that numerical description.

But which line best describes our data?

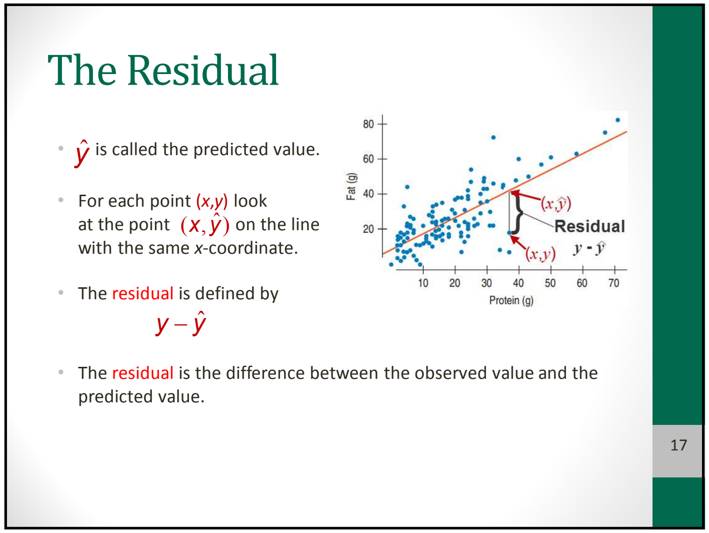
15

The Linear Model

- Fat and Protein at Burger King
  - The correlation is 0.76.
  - This indicates a strong linear fit, but what line?
  - The line should be "closest" to the points.



16

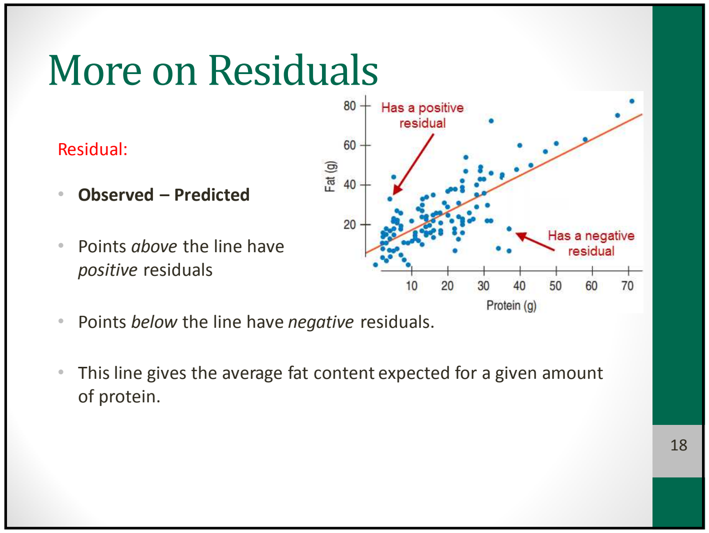


The Residual

- $\hat{y}$ is called the predicted value.
- For each point (x,y) look at the point $(x,\hat{y})$ on the line with the same *x*-coordinate.
- The residual is defined by

  $y - \hat{y}$
- The residual is the difference between the observed value and the predicted value.

17

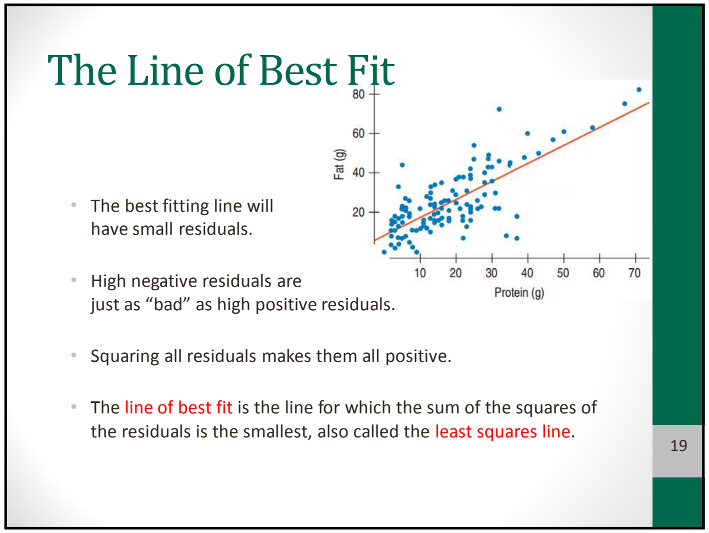


More on Residuals

Residual:

- **Observed – Predicted**
- Points *above* the line have *positive* residuals
- Points *below* the line have *negative* residuals.
- This line gives the average fat content expected for a given amount of protein.

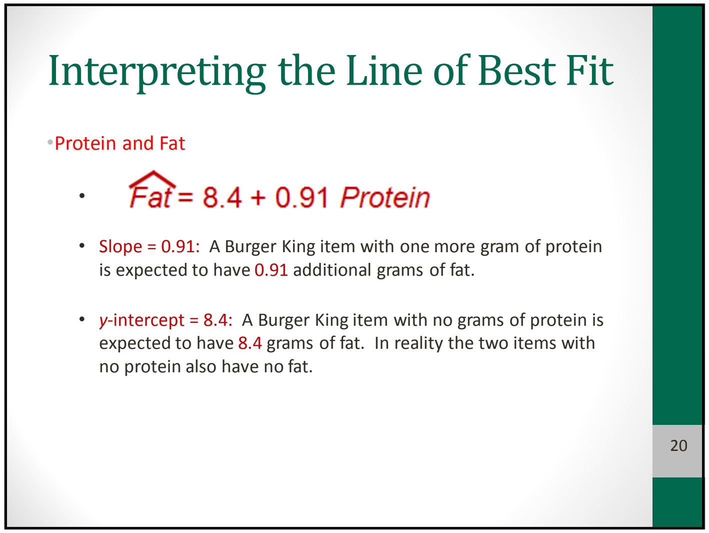
18



The Line of Best Fit

- The best fitting line will have small residuals.
- High negative residuals are just as “bad” as high positive residuals.
- Squaring all residuals makes them all positive.
- The line of best fit is the line for which the sum of the squares of the residuals is the smallest, also called the least squares line.

