

Chapter 29 Summary

Multiple Regression

What have we learned?

- There are many similarities between simple and multiple regression:
 - We fit the model by least squares.
 - The assumptions and conditions are essentially the same.
 - R^2 still gives us the fraction of the total variation in y accounted for by the model.
 - We perform inference on the coefficients by looking at the t -values, created from the ratio of the coefficients to their standard errors.
- There are some profound differences in interpretation when adding more predictors:
 - The coefficient of each x indicates the average change in y we'd expect to see for a unit change in that x for particular values of all the other x -variables.
 - The coefficient of a predictor variable can change sign when another variable is entered or dropped from the model.
 - Finding a suitable model from among the possibly hundreds of potential models is not straightforward.

Just Do It

- The method of least squares can be expanded to include more than one predictor. The method is known as multiple regression.
- For simple regression we found the Least Squares solution, the one whose coefficients made the sum of the squared residuals as small as possible.
- For multiple regression, we'll do the same thing, but this time with more coefficients.
- You should recognize most of the numbers in the following example (*%body fat*) of a multiple regression table:

Dependent variable is: Pct BF

R-squared = 71.3% R-squared (adjusted) = 71.1%

$s = 4.460$ with $250 - 3 = 247$ degrees of freedom

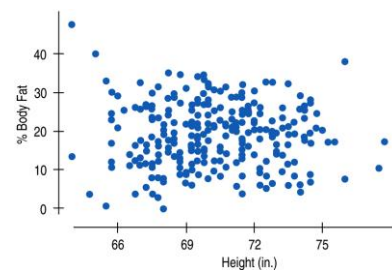
Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	-3.10088	7.686	-0.403	0.6870
Waist	1.77309	0.0716	24.8	≤ 0.0001
Height	-0.60154	0.1099	-5.47	≤ 0.0001

So What's New?

- The *meaning* of the coefficients in the regression model has changed in a subtle but important way.
- Multiple regression is an extraordinarily versatile calculation, underlying many widely used Statistics methods.
- Multiple regression offers our first glimpse into statistical methods that use more than two quantitative variables.

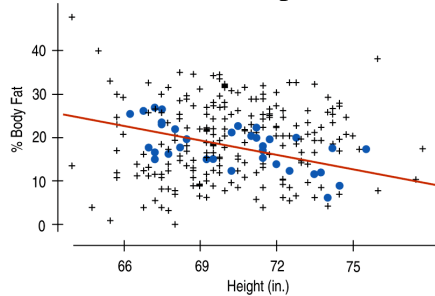
What Multiple Regression Coefficients Mean

- We said that height might be important in predicting body fat in men.
- What's the relationship between *%body fat* and *height* in men? Here's the scatterplot:

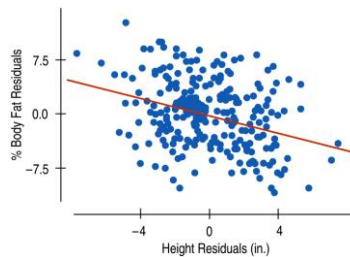


What Multiple Regression Coefficients Mean (cont.)

- It doesn't look like *height* tells us much about *%body fat*. Or does it?
- The coefficient of *height* in the multiple regression model was statistically significant, so it *did* contribute to the *multiple* regression model.
- How can this be?
 - The multiple regression coefficient of *height* takes account of the other predictor (*waist size*) in the regression model.
- For example, when we restrict our attention to men with waist sizes between 36 and 38 inches (points in blue), we can see a relationship between *%body fat* and *height*:



- So, overall there's little relationship between *%body fat* and *height*, but when we focus on *particular* waist sizes there is a relationship.
 - This relationship is conditional because we've restricted our set to only those men with a certain range of waist sizes.
 - For men with that waist size, an extra inch of height is associated with a decrease of about 0.60% in body fat.
 - If that relationship is consistent for each *waist* size, then the multiple regression coefficient will estimate it.
- The following is a partial regression plot shows the coefficient of *height* in the regression model has a slope equal to the coefficient value in the multiple regression model:



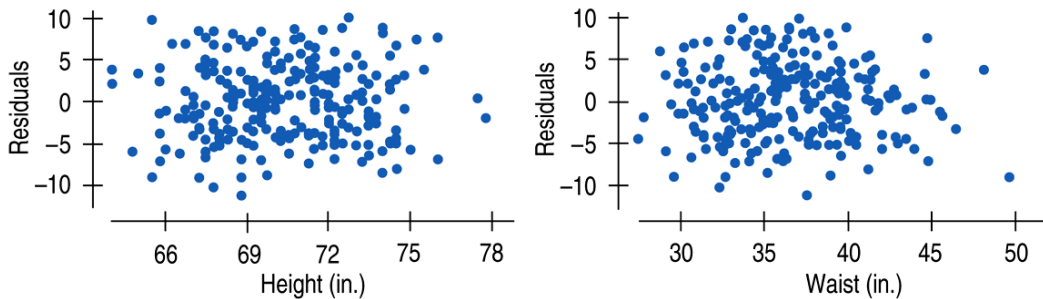
- For a multiple regression with k predictors, the model is: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon$
- The assumptions and conditions for the multiple regression model sound nearly the same as for simple regression, but with more variables in the model, we'll have to make a few changes.

