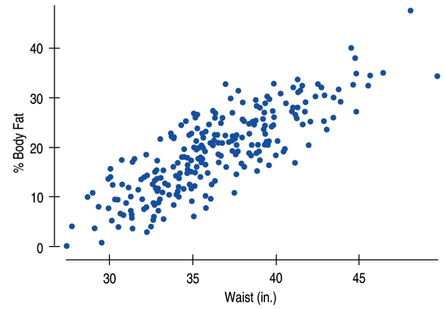
Chapter 27 Summary
Inferences for Regression

What have we learned?

- We have now applied inference to regression models.
- Like in all inference situations, there are conditions that we must check.
- We can test a hypothesis about the slope and find a confidence interval for the true slope.
- And, again, we are reminded never to mistake the presence of an association for proof of causation.

An Example: Body Fat and Waist Size

- Our chapter example revolves around the relationship between *% body fat* and *waist size* (in inches). Here is a scatterplot of our data set:

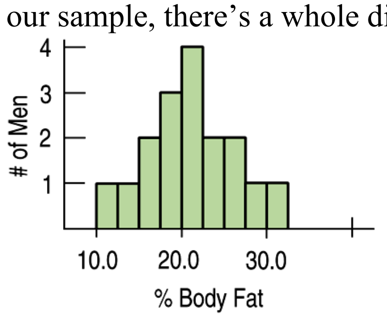


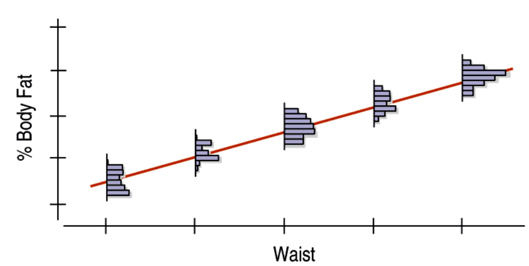
Remembering Regression

- In regression, we want to *model* the relationship between two quantitative variables, one the predictor and the other the response.
- To do that, we imagine an idealized regression line, which assumes that the means of the distributions of the response variable fall along the line even though individual values are scattered around it.
- Now we'd like to know what the regression model can tell us beyond the individuals in the study.
- We want to make confidence intervals and test hypotheses about the slope and intercept of the regression line.

The Population and the Sample

- When we found a confidence interval for a mean, we could imagine a single, true underlying value for the mean.
- When we tested whether two means or two proportions were equal, we imagined a true underlying difference.
- What does it mean to do inference for regression?
- We know better than to think that even if we know every population value, the data would line up perfectly on a straight line.
- In our sample, there's a whole distribution of *%body fat* for men with 38-inch waists:





- This is true at each waist size.
- We could depict the distribution of *%body fat* at different *waist* sizes (see above)

The Population and the Sample (cont.)

- The model assumes that the *means* of the distributions of *%body fat* for each *waist* size fall along the line even though the individuals are scattered around it.
- The model is not a perfect description of how the variables are associated, but it may be useful.
- If we had all the values in the population, we could find the slope and intercept of the *idealized regression line* explicitly by using least squares.
- We write the idealized line with Greek letters and consider the coefficients to be *parameters:* β_0 is the intercept and β_1 is the slope.
- Corresponding to our fitted line of $\hat{y}=b_0+b_1x$, we write $\mu_y=\beta_0+\beta_1x$
- Now, not all the individual *y*'s are at these means—some lie above the line and some below. Like all models, there are errors.
- Denote the errors by ε and write $\varepsilon = y - \mu_y$ for each data point (x, y).
- When we add error to the model, we can talk about individual *y*'s instead of means:

$$y=\beta_0+\beta_1x+\varepsilon$$

- This equation is now true for each data point (since the individual ε's soak up the deviations) and gives a value of *y* for each *x*.

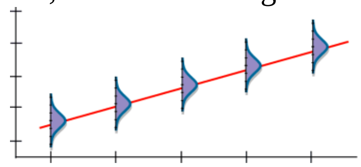
Assumptions and Conditions

- In Chapter 8 when we fit lines to data, we needed to check only the Straight Enough Condition.
- Now, when we want to make inferences about the coefficients of the line, we'll have to make more assumptions (and thus check more conditions).
- We need to be careful about the order in which we check conditions. If an initial assumption is not true, it makes no sense to check the later ones.

