AP Statistics
Chapter 8

Linear Regression

Did You Mean Association Or Correlation?

- Be careful not to use the word **correlation** when you really mean **association**. Often times people will incorrectly use the word **correlation** when talking about relationships in order to sound scientific. However, **associations** just describe a general relationship between two variable whereas **correlations** specifically describes the **<u>linear</u>** relationship between the two variables if any.

Always Check Your Conditions

- The conditions for correlation:
  - ❖ **The variables must be numerical**.
    - ⇨ People who misuse correlation to mean association often fail to notice whether the variables they discuss are quantitative
  - ❖ **The association is linear**.
    - ⇨ Correlations only describe linear associations
  - ❖ **No outliers**.
    - ⇨ Outliers can drastically change your data. Always be aware of any points that may sway your data.

Linear Regression

- It would be great to be able to look at multi-variable data and reduce it to a single equation that might help us make predictions
- "Given the data of tuition at Arizona State University during the 1990's, can you predict the tuition in 2002?"
- Let's take this step by step to see how to perform a linear regression

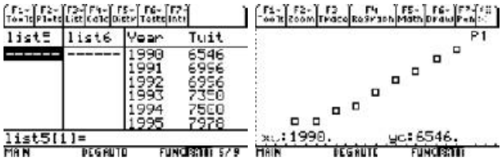
Linear Regression

- Make a new list labeled Year and Tuit (for tuition)
- Then, input the following data into your calculator

Year	Tuition	Year	Tuition
1990	6546	1996	8377
1991	6996	1997	8710
1992	6996	1998	9110
1993	7350	1999	9411
1994	7500	2000	9800
1995	7978		

Linear Regression

- Next, check your conditions.

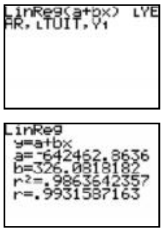


```
list5   list6   Year  Tuit
------- ------- 1990  6546
                1991  6956
                1992  6956
                1993  7350
                1994  7500
                1995  7978
list5[1]=
```

- Are the variables quantitative?
- Does the data look somewhat linear?
- Are there outliers?

Linear Regression

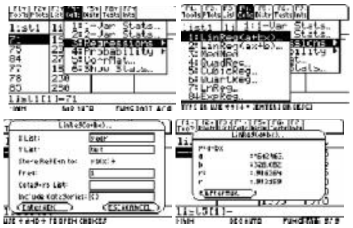
- Now, let’s calculate the linear regression line



TI-84

```
LinReg(a+bx) ʟYE
AR,ʟTUIT,Y1
```

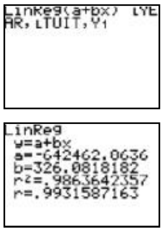
```
LinReg
 y=a+bx
 a=-642462.8636
 b=326.0818182
 r²=.9863642357
 r=.9931587163
```



TI-89

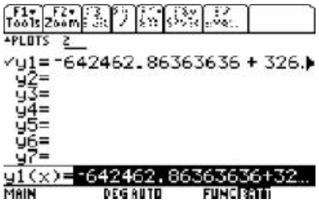
Linear Regression

- The Y_1 variable automatically inputs the “**Least-Squared Regression Line**” (also called the **LSRL**) into the Y_1 function in your calculator:



```
LinReg(a+bx) ʟYE
AR,ʟTUIT,Y1
```

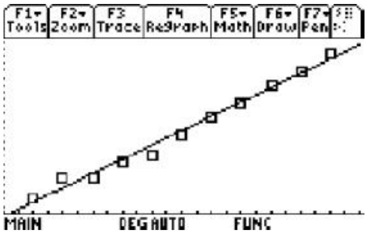
```
LinReg
 y=a+bx
 a=-642462.0636
 b=326.0818182
 r²=.9863642357
 r=.9931587163
```



```
▴PLOTS 2
✓y1=-642462.86363636 + 326.▸
 y2=
 y3=
 y4=
 y5=
 y6=
 y7=
y1(x)=-642462.86363636+32...
MAIN          DEG AUTO       FUNC
```

The Least-Square Regression

- The LSRL finds the best fit line by trying to minimize the areas formed by the difference of the real data from the predicted data.



