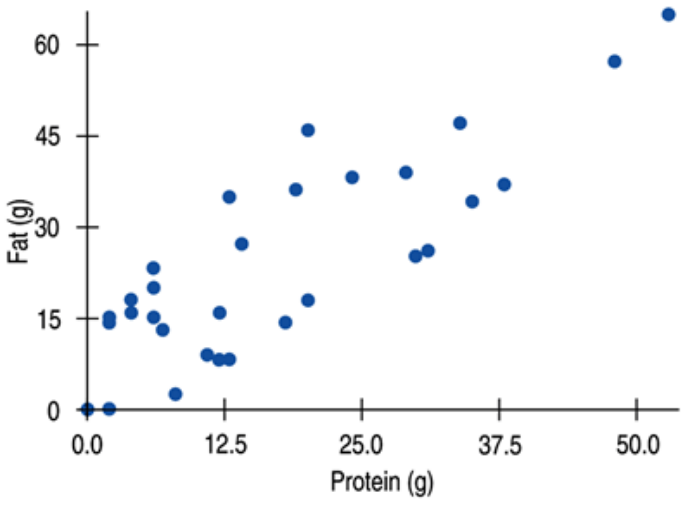
Chapter 8: Linear Regression

Obj - SWBAT understand how the correlation coefficient and the regression slope are related, as well as know how R^2 describes how much of the variation in y is accounted for by its linear relationship with x.

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. It 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 (cont.)

- 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.

Residuals

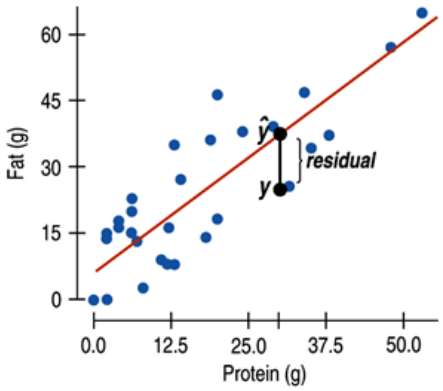
- The model won't be perfect, regardless of the line we draw.
- Some points will be above the line and some will be below.
- The estimate made from a model is the predicted value (denoted as $\hat{y}$).

Residuals (cont.)

The difference between the observed value and its associated predicted value is called the residual.

Residuals (cont.)

- 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.

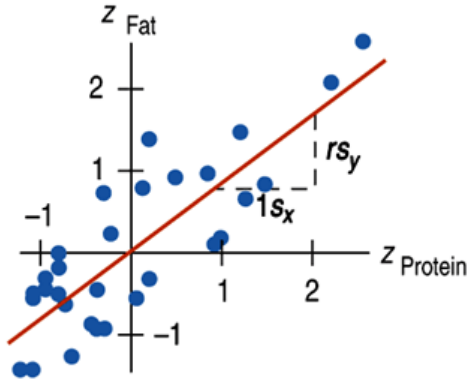


"Best Fit" Means Least Squares

- 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.

Correlation and the Line

- The figure shows the scatterplot of z-scores for fat and protein.
- If a burger has average protein content, it should have about average fat content too.
- Moving one standard deviation away from the mean in x moves us r standard deviations away from the mean in y.



Correlation and the Line (cont.)

- 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.

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:

- 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 and b_0 for the slope and 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.

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:

The Regression Line in Real Units (cont.)

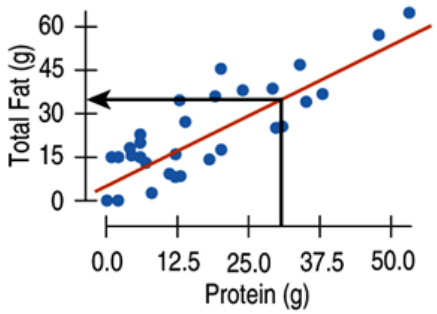
- In our model, we also have an intercept (b_0).
- The intercept is built from the means and the slope:

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.

The Regression Line in Real Units (cont.)

- 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

Residuals Revisited

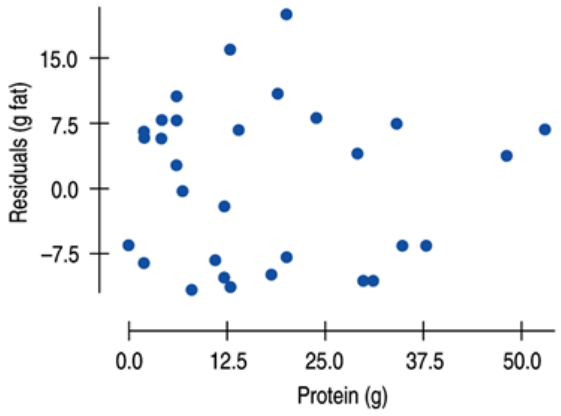
- The linear model assumes that the relationship between the two variables is a perfect straight line. The residuals are the part of the data that hasn't been modeled.
- Data = Model + Residual
- or (equivalently)
- Residual = Data – Model
- Or, in symbols, $e = y - \hat{y}$

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.

Residuals Revisited (cont.)

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



The Residual Standard Deviation

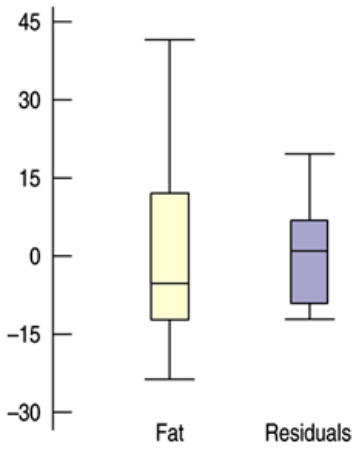
- The standard deviation of the residuals, s_e, measures how much the points spread around the regression line.
- Check to make sure the residual plot has about the same amount of scatter throughout. Check the Equal Variance Assumption with the Does the Plot Thicken? Condition.
- We estimate the SD of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

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.)

- 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.

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

- 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, $r^2 = (0.832)^2 = 0.69$, so 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 (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 variation in the protein content.

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 (and not two categorical 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.

Assumptions and Conditions (cont.)

- 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.

Assumptions and Conditions (cont.)

It's a good idea to check linearity again after computing the regression when we can examine the residuals.

- Does the Plot Thicken? Condition (Equal Variance Assumption):
  - > 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.

Assumptions and Conditions (cont.)

- 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.
  - > If the data seem to clump or cluster in the scatterplot, that could be a sign of trouble worth looking into further.

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.

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.

What have we learned? (cont.)

- 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).
- R2 gives us the fraction of the response accounted for by the regression model.

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.

What have we learned? (cont.)

- The linear model makes no sense unless the Linear Relationship Assumption is satisfied.
- Also, we need to check the Straight Enough Condition and Outlier Condition with a scatterplot.
- For the standard deviation of the residuals, we must make the Equal Variance Assumption. We check it by looking at both the original scatterplot and the residual plot for Does the Plot Thicken? Condition.

Steps for using Linear Regression

1.) State the problem

2.) Identify the W's in the study

3.) Check the conditions for regression:
- > Quantitative Variables Condition
- > Outlier Condition
- > Straight Enough Condition
- > Does the plot thicken Condition

4.) Fit the straight line model of the form:
$\hat{y} = b_0 + b_1 x$

5.) Calculate R^2 and s_e

6.) Check the plot of the residuals

7.) State the conclusion