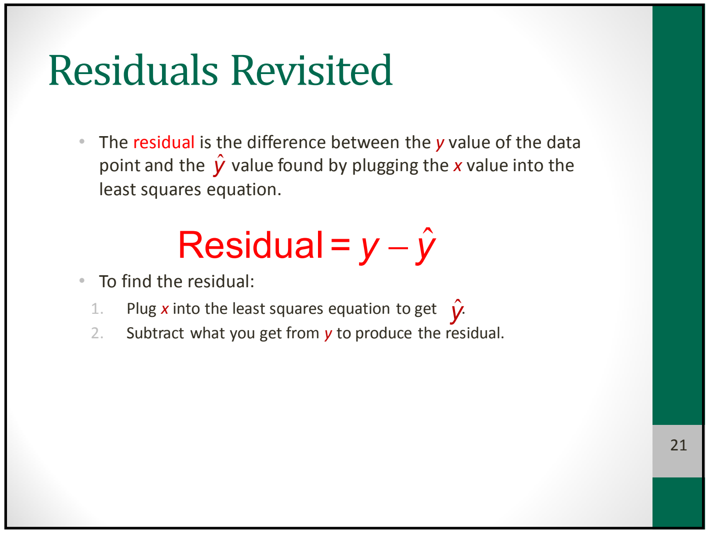
19



Interpreting the Line of Best Fit

- Protein and Fat
  - $\widehat{Fat}$ = 8.4 + 0.91 *Protein*
  - Slope = 0.91: A Burger King item with one more gram of protein is expected to have 0.91 additional grams of fat.
  - *y*-intercept = 8.4: A Burger King item with no grams of protein is expected to have 8.4 grams of fat. In reality the two items with no protein also have no fat.

20



Residuals Revisited

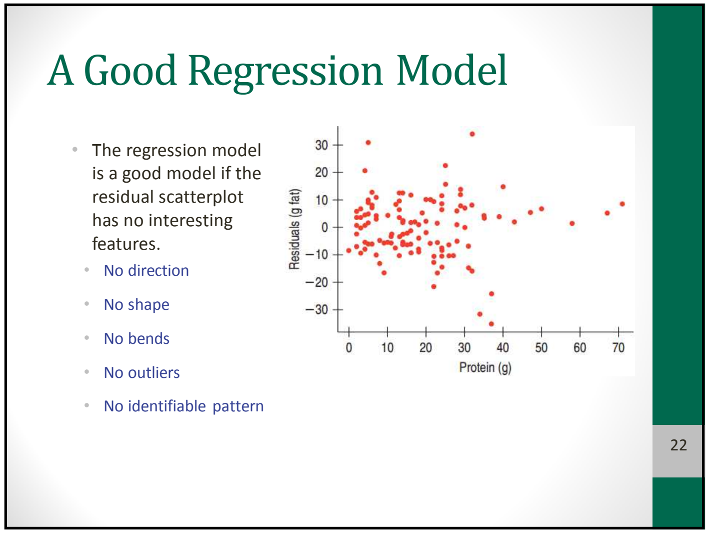
- The residual is the difference between the *y* value of the data point and the $\hat{y}$ value found by plugging the *x* value into the least squares equation.

$$\text{Residual} = y - \hat{y}$$

- To find the residual:

1. Plug *x* into the least squares equation to get $\hat{y}$.
2. Subtract what you get from *y* to produce the residual.

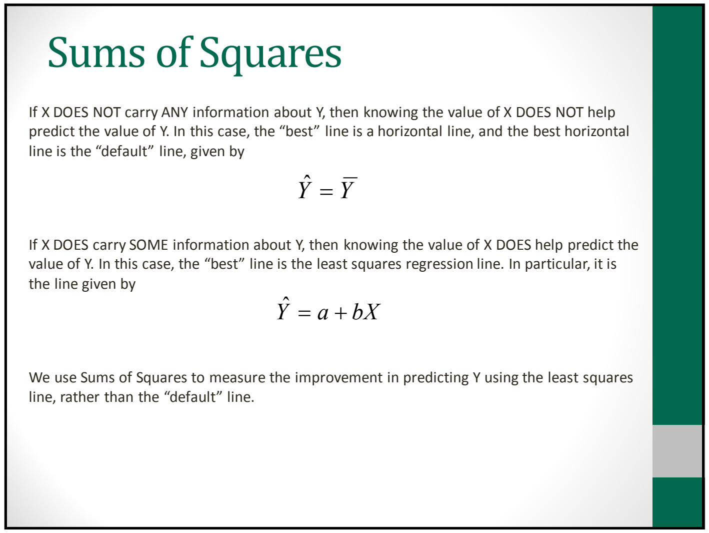
21



A Good Regression Model

- The regression model is a good model if the residual scatterplot has no interesting features.
  - No direction
  - No shape
  - No bends
  - No outliers
  - No identifiable pattern

22



Sums of Squares

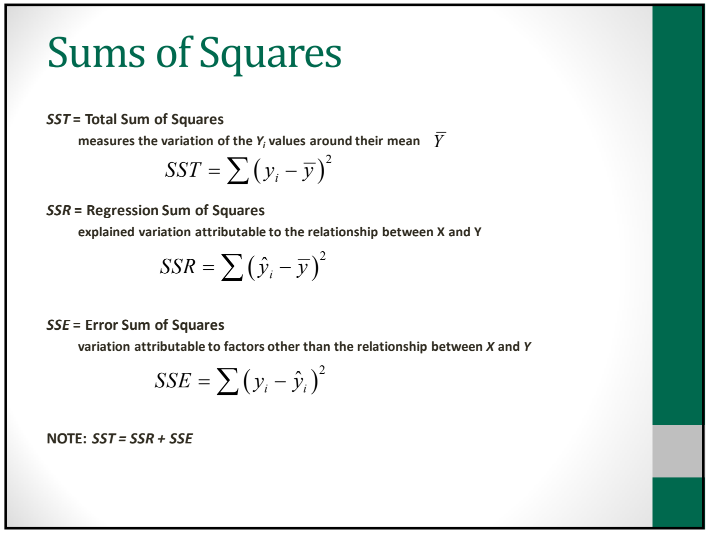
If X DOES NOT carry ANY information about Y, then knowing the value of X DOES NOT help predict the value of Y. In this case, the "best" line is a horizontal line, and the best horizontal line is the "default" line, given by

$$\hat{Y} = \overline{Y}$$

If X DOES carry SOME information about Y, then knowing the value of X DOES help predict the value of Y. In this case, the "best" line is the least squares regression line. In particular, it is the line given by

$$\hat{Y} = a + bX$$

We use Sums of Squares to measure the improvement in predicting Y using the least squares line, rather than the "default" line.