The Least-Square Regression

- The LSRL helps us make predictions and creates a line that “best fits” the data.
- It is called the Least Squares Regression Line because it is the ONE line that has the smallest Least Squares Error – it gives the smallest sum of squared deviation.
- The LSRL equation that we received from the Arizona Tuition problem was:

$$\hat{y} = -642{,}463 + 326.08x$$

- The LSRL helps us make predictions and creates a line that “best fits” the data.
- What is the y-intercept of the line? a = -642463
- What is the slope of the line? b = 326.08

The Least-Square Regression

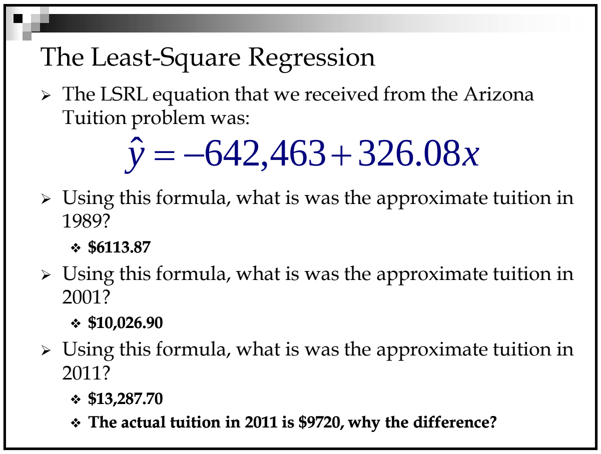
- The LSRL equation that we received from the Arizona Tuition problem was:

$$\hat{y} = -642{,}463 + 326.08x$$

- What does the y-intercept represent?
  - **It represents the tuition at year 0**
- Does the y-intercept make sense in the context of this problem?
  - **No, since at year 0 was during Jesus’ time, it doesn’t make sense to speak of the tuition of Arizona State during this time frame! Plus, it means they would pay you to attend!!!**
- What does the slope represent?
  - **It represents the amount of money that tuition will raise for every increase of 1 year. In this example, the model predicts that tuition will raise $326.08 every year at Arizona State.**

The Least-Square Regression

- **Note: when asked about the y-intercept (*a*) and the slope (*b*), you should memorize this phrase:**
  - ***y-intercept (a)*:** at an (**explanatory variable**) value of 0 (**units**), our model predicts a (**response variable**) of (**y units**).
    - ⇨ Always ask if this makes sense!!!
  - ***Slope (b)*:** for every (**1 unit**) increase in the (**explanatory variable**), our model predicts an average (**increase/decrease**) of (**y units**) in the (**response variable**).
- Let’s apply these phrases with our Arizona State example…
  - ***y-intercept (a)*:** at **year 0**, our model predicts a tuition of -$642,463.
    - ⇨ This makes no sense at all!!!
  - ***Slope (b)*:** for every **1 year** increase, our model predicts an average **increase** of **$326.08** in the **tuition**.

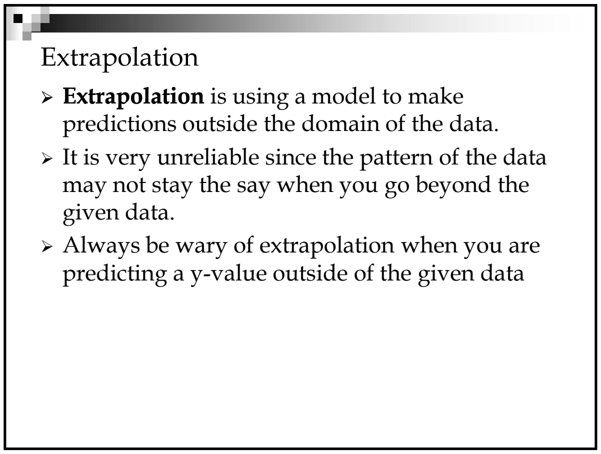


The Least-Square Regression

- The LSRL equation that we received from the Arizona Tuition problem was:

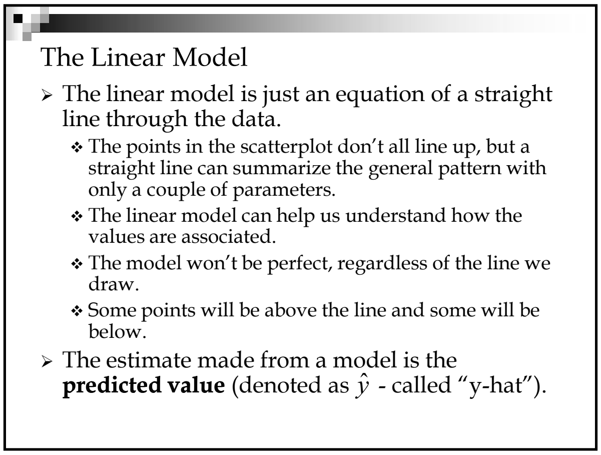
$$\hat{y} = -642{,}463 + 326.08x$$

- Using this formula, what is was the approximate tuition in 1989?
  - **$6113.87**
- Using this formula, what is was the approximate tuition in 2001?
  - **$10,026.90**
- Using this formula, what is was the approximate tuition in 2011?
  - **$13,287.70**
  - **The actual tuition in 2011 is $9720, why the difference?**



