

Sliced Bread is the Best Thing Since Linear Regression

Ben Babcock
University of Minnesota

Regression Vocabulary

The criteria variable is the variable that we are trying to predict. We usually plot this on the y-axis of a scatterplot.

The predictor variable is the variable we are using to predict the criteria. We generally plot this on the x-axis.

The slope is the steepness of the line.

The intercept is the predicted value of the criteria when the predictor variable is 0.

We regress the criteria (Y) on the predictor (X).

A Major Truth and Some History

What would you think if I told you that correlation, the independent samples t -test, one-way ANOVA, and factorial ANOVA were all special cases of one single statistical model?