Sums of Squares

SST **= Total Sum of Squares**

measures the variation of the Y_i values around their mean $\overline{Y}$

$$SST = \sum (y_i - \overline{y})^2$$

SSR **= Regression Sum of Squares**

explained variation attributable to the relationship between X and Y

$$SSR = \sum (\hat{y}_i - \overline{y})^2$$

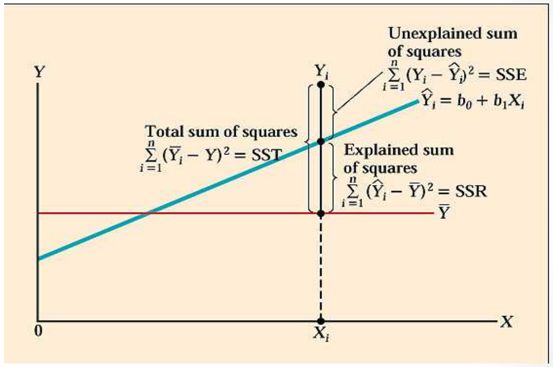
SSE **= Error Sum of Squares**

variation attributable to factors other than the relationship between *X* and *Y*

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

NOTE: *SST = SSR + SSE*

Sums of Squares


Y
Unexplained sum of squares
$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \text{SSE}$
Y_i
$\hat{Y}_i = b_0 + b_1X_i$
Total sum of squares
$\sum_{i=1}^{n}(\overline{Y}_i - Y)^2 = \text{SST}$
Explained sum of squares
$\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = \text{SSR}$
$\overline{Y}$
0
X_i
X

Sums of Squares

Note that the formula for SST is *almost* identical to the formula for the variance of Y:

$$SST = \sum (y_i - \bar{y})^2$$

In other words, SST measures the variability present in the Y values, *without taking the X values into account.*

SSE measures the variability in the residuals; this is the variability that remains after the information about Y that is “carried” by X has been removed. In other words, SSE measures the variability in the Y values that IS NOT “explained” by the linear relationship between Y and X.

Thus, the difference SST – SSE = SSR measures the variability in the Y values that IS “explained” by the linear relationship between Y and X.

Coefficient of determination

Since SST measures the total initial variability in the Y values, and SSE measures the variability of the residuals (the Y values with the linear model subtracted out), the ratio of these values gives the PROPORTION of the variability in the Y values that is “explained by the linear relationship between Y and X. This ratio is called the Coefficient of Determination.

$$\text{Coefficient of Determination} = \frac{\text{Variability "explained" by X}}{\text{Total Variability of Y values}}$$

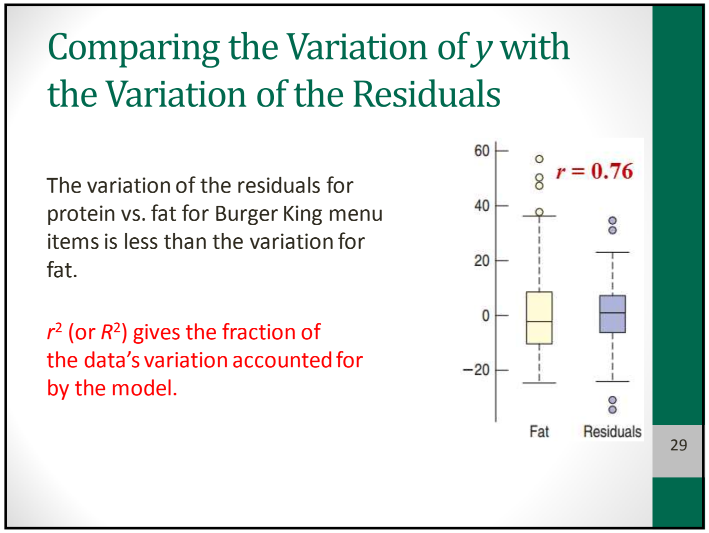
$$= \frac{\text{SSR}}{\text{SST}}$$

$$= \left\{\begin{array}{l}\text{Proportion of Variability in Y values} \\ \text{"explained" by the linear relationship} \\ \text{between Y and X}\end{array}\right.$$

Coefficient of determination, r^2

***r*²**, the coefficient of determination, is the square of the correlation coefficient.

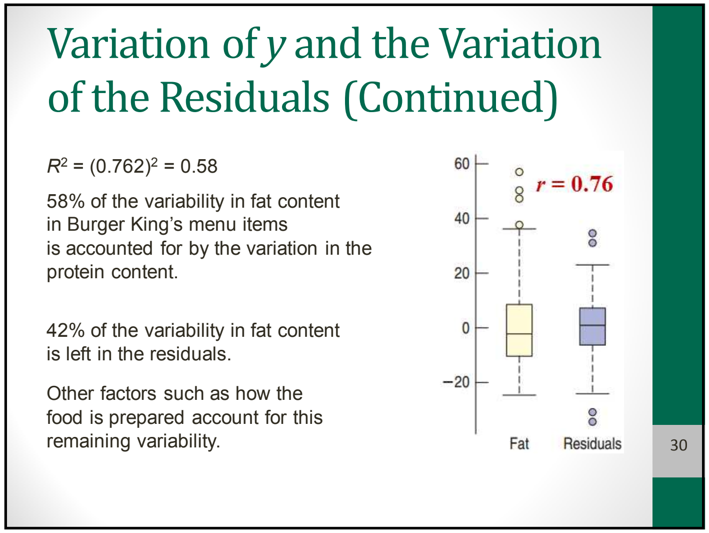
r*²** represents **the fraction of the variance in *y (vertical scatter from the regression line) **that can be explained by changes in *x***.



Comparing the Variation of *y* with the Variation of the Residuals

The variation of the residuals for protein vs. fat for Burger King menu items is less than the variation for fat.

r^2 (or R^2) gives the fraction of the data's variation accounted for by the model.



Variation of *y* and the Variation of the Residuals (Continued)

$R^2 = (0.762)^2 = 0.58$

58% of the variability in fat content in Burger King's menu items is accounted for by the variation in the protein content.

42% of the variability in fat content is left in the residuals.