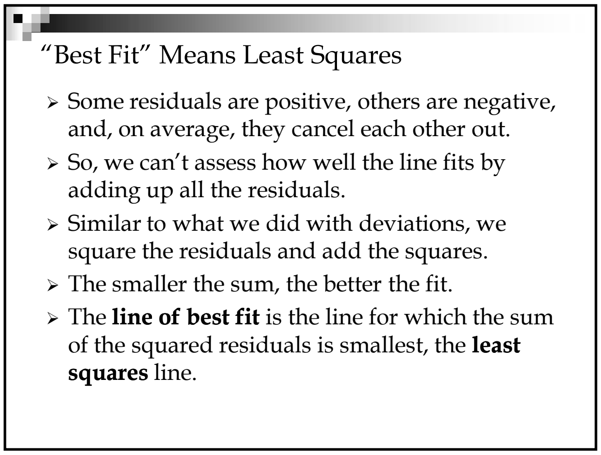
Extrapolation

- **Extrapolation** is using a model to make predictions outside the domain of the data.
- It is very unreliable since the pattern of the data may not stay the say when you go beyond the given data.
- Always be wary of extrapolation when you are predicting a y-value outside of the given data



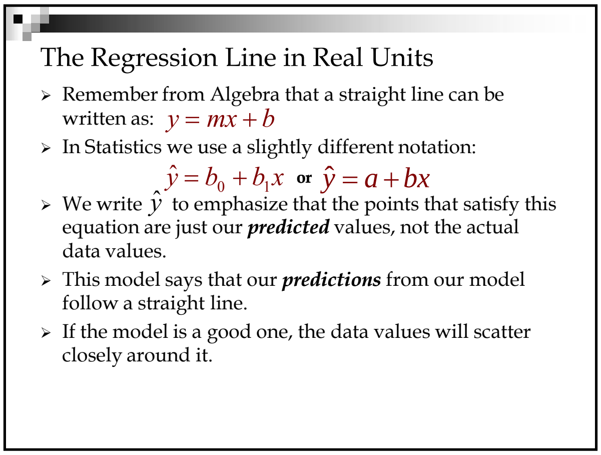
The Linear Model

- The linear model is just an equation of a straight line through the data.
  - The points in the scatterplot don't all line up, but a straight line can summarize the general pattern with only a couple of parameters.
  - The linear model can help us understand how the values are associated.
  - The model won't be perfect, regardless of the line we draw.
  - Some points will be above the line and some will be below.
- The estimate made from a model is the **predicted value** (denoted as $\hat{y}$ - called "y-hat").



"Best Fit" Means Least Squares

- Some residuals are positive, others are negative, and, on average, they cancel each other out.
- So, we can't assess how well the line fits by adding up all the residuals.
- Similar to what we did with deviations, we square the residuals and add the squares.
- The smaller the sum, the better the fit.
- The **line of best fit** is the line for which the sum of the squared residuals is smallest, the **least squares** line.

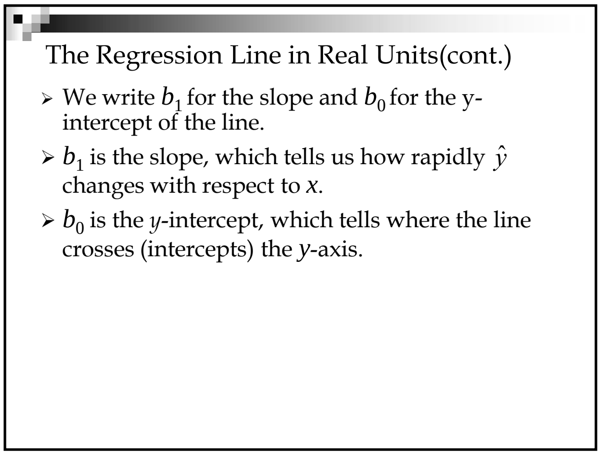


The Regression Line in Real Units

- Remember from Algebra that a straight line can be written as: $y = mx + b$
- In Statistics we use a slightly different notation:

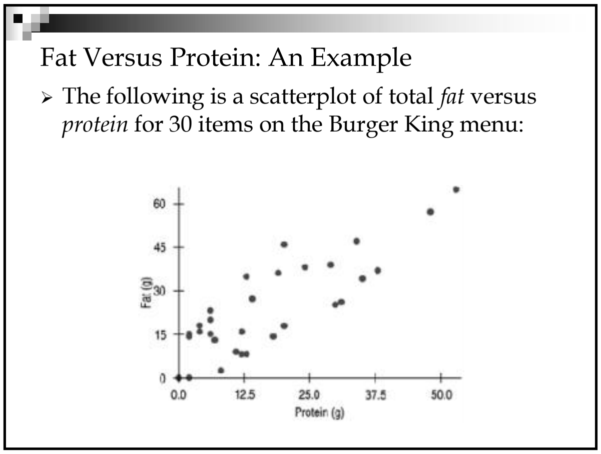
$$\hat{y} = b_0 + b_1 x \quad \textbf{or} \quad \hat{y} = a + bx$$

- We write $\hat{y}$ to emphasize that the points that satisfy this equation are just our ***predicted*** values, not the actual data values.
- This model says that our ***predictions*** from our model follow a straight line.
- If the model is a good one, the data values will scatter closely around it.