Assumptions and Conditions

- Linearity Assumption:
 - Straight Enough Condition: Check the scatterplot for each candidate predictor variable—the shape must not be obviously curved or we can't consider that predictor in our multiple regression model.
- Independence Assumption:
 - Randomization Condition: Check the residuals plot (part 1)—the residuals should appear to be randomly scattered.

Assumptions and Conditions (cont.)

- Equal Variance Assumption:
 - Does the Plot Thicken? Condition: Check the residuals plot (part 2)—the spread of the residuals should be uniform.



- Normality Assumption:
 - Nearly Normal Condition: Check a histogram of the residuals—the distribution of the residuals should be unimodal and symmetric, and the Normal probability plot should be straight.
- Summary of the checks of conditions in order:
 1. Check the Straight Enough Condition.
 2. If the scatterplots are straight enough, fit a multiple regression model to the data.
 3. Find the residuals and predicted values.
 4. Make and check a scatterplot of the residuals against the predicted values.
 5. Think about how the data were collected.
 6. If the conditions check out this far, feel free to interpret the regression model and use it for prediction.
 7. If you wish to test hypotheses about the coefficients or about the overall regression, then make a histogram and Normal probability plot of the residuals to check the Nearly Normal Condition.

Multiple Regression Inference I: I Thought I Saw an ANOVA Table...

- Now that we have more than one predictor, there's an overall test we should consider before we do more inference on the coefficients.
- We ask the global question "Is this multiple regression model any good at all?"
- We test $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$
- The F -statistic and associated P-value from the ANOVA table are used to answer our question.

Multiple Regression Inference II: Testing the Coefficients

- Once we check the F -test and reject the null hypothesis, we can move on to checking the test statistics for the individual coefficients.
- For each coefficient, we test $H_0: \beta_j = 0$
- If the assumptions and conditions are met (including the Nearly Normal Condition), these ratios follow a Student's t -distribution.

$$t_{n-k-1} = \frac{b_j - 0}{SE(b_j)}$$
- We can also find a confidence interval for each coefficient:

Multiple Regression Inference II: Testing the Coefficients (cont.)

- Keep in mind that the meaning of a multiple regression coefficient depends on all the *other* predictors in the multiple regression model.
 - If we fail to reject the null hypothesis for a multiple regression coefficient, it does **not** mean that the corresponding predictor variable has no linear relationship to y .
 - It means that the corresponding predictor contributes nothing to modeling y *after allowing for all the other predictors*.

How's That, Again?

- It *looks* like each coefficient in the multiple regression tells us the effect of its associated predictor on the response variable.
- But, that is not so.
- The coefficient of a predictor in a multiple regression depends as much on the *other* predictors as it does on the predictor itself.

Comparing Multiple Regression Models

- How do we know that some other choice of predictors might not provide a better model?
- What exactly *would* make an alternative model better?
- These questions are not easy—there's no simple measure of the success of a multiple regression model.
- Regression models should make sense.
 - Predictors that are easy to understand are usually better choices than obscure variables.
 - Similarly, if there is a known mechanism by which a predictor has an effect on the response variable, that predictor is usually a good choice for the regression model.
- The simple answer is that we can't know whether we have the best possible model.

Adjusted R^2

- There is another statistic in the full regression table called the adjusted R^2 .
- This statistic is a rough attempt to adjust for the simple fact that when we add another predictor to a multiple regression, the R^2 can't go down and will most likely get larger.
- This fact makes it difficult to compare alternative regression models that have different numbers of predictors.

- The formula for R^2 is $R^2 = \frac{SS_{Regression}}{SS_{Total}}$

while the formula for adjusted R^2 is $R^2_{adj} = \frac{MS_{Regression}}{MS_{Total}}$

- Because the mean squares are sums of squares divided by their degrees of freedom, they are adjusted for the number of predictors in the model.
 - As a result, the adjusted R^2 value won't necessarily increase when a new predictor is added to the multiple regression model.
 - That's fine, but adjusted R^2 no longer tells the fraction of variability accounted for by the model—it isn't even bounded by 0 and 100%.

Adjusted R^2 (cont.)

- Comparing alternative regression models is a challenge, especially when they have different numbers of predictors.
- Adjusted R^2 is one way to help you choose your model.
- But, don't use it as the sole decision criterion when you compare different regression models.

What Can Go Wrong?

- Interpreting Coefficients
 - Don't claim to "hold everything else constant" for a single individual.
 - Don't interpret regression causally.
 - Be cautious about interpreting a regression model as predictive.
 - Don't think that the sign of a coefficient is special.
 - If a coefficient's t -statistic is not significant, don't interpret it at all.
- Don't fit a linear regression to data that aren't straight.
- Watch out for the plot thickening.
- Make sure the errors are nearly Normal.
- Watch out for high-influence points and outliers.