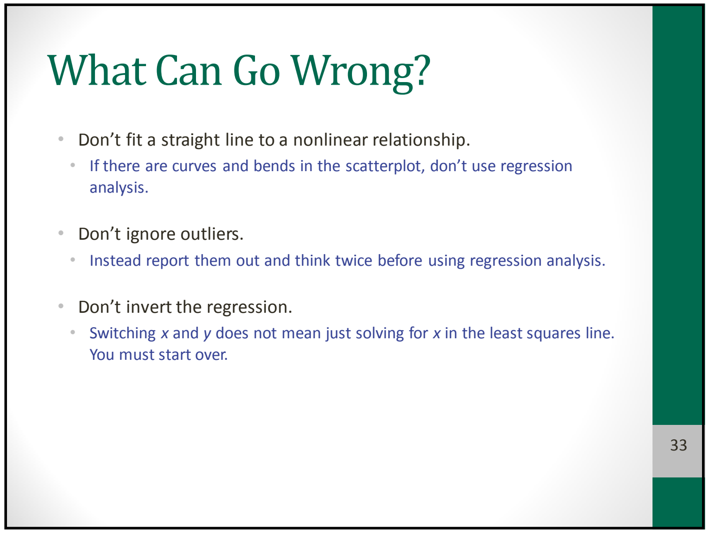
Other factors such as how the food is prepared account for this remaining variability.

Conditions to Check

- Quantitative Variable Condition: Regression analysis cannot be used for qualitative variables.
- Straight Enough Condition: The scatterplot should indicate a relatively straight pattern.
- Outlier Condition: Outliers dramatically influence the fit of the least squares line.
- Does the Plot Thicken? Condition: The data should not become more spread out as the values of *x* increase. The spread should be relatively consistent for all *x*.

What Can Go Wrong?

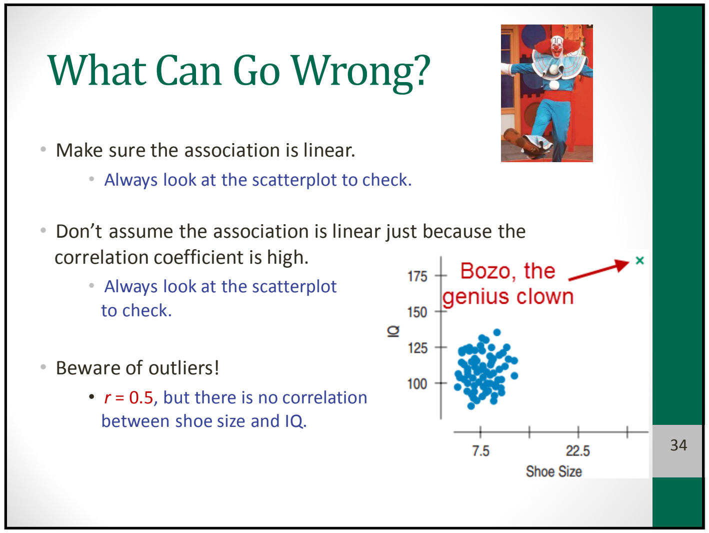
- Don't say "correlation" when you mean "association."
  - Correlation implies a **linear** relationship. Association means any relationship.
- Don't correlate categorical variables.
  - It makes no sense to say *car model* and *personality type* are correlated.
- Don't confuse correlation with causation.
  - Correlation only implies general tendencies.



What Can Go Wrong?

- Don’t fit a straight line to a nonlinear relationship.
  - If there are curves and bends in the scatterplot, don’t use regression analysis.
- Don’t ignore outliers.
  - Instead report them out and think twice before using regression analysis.
- Don’t invert the regression.
  - Switching *x* and *y* does not mean just solving for *x* in the least squares line. You must start over.

33

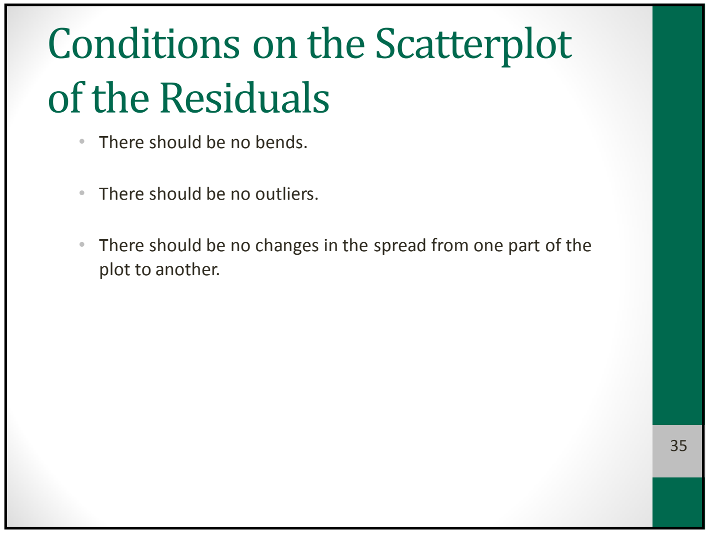


What Can Go Wrong?

- Make sure the association is linear.
  - Always look at the scatterplot to check.
- Don’t assume the association is linear just because the correlation coefficient is high.
  - Always look at the scatterplot to check.
- Beware of outliers!
  - *r* = 0.5, but there is no correlation between shoe size and IQ.

Bozo, the genius clown
IQ
175
150
125
100
7.5
22.5
Shoe Size

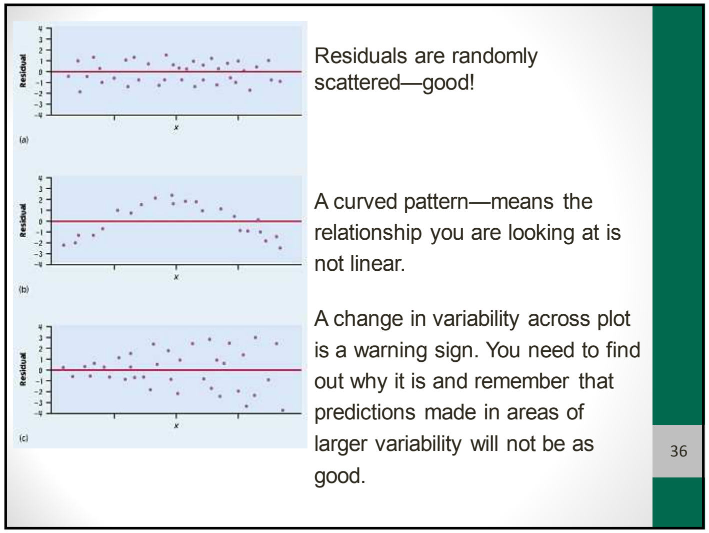
34



Conditions on the Scatterplot of the Residuals

- There should be no bends.
- There should be no outliers.
- There should be no changes in the spread from one part of the plot to another.