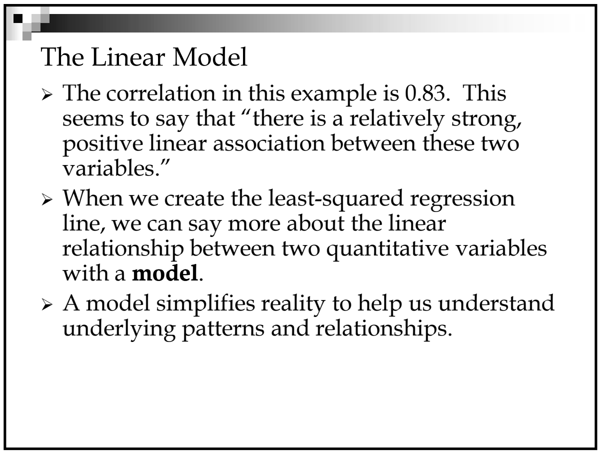
The Regression Line in Real Units(cont.)

- We write b_1 for the slope and b_0 for the y-intercept of the line.
- b_1 is the slope, which tells us how rapidly $\hat{y}$ changes with respect to *x*.
- b_0 is the *y*-intercept, which tells where the line crosses (intercepts) the *y*-axis.



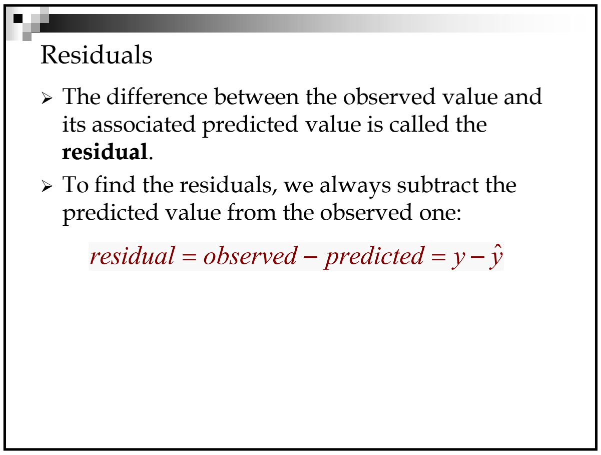
Fat Versus Protein: An Example

- The following is a scatterplot of total *fat* versus *protein* for 30 items on the Burger King menu:



The Linear Model

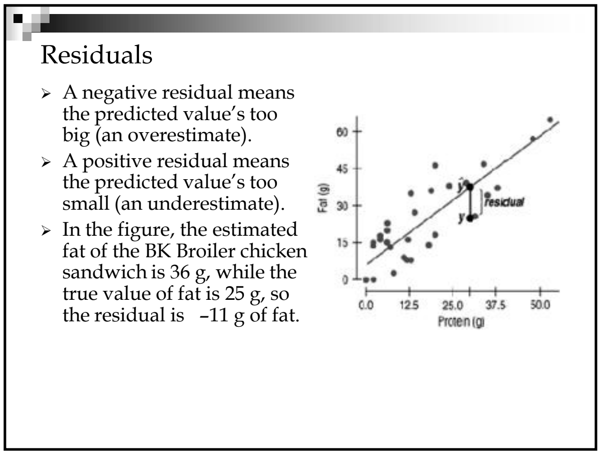
- The correlation in this example is 0.83. This seems to say that "there is a relatively strong, positive linear association between these two variables."
- When we create the least-squared regression line, we can say more about the linear relationship between two quantitative variables with a **model**.
- A model simplifies reality to help us understand underlying patterns and relationships.



Residuals

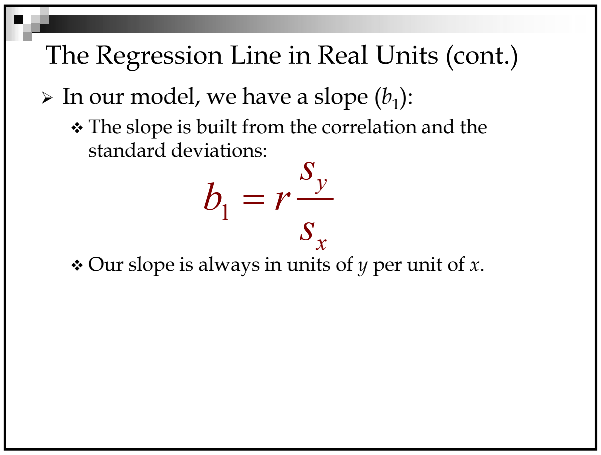
- The difference between the observed value and its associated predicted value is called the **residual**.
- To find the residuals, we always subtract the predicted value from the observed one:

