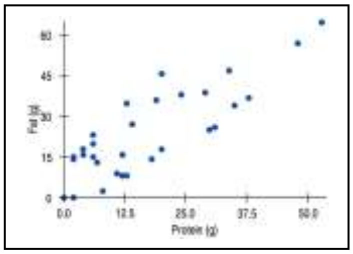
Chapter 8 Summary
Linear Regression

What have we learned?

- When the relationship between two quantitative variables is fairly straight, a linear model can help summarize that relationship.
  - The regression line doesn't pass through all the points, but it is the best compromise in the sense that it has the smallest sum of squared residuals.
- The correlation tells us several things about the regression:
  - The slope of the line is based on the correlation, adjusted for the units of *x* and *y*.
  - For each SD in *x* that we are away from the *x* mean, we expect to be *r* SDs in *y* away from the *y* mean.
  - Since *r* is always between -1 and +1, each predicted *y* is fewer SDs away from its mean than the corresponding *x* was (regression to the mean).
  - R^2 gives us the fraction of the response accounted for by the regression model.
- 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.

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

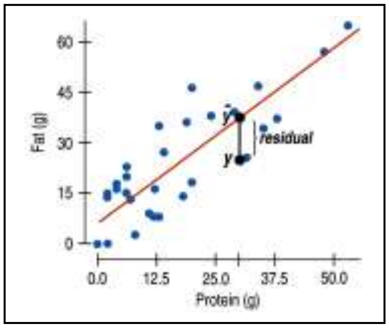
- Correlation says "There seems to be a linear association between these two variables," but it doesn't tell *what that association is*.
- We can say more about the linear relationship between two quantitative variables with a model.
- A model simplifies reality to help us understand underlying patterns and relationships.
- 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.
- The linear model can help us understand how the values are associated.

Residuals

- 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 $\hat{y}$).
- 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:

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

- A negative residual means the predicted value's too big (an overestimate).
- A positive residual means the predicted value's too small (an underestimate).



"Best Fit" Means Least Squares

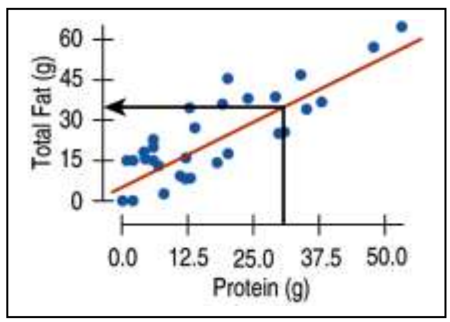
- Some residuals are positive, others are negative, and, on average, they cancel each other out.
- So, we can't assess how well the line fits by adding up all the residuals.
- Similar to what we did with deviations, we square the residuals and add the squares.
- The smaller the sum, the better the fit.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

The Least Squares Line

- We write our model as $\hat{y} = b_0 + b_1 x$
- 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.
- 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*.
- 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*.
- Since regression and correlation are closely related, we need to check the same conditions for regressions as we did for correlations:
  - Quantitative Variables Condition
  - Straight Enough Condition
  - Outlier Condition

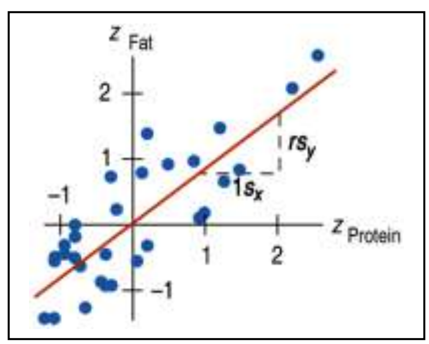
Fat Versus Protein: An Example

- 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 is
    6.8 + 0.97(30) = 35.9 grams of fat.



Correlation and the Line

- Moving one standard deviation away from the mean in *x* moves us *r* standard deviations away from the mean in *y*.
- This relationship is shown in a scatterplot of *z*-scores for *fat* and *protein*:
- Put generally, moving any number of standard deviations away from the mean in *x* moves us *r* times that number of standard deviations away from the mean in *y*.



How Big Can Predicted Values Get?

- *r* cannot be bigger than 1 (in absolute value), so each predicted *y* tends to be closer to its mean (in standard deviations) than its corresponding *x* was.
- This property of the linear model is called regression to the mean; the line is called the regression line.

Residuals Revisited

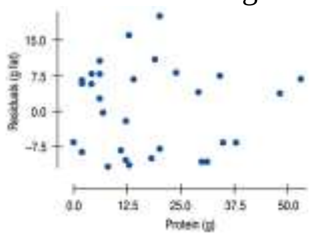
The linear model assumes that the relationship between the two variables is a perfect straight line. The residuals are the part of the data that *hasn't* been modeled.

Data = *Model* + *Residual* or (equivalently) *Residual* = *Data* – *Model*

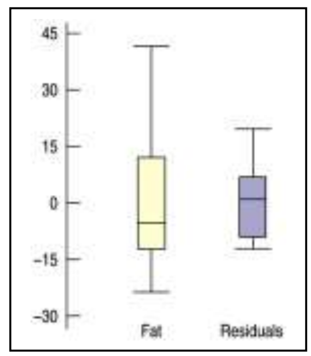
Or, in symbols, $e = y - \hat{y}$

Residuals Revisited (cont.)

- Residuals help us to see whether the model makes sense.
- 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…nothing.
- The residuals for the BK menu regression look appropriately boring:



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.
- If the correlation were 1.0 and the model predicted the *fat* values perfectly, the residuals would all be zero and have no variation.
- As it is, the correlation is 0.83—not perfection.
- However, we did see that the model residuals had less variation than total *fat* alone.
- 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, gives the fraction of the data's variance accounted for by the model.
- Thus, $1 - r^2$ is the fraction of the original variance left in the residuals.
- For the BK model, $r^2 = 0.83^2 = 0.69$, so 31% of the variability in total *fat* has been left in the residuals.
- All regression analyses include this statistic, although by tradition, it is written R^2 (pronounced "*R*-squared"). An R^2 of 0 means that none of the variance in the data is in the model; all of it is still in the residuals.
- When interpreting a regression model you need to *Tell* what R^2 means.
- In the BK example, 69% of the variation in total *fat* is accounted for by the model.

How Big Should R^2 Be?

- 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.
- The standard deviation of the residuals can give us more information about the usefulness of the regression by telling us how much scatter there is around the line.

Reporting R^2

- Along with the slope and intercept 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.

Assumptions and Conditions

- Quantitative Variables Condition:
  - Regression can only be done on two quantitative variables, so make sure to check this condition.
- Straight Enough Condition:
  - The linear model assumes that the relationship between the variables is linear.
  - A scatterplot will let you check that the assumption is reasonable.
- It's a good idea to check linearity again *after* computing the regression when we can examine the residuals.
- You should also check for outliers, which could change the regression.
- If the data seem to clump or cluster in the scatterplot, that could be a sign of trouble worth looking into further.
- 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:
  - Watch out for outliers.
  - Outlying points can dramatically change a regression model.
  - Outliers can even change the sign of the slope, misleading us about the underlying relationship between the variables.

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?

- Don't fit a straight line to a nonlinear relationship.
- 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.