35



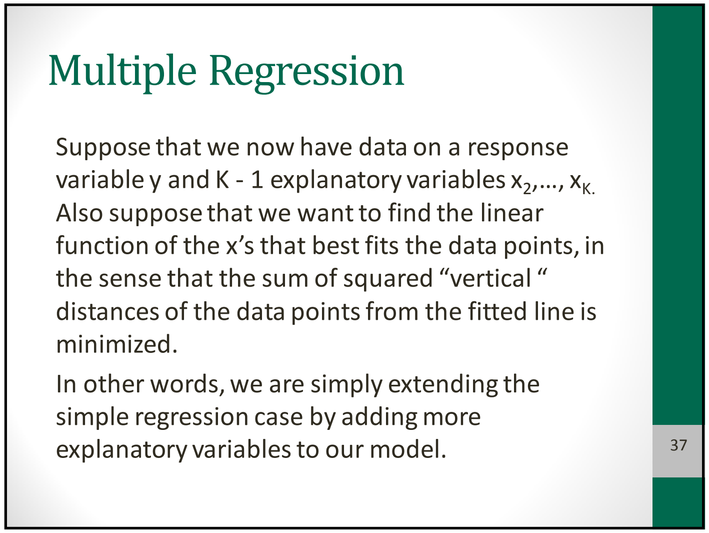
Residual
x
(a)
Residual
x
(b)
Residual
x
(c)

Residuals are randomly scattered—good!

A curved pattern—means the relationship you are looking at is not linear.

A change in variability across plot is a warning sign. You need to find out why it is and remember that predictions made in areas of larger variability will not be as good.

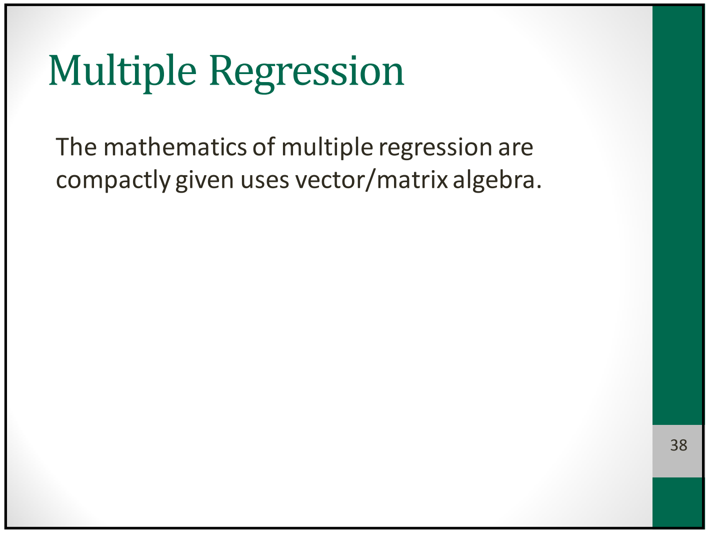
36



Multiple Regression

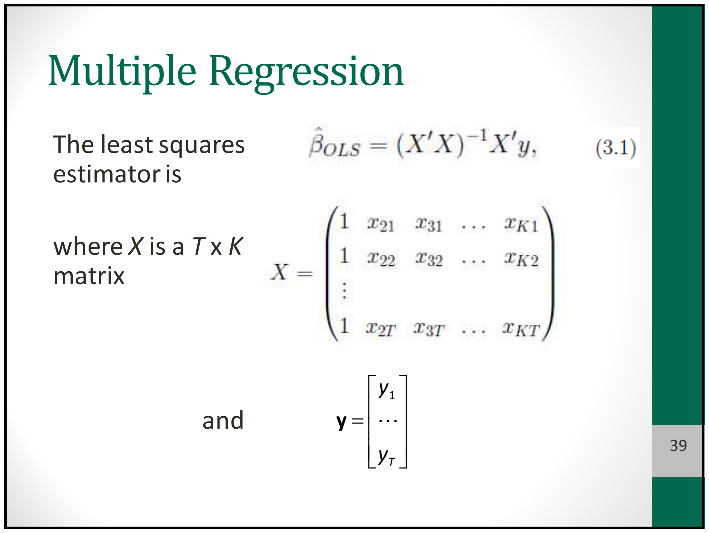
Suppose that we now have data on a response variable y and K - 1 explanatory variables $x_2,\ldots, x_{K.}$ Also suppose that we want to find the linear function of the x's that best fits the data points, in the sense that the sum of squared "vertical " distances of the data points from the fitted line is minimized.

In other words, we are simply extending the simple regression case by adding more explanatory variables to our model.



Multiple Regression

The mathematics of multiple regression are compactly given uses vector/matrix algebra.



Multiple Regression

The least squares estimator is

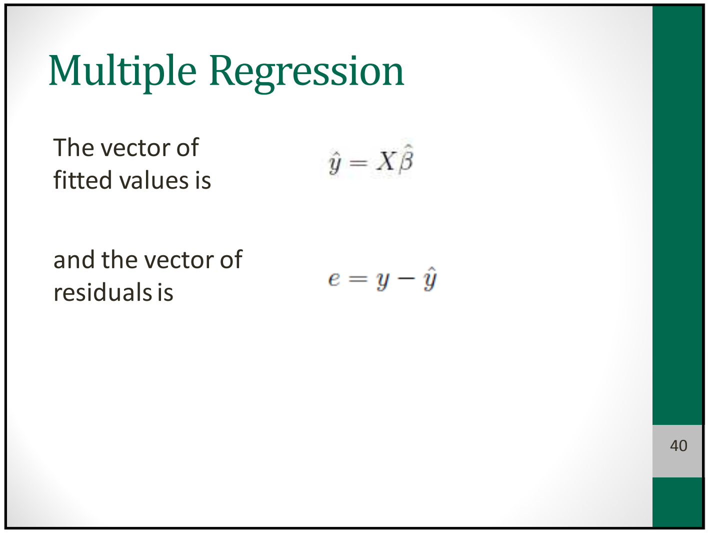
$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y, \qquad (3.1)$$

where *X* is a *T* x *K* matrix

$$X = \begin{pmatrix} 1 & x_{21} & x_{31} & \ldots & x_{K1} \\ 1 & x_{22} & x_{32} & \ldots & x_{K2} \\ \vdots & & & & \\ 1 & x_{2T} & x_{3T} & \ldots & x_{KT} \end{pmatrix}$$

and

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \cdots \\ y_T \end{bmatrix}$$