$$\textit{residual} = \textit{observed} - \textit{predicted} = y - \hat{y}$$



Residuals

- A negative residual means the predicted value's too big (an overestimate).
- A positive residual means the predicted value's too small (an underestimate).
- In the figure, the estimated fat of the BK Broiler chicken sandwich is 36 g, while the true value of fat is 25 g, so the residual is –11 g of fat.

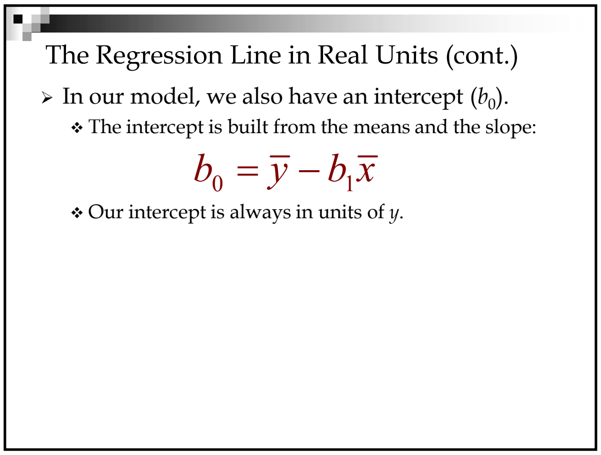


The Regression Line in Real Units (cont.)

- In our model, we have a slope (b_1):
  - The slope is built from the correlation and the standard deviations:

$$b_1 = r\frac{s_y}{s_x}$$

  - Our slope is always in units of *y* per unit of *x*.

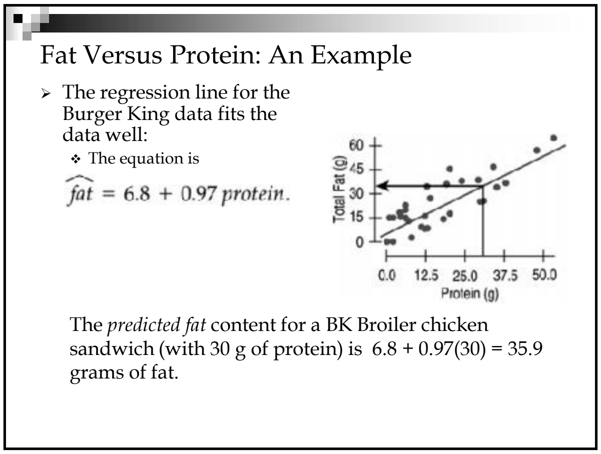


The Regression Line in Real Units (cont.)

- In our model, we also have an intercept (b_0).
  - The intercept is built from the means and the slope:

$$b_0 = \bar{y} - b_1\bar{x}$$

  - Our intercept is always in units of *y*.

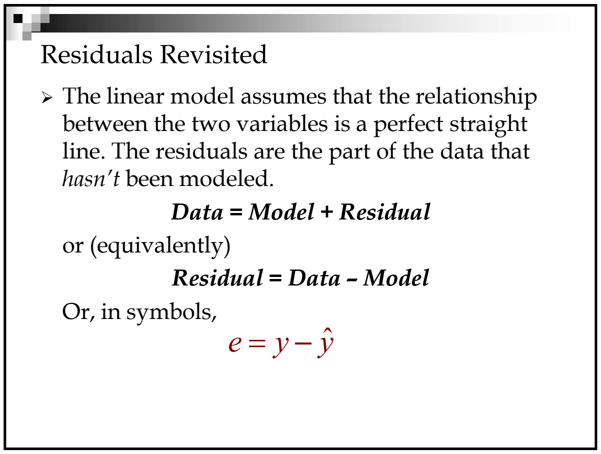


Fat Versus Protein: An Example

- The regression line for the Burger King data fits the data well:
  - The equation is

$$\widehat{fat} = 6.8 + 0.97\,protein.$$

The *predicted fat* content for a BK Broiler chicken sandwich (with 30 g of protein) is 6.8 + 0.97(30) = 35.9 grams of fat.



Residuals Revisited

- The linear model assumes that the relationship between the two variables is a perfect straight line. The residuals are the part of the data that *hasn't* been modeled.