1. Linearity Assumption:
    - Straight Enough Condition: Check the scatterplot—the shape must be linear or we can't use regression at all.
    - If the scatterplot is straight enough, we can go on to some assumptions about the errors. If not, stop here, or consider re-expressing the data to make the scatterplot more nearly linear.
2. Independence Assumption:
    - Randomization Condition: the individuals are a representative sample from the population.
    - Check the residual plot (part 1)—the residuals should appear to be randomly scattered.
3. Equal Variance Assumption:
    - Does The Plot Thicken? Condition: Check the residual plot (part 2)—the spread of the residuals should be uniform.
4. Normal Population Assumption:
    - Nearly Normal Condition: Check a histogram of the residuals. The distribution of the residuals should be unimodal and symmetric.

Assumptions and Conditions (cont.)

- If all four assumptions are true, the idealized regression model would look like this:



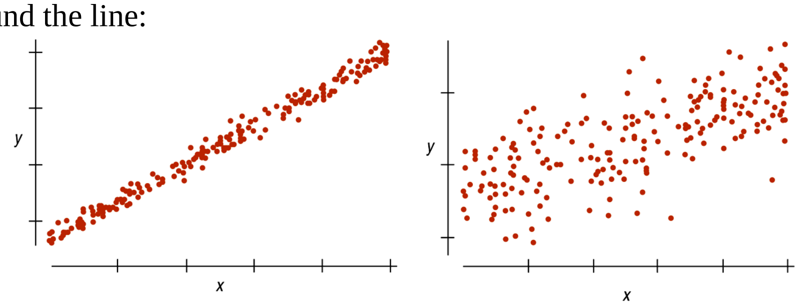
- At each value of x there is a distribution of y-values that follows a Normal model, and each of these Normal models is centered on the line and has the same standard deviation.

Which Come First: the Conditions or the Residuals?

- There's a catch in regression—the best way to check many of the conditions is with the residuals, but we get the residuals only *after* we compute the regression model.
- To compute the regression model, however, we should check the conditions.
- So we work in this order:
    - Make a scatterplot of the data to check the Straight Enough Condition. (If the relationship isn't straight, try re-expressing the data. Or stop.)
    - If the data are straight enough, fit a regression model and find the residuals, e, and predicted values, .
    - Make a scatterplot of the residuals against x or the predicted values.
        - This plot should have no pattern. Check in particular for any bend, any thickening, or any outliers.
    - If the data are measured over time, plot the residuals against time to check for evidence of patterns that might suggest they are not independent.
    - If the scatterplots look OK, then make a histogram and Normal probability plot of the residuals to check the Nearly Normal Condition.
    - If all the conditions seem to be satisfied, go ahead with inference.

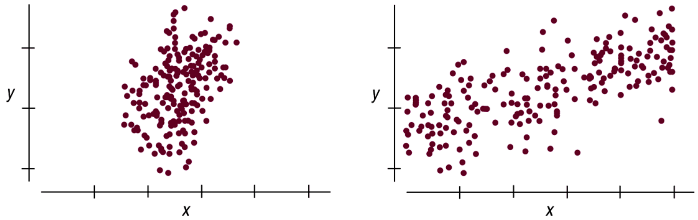
Intuition About Regression Inference

- We expect any sample to produce a b_1 whose expected value is the true slope, β_1.
- What about its standard deviation?
- What aspects of the data affect how much the slope and intercept vary from sample to sample?
    - Spread around the line:
        - Less scatter around the line means the slope will be more consistent from sample to sample.
        - The spread around the line is measured with the residual standard deviation s_e.
        - You can always find s_e in the regression output, often just labeled s.
- Spread around the line:



Intuition About Regression Inference (cont.)

- Spread of the x's: A large standard deviation of x provides a more stable regression.



- Sample size: Having a larger sample size, n, gives more consistent estimates.