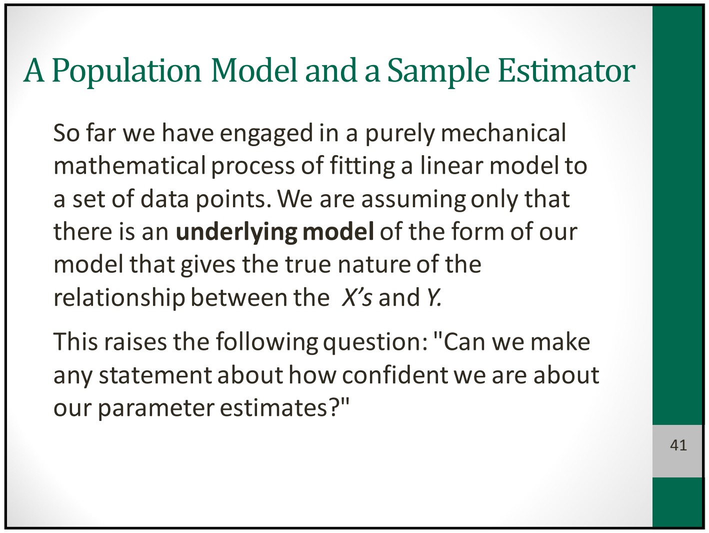
Multiple Regression

The vector of fitted values is

$$\hat{y} = X\hat{\beta}$$

and the vector of residuals is

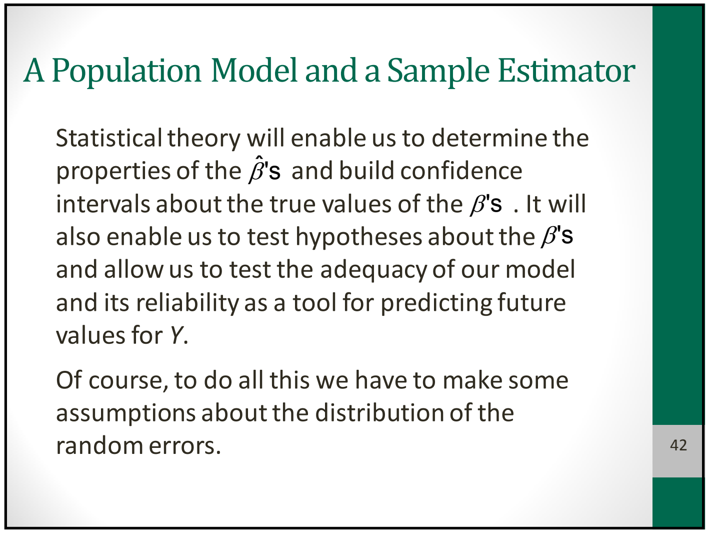
$$e = y - \hat{y}$$



A Population Model and a Sample Estimator

So far we have engaged in a purely mechanical mathematical process of fitting a linear model to a set of data points. We are assuming only that there is an **underlying model** of the form of our model that gives the true nature of the relationship between the *X's* and *Y.*

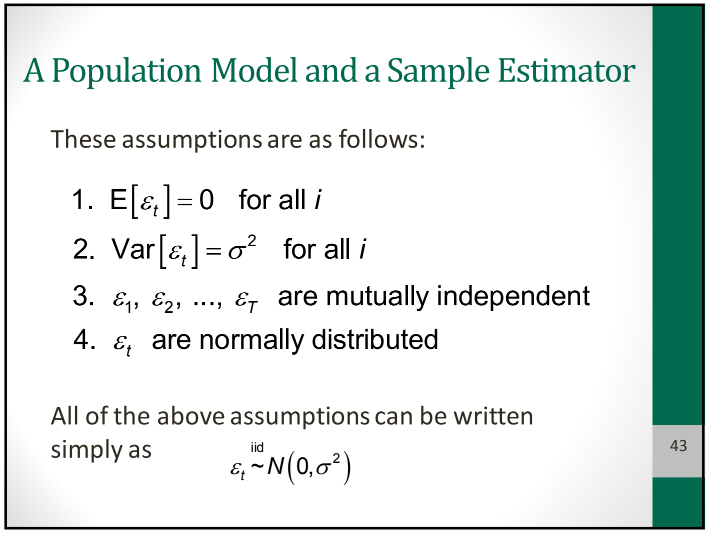
This raises the following question: "Can we make any statement about how confident we are about our parameter estimates?"



A Population Model and a Sample Estimator

Statistical theory will enable us to determine the properties of the $\hat{\beta}$'s and build confidence intervals about the true values of the β's . It will also enable us to test hypotheses about the β's and allow us to test the adequacy of our model and its reliability as a tool for predicting future values for *Y*.

Of course, to do all this we have to make some assumptions about the distribution of the random errors.

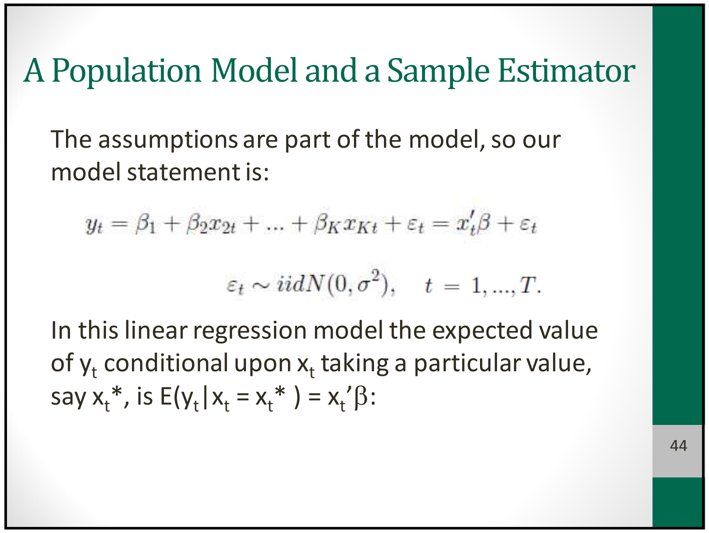


A Population Model and a Sample Estimator

These assumptions are as follows:

1. $\mathrm{E}[\varepsilon_t] = 0$ for all *i*
2. $\mathrm{Var}[\varepsilon_t] = \sigma^2$ for all *i*
3. $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_T$ are mutually independent
4. ε_t are normally distributed

All of the above assumptions can be written simply as $\varepsilon_t \overset{\text{iid}}{\sim} N(0,\sigma^2)$