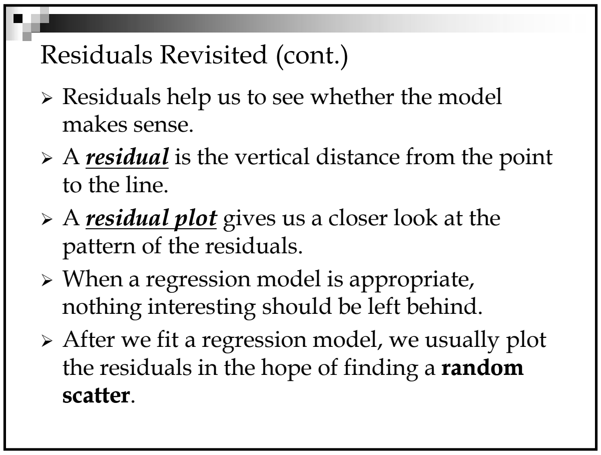
  Data = Model + Residual

  or (equivalently)

  Residual = Data – Model

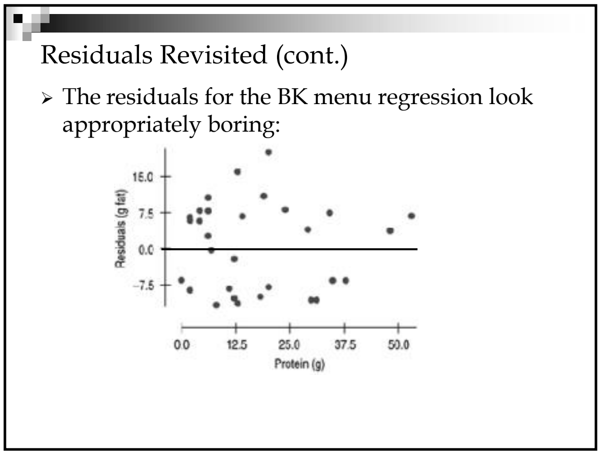
  Or, in symbols,

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



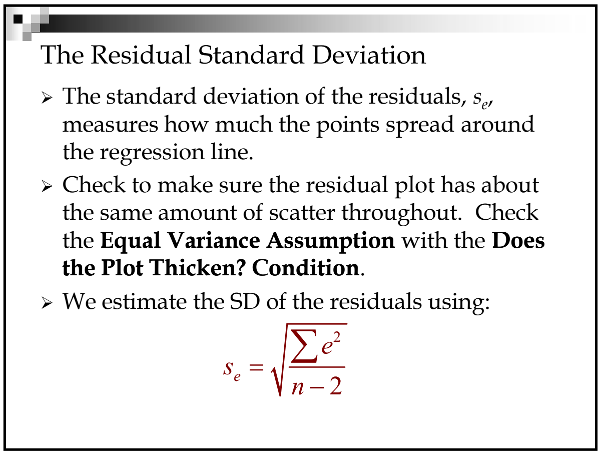
Residuals Revisited (cont.)

- Residuals help us to see whether the model makes sense.
- A ***residual*** is the vertical distance from the point to the line.
- A ***residual plot*** gives us a closer look at the pattern of the residuals.
- When a regression model is appropriate, nothing interesting should be left behind.
- After we fit a regression model, we usually plot the residuals in the hope of finding a **random scatter**.



Residuals Revisited (cont.)

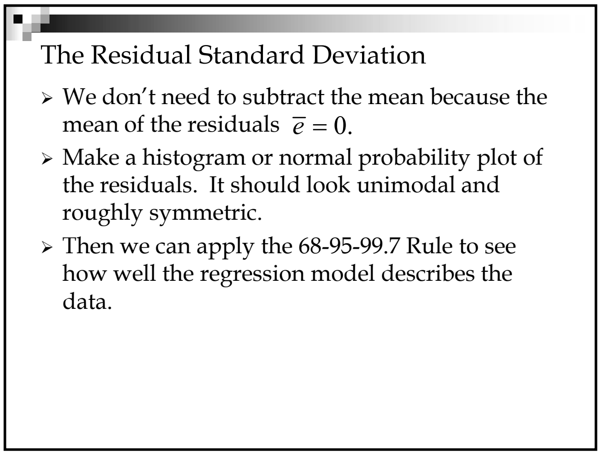
- The residuals for the BK menu regression look appropriately boring:



The Residual Standard Deviation

- The standard deviation of the residuals, s_e, measures how much the points spread around the regression line.
- Check to make sure the residual plot has about the same amount of scatter throughout. Check the **Equal Variance Assumption** with the **Does the Plot Thicken? Condition**.
- We estimate the SD of the residuals using:

$$s_e = \sqrt{\frac{\sum e^2}{n-2}}$$

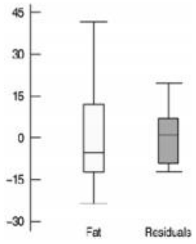


The Residual Standard Deviation

- We don't need to subtract the mean because the mean of the residuals $\bar{e} = 0$.
- Make a histogram or normal probability plot of the residuals. It should look unimodal and roughly symmetric.
- Then we can apply the 68-95-99.7 Rule to see how well the regression model describes the data.

R^2—The Variation Accounted For

- The variation in the residuals is the key to assessing how well the model fits.
- In the BK menu items example, total *fat* has a standard deviation of 16.4 grams. The standard deviation of the residuals is 9.2 grams.



R^2—The Variation Accounted For (cont.)

- Since the correlation (r) is 0.83 (not perfect) we cannot **perfectly** predict fat from protein.
- However, we can determine how much of the variation is accounted for by the model and how much is left in the residuals.
- The squared correlation, ***r^2* – called the Coefficient of Determination**, gives the fraction of the data's variance accounted for by the model.
- **If *r^2* = 1.0,** the model would predict the ***fat*** values ***perfectly*** without error from ***protein***.
- **If *r^2* = 0**, **fat** could not be predicted from **protein** at all.

R^2—The Variation Accounted For (cont.)

- ***r^2*** is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable.
- Thus, **1 – *r^2*** is the fraction of the original variance left in the residuals.
- **Note: when asked about *r^2*, you should memorize this phrase:**
  - ***r^2*:** (**x**) percent of the variation in the (**response variable**) can be explained by the approximate linear relationship with the (**explanatory variable**).
- Let's apply this phrase with our Burger King example…

R^2—The Variation Accounted For (cont.)

- ***Apply the phrase*:** (**x**) percent of the variation in the (**response variable**) can be explained by the linear relationship with the (**explanatory variable**).
- For the BK model, ***r*** = 0.83 and ***r^2*** = 0.83^2 = 0.69, so 69% of the variation in total fat can be explained by the linear relationship with protein.
- Always try to understand what we are talking about and don't get lost in the wording. 69% of the variability is accounted for by the Linear Regression Line and 31% of the variability in total *fat* has been left in the residuals.

R^2—The Variation Accounted For (cont.)

- All regression analyses include this statistic, although by tradition, it is written ***R^2** or **r^2*** (pronounced "*r*-squared").
- ***r^2*** is always between 0% and 100%. What makes a "good" ***r^2*** value depends on the kind of data you are analyzing and on what you want to do with it.
- An ***r^2*** between 0 and 1 indicates the extent to which the dependent variable is predictable. An ***r^2*** of 0.10 means that 10 percent of the variance in *Y* is predictable from *X*; an ***r^2*** of 0.20 means that 20 percent is predictable; and so on.

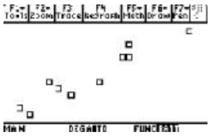
Reporting R^2

- Along with the slope, intercept, and correlation for a regression, you should always report ***r^2*** so that readers can judge for themselves how successful the regression is at fitting the data.
- Statistics is about variation, and ***r^2*** measures the success of the regression model in terms of the fraction of the variation of *y* accounted for by the regression model.

SAT Scores vs. College GPA

SAT	1290	1335	1425	1470	1530	1665	1770	1800	1800	2085
GPA	2.8	2.7	3.2	3.1	3.0	3.2	3.6	3.8	3.6	4.0

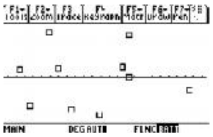
- Is it reasonable to perform linear regression on this data? How do you know?
  - **Yes, it is reasonable since we satisfy our 3 conditions: the data is quantitative, the scatter is linear, and there are no outliers.**



SAT Scores vs. College GPA

SAT	1290	1335	1425	1470	1530	1665	1770	1800	1800	2085
GPA	2.8	2.7	3.2	3.1	3.0	3.2	3.6	3.8	3.6	4.0

- Looking at the residuals, is it still appropriate to perform linear regression?
  - **Yes, it is appropriate to perform linear regression since the residuals show a random scatter about the line.**



SAT Scores vs. College GPA

SAT	1290	1335	1425	1470	1530	1665	1770	1800	1800	2085
GPA	2.8	2.7	3.2	3.1	3.0	3.2	3.6	3.8	3.6	4.0

- Describe the association.
  - **There is a strong, positive, linear association between SAT score and GPA. The correlation is 0.948 indicating that the strength is very strong.**
- Describe the variation.
  - **The variation can be analyzed by looking at r^2.**
  - **Since r = 0.948, r^2 = 0.899. So, 89.9% of the variation in the college GPA can be explained by the linear relationship with SAT score.**

SAT Scores vs. College GPA

SAT	1290	1335	1425	1470	1530	1665	1770	1800	1800	2085
GPA	2.8	2.7	3.2	3.1	3.0	3.2	3.6	3.8	3.6	4.0

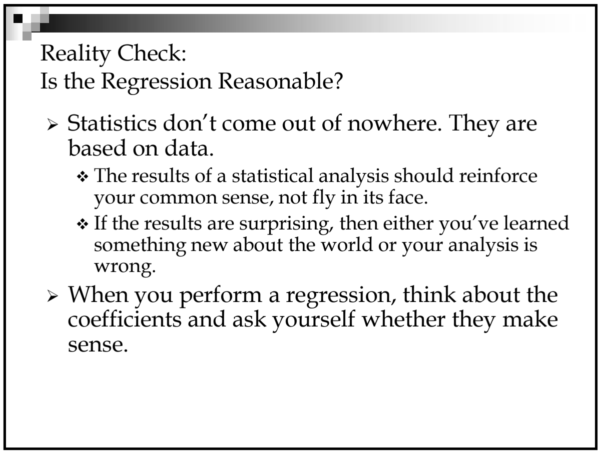
- If a student has an SAT score of 2000, what would you predict his college GPA will be?
  - **Approximately 3.93**
- If a student has an SAT score of 2350, what would you predict his college GPA will be?
  - **Approximately 4.50**
- Does this seem reasonable?
  - **No, in college, you can only get a 4.0**
- What went wrong? Why?

Assumptions and Conditions

- Before performing any linear regression, check the conditions:
  - Quantitative Variables Condition:
  - Straight Enough Condition:
    - If the scatterplot is not straight enough, stop here.
    - You **can't** use a linear model for *any* two variables, even if they are related.
    - They must have a *linear* association or the model won't mean a thing.
    - Some nonlinear relationships can be saved by re-expressing the data to make the scatterplot more linear.
  - Outlier Condition:

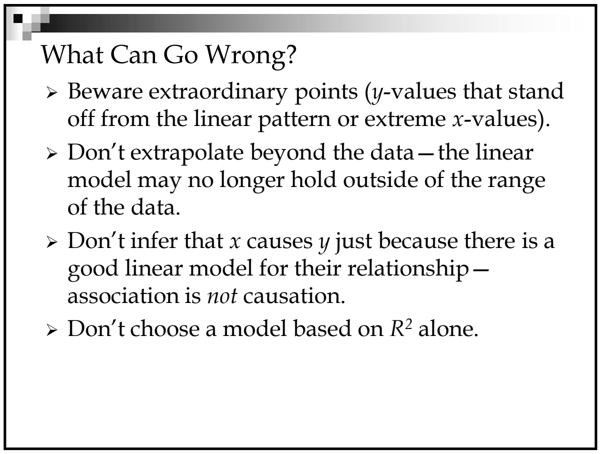
Assumptions and Conditions (cont.)

- It's a good idea to check linearity again *after* computing the regression when we can examine the residuals.
- Equal Variance Condition:
  - Check to see if the plot thickens! Look at the residual plot -- for the standard deviation of the residuals to summarize the scatter, the residuals should share the same spread. Check for changing spread in the residual scatterplot.



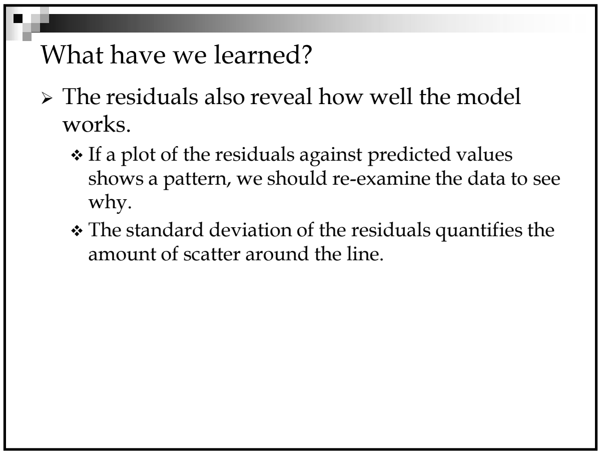
Reality Check:
Is the Regression Reasonable?

- Statistics don’t come out of nowhere. They are based on data.
  - The results of a statistical analysis should reinforce your common sense, not fly in its face.
  - If the results are surprising, then either you’ve learned something new about the world or your analysis is wrong.
- When you perform a regression, think about the coefficients and ask yourself whether they make sense.



What Can Go Wrong?

- Beware extraordinary points (*y*-values that stand off from the linear pattern or extreme *x*-values).
- Don’t extrapolate beyond the data—the linear model may no longer hold outside of the range of the data.
- Don’t infer that *x* causes *y* just because there is a good linear model for their relationship—association is *not* causation.
- Don’t choose a model based on R^2 alone.



What have we learned?

- The residuals also reveal how well the model works.
  - If a plot of the residuals against predicted values shows a pattern, we should re-examine the data to see why.
  - The standard deviation of the residuals quantifies the amount of scatter around the line.