Standard Error for the Slope

- Three aspects of the scatterplot affect the standard error of the regression slope:
  - spread around the line, s_e
  - spread of x values, s_x
  - sample size, n.
- The formula for the standard error (which you will probably never have to calculate by hand) is: $SE(b_1) = \frac{s_e}{\sqrt{n-1}\, s_x}$

Sampling Distribution for Regression Slopes

- When the conditions are met, the standardized estimated regression slope $t = \frac{b_1 - \beta_1}{SE(b_1)}$ follows a Student's t-model with $n - 2$ degrees of freedom.
- We estimate the standard error with $SE(b_1) = \frac{s_e}{\sqrt{n-1}\, s_x}$
- where:
  - $s_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}}$
  - n is the number of data values
  - s_x is the ordinary standard deviation of the x-values.

What About the Intercept?

- The same reasoning applies for the intercept.
- We can write $\frac{b_0 - \beta_0}{SE(b_0)} \sim t_{n-2}$ but we rarely use this fact for anything.
- The intercept usually isn't interesting. Most hypothesis tests and confidence intervals for regression are about the slope.

Regression Inference

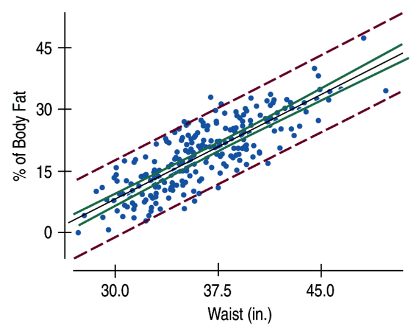
- A null hypothesis of a zero slope questions the entire claim of a linear relationship between the two variables—often just what we want to know.
- To test H_0: $\beta_1 = 0$, we find $t = \frac{b_1 - 0}{SE(b_1)}$ and continue as we would with any other t-test.
- The formula for a confidence interval for β_1 is $b_1 \pm t^*_{n-2} \times SE(b_1)$

Standard Errors for Predicted Values

- Once we have a useful regression, how can we indulge our natural desire to predict, without being irresponsible?
- Now we have standard errors—we can use those to construct a confidence interval for the predictions, smudging the results in the right way to report our uncertainty honestly.
- For our *%body fat* and *waist* size example, there are two questions we could ask:
  - Do we want to know the mean *%body fat* for *all* men with a *waist* size of, say, 38 inches?
  - Do we want to estimate the *%body fat* for a particular man with a 38-inch *waist*?
- The predicted *%body fat* is the same in both questions, but we can predict the *mean %body fat* for *all* men whose *waist* size is 38 inches with a lot more precision than we can predict the *%body fat* of a *particular individual* whose *waist* size happens to be 38 inches.
- We start with the same prediction in both cases.
  - We are predicting for a new individual, one that was not in the original data set.
  - Call his *x*-value x_ν.
  - The regression predicts *%body fat* as $\hat{y}_\nu = b_0 + b_1 x_\nu$
- Both intervals take the form $\hat{y}_\nu \pm t^*_{n-2} \times SE$
- The *SE*'s will be different for the two questions we have posed.
- The standard error of the *mean* predicted value is: $SE(\hat{\mu}_\nu) = \sqrt{SE^2(b_1) \times (x_\nu - \bar{x})^2 + \frac{s_e^2}{n}}$
- Individuals vary more than means, so the standard error for a single predicted value is larger than the standard error for the mean: $SE(\hat{y}_\nu) = \sqrt{SE^2(b_1) \times (x_\nu - \bar{x})^2 + \frac{s_e^2}{n} + s_e^2}$

Confidence Intervals for Predicted Values

- Here's a look at the difference between predicting for a mean and predicting for an individual.
- The solid green lines near the regression line show the 95% confidence interval for the mean predicted value, and the dashed red lines show the prediction intervals for individuals.



What Can Go Wrong?

- Don't fit a linear regression to data that aren't straight.
- Watch out for the plot thickening.
  - If the spread in *y* changes with *x*, our predictions will be very good for some *x*-values and very bad for others.
- Make sure the errors are Normal.
  - Check the histogram and Normal probability plot of the residuals to see if this assumption looks reasonable.
- Watch out for extrapolation.
  - It's always dangerous to predict for *x*-values that lie far from the center of data.
- Watch out for high-influence points and outliers.
- Watch out for one-tailed tests.
  - Tests of hypotheses about regression coefficients are usually two-tailed, so software packages report two-tailed P-values.
  - If you are using software to conduct a one-tailed test about slope, you'll need to divide the reported P-value in half.