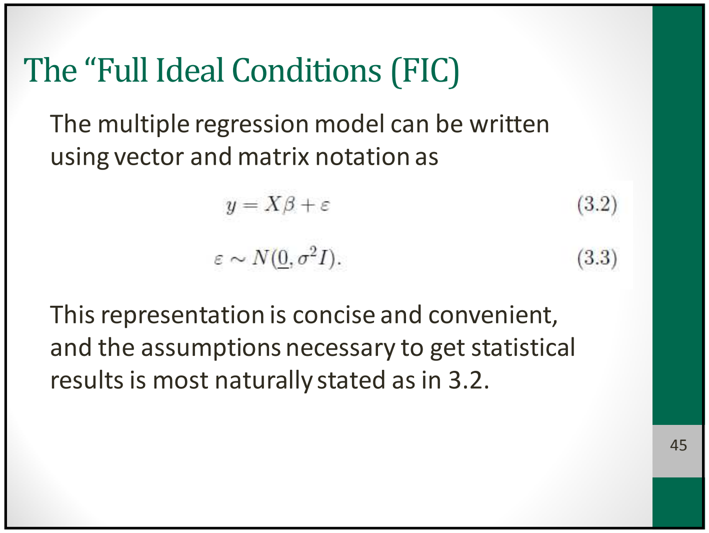
A Population Model and a Sample Estimator

The assumptions are part of the model, so our model statement is:

$$y_t = \beta_1 + \beta_2 x_{2t} + \ldots + \beta_K x_{Kt} + \varepsilon_t = x_t'\beta + \varepsilon_t$$

$$\varepsilon_t \sim iidN(0,\sigma^2), \quad t = 1, \ldots, T.$$

In this linear regression model the expected value of y_t conditional upon x_t taking a particular value, say x_t^*, is $E(y_t | x_t = x_t^*) = x_t'\beta$:



The “Full Ideal Conditions (FIC)

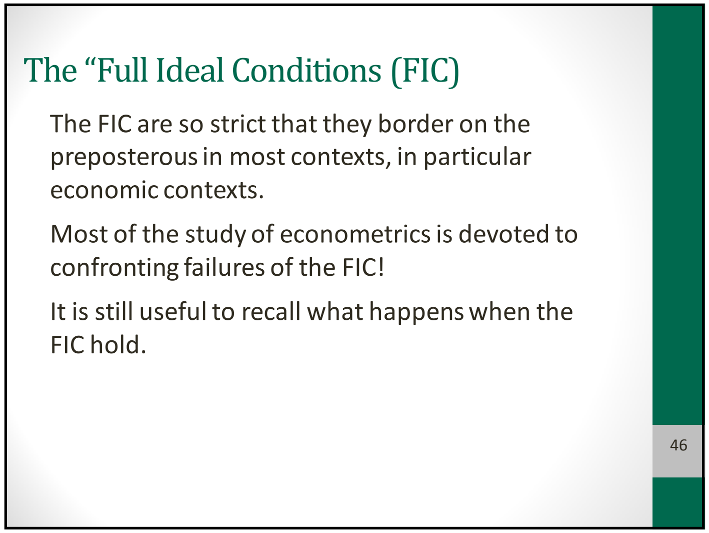
The multiple regression model can be written using vector and matrix notation as

$$y = X\beta + \varepsilon \tag{3.2}$$

$$\varepsilon \sim N(\underline{0}, \sigma^2 I). \tag{3.3}$$

This representation is concise and convenient, and the assumptions necessary to get statistical results is most naturally stated as in 3.2.

45



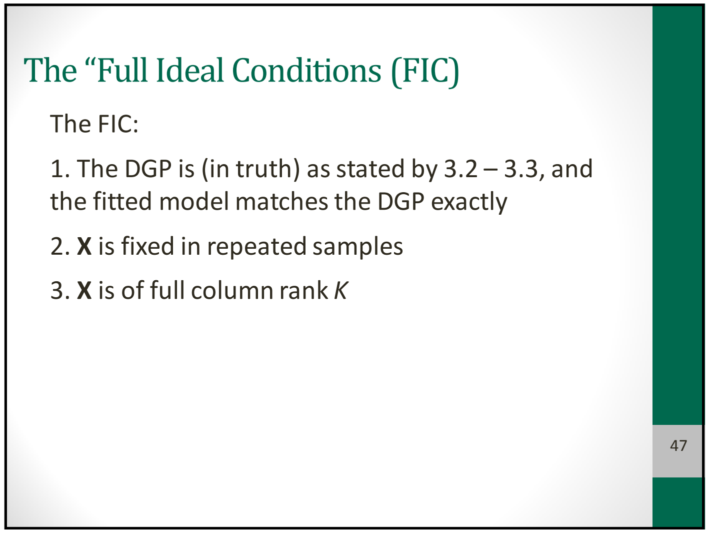
The “Full Ideal Conditions (FIC)

The FIC are so strict that they border on the preposterous in most contexts, in particular economic contexts.

Most of the study of econometrics is devoted to confronting failures of the FIC!

It is still useful to recall what happens when the FIC hold.

46



The “Full Ideal Conditions (FIC)

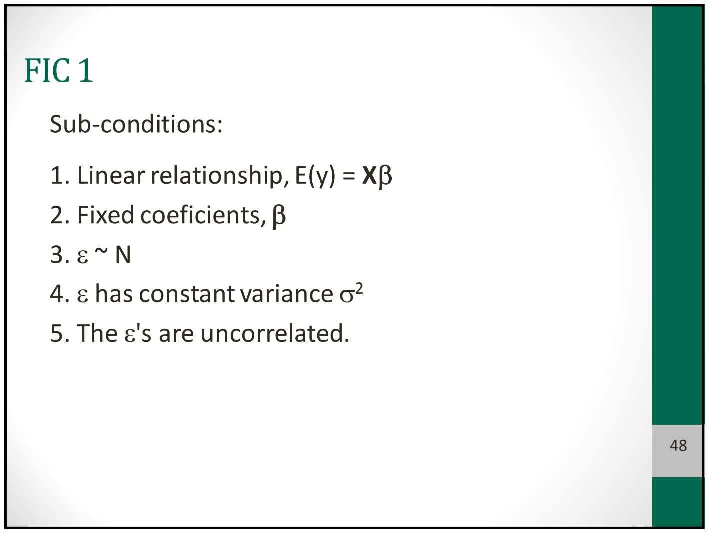
The FIC:

1. The DGP is (in truth) as stated by 3.2 – 3.3, and the fitted model matches the DGP exactly

2. **X** is fixed in repeated samples

3. **X** is of full column rank *K*

47



FIC 1

Sub-conditions:

1. Linear relationship, E(y) = **X**β

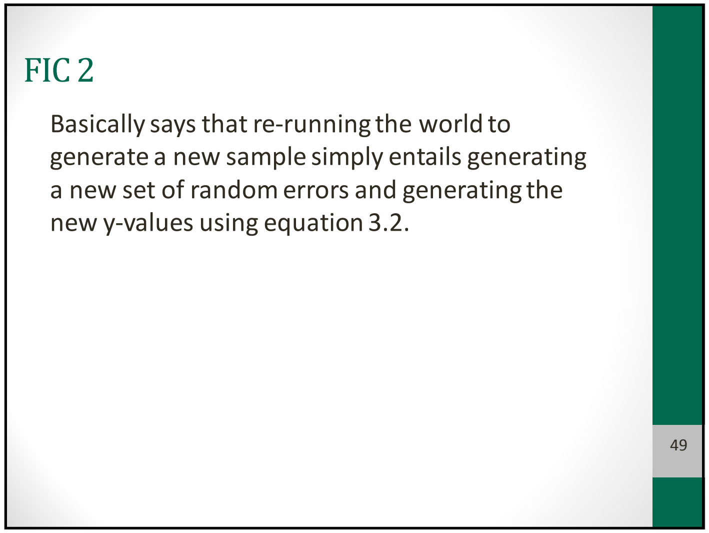
2. Fixed coeficients, β

3. ε ~ N

4. ε has constant variance σ^2

5. The ε's are uncorrelated.

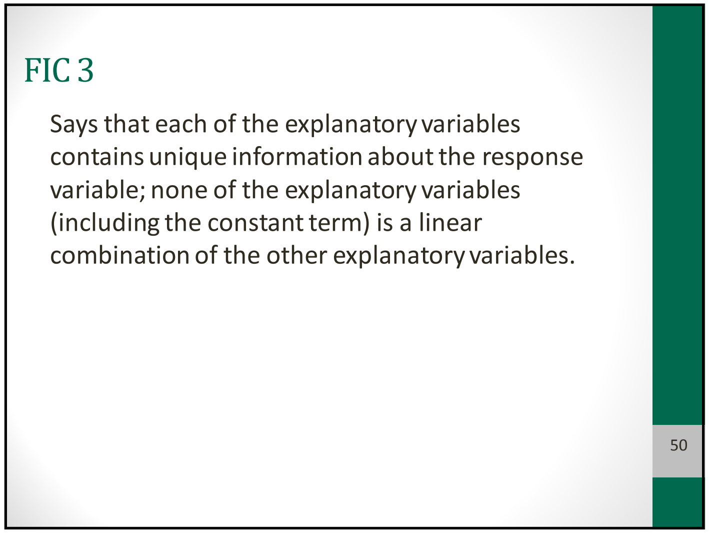
48



FIC 2

Basically says that re-running the world to generate a new sample simply entails generating a new set of random errors and generating the new y-values using equation 3.2.

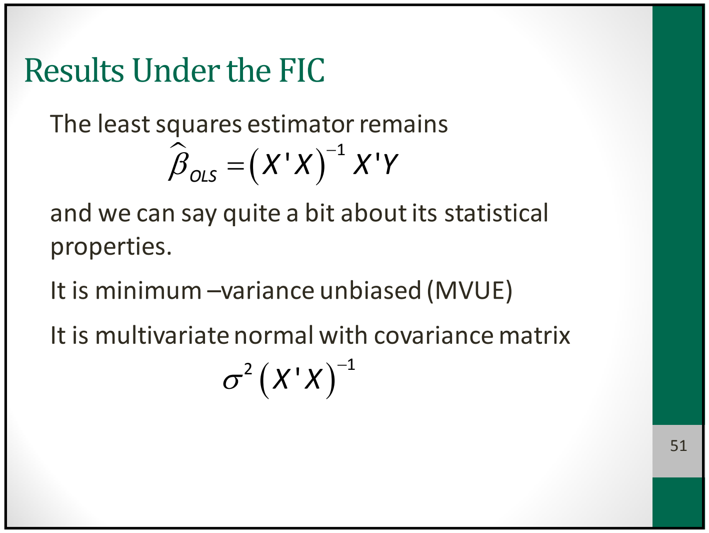
49



FIC 3

Says that each of the explanatory variables contains unique information about the response variable; none of the explanatory variables (including the constant term) is a linear combination of the other explanatory variables.

50



Results Under the FIC

The least squares estimator remains

$$\hat{\beta}_{OLS} = \left(X'X\right)^{-1} X'Y$$

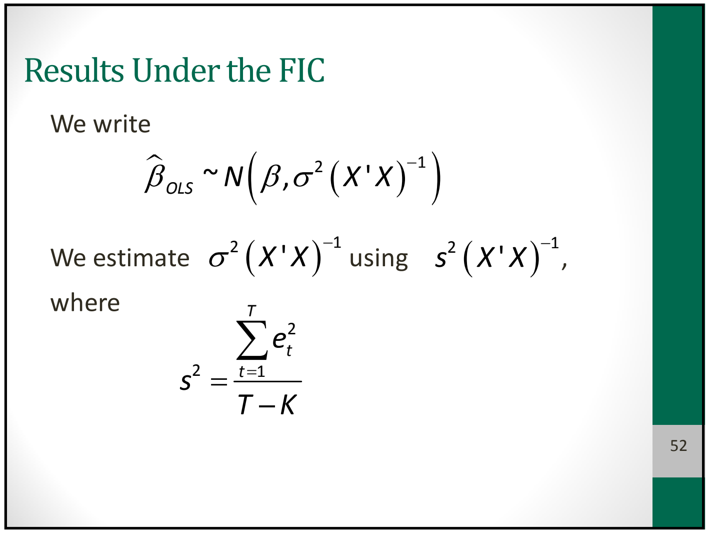
and we can say quite a bit about its statistical properties.

It is minimum –variance unbiased (MVUE)

It is multivariate normal with covariance matrix

$$\sigma^2 \left(X'X\right)^{-1}$$

51



Results Under the FIC

We write

$$\hat{\beta}_{OLS} \sim N\left(\beta, \sigma^2 \left(X'X\right)^{-1}\right)$$

We estimate $\sigma^2 \left(X'X\right)^{-1}$ using $s^2 \left(X'X\right)^{-1}$,

where

$$s^2 = \frac{\sum_{t=1}^{T} e_t^2}{T-K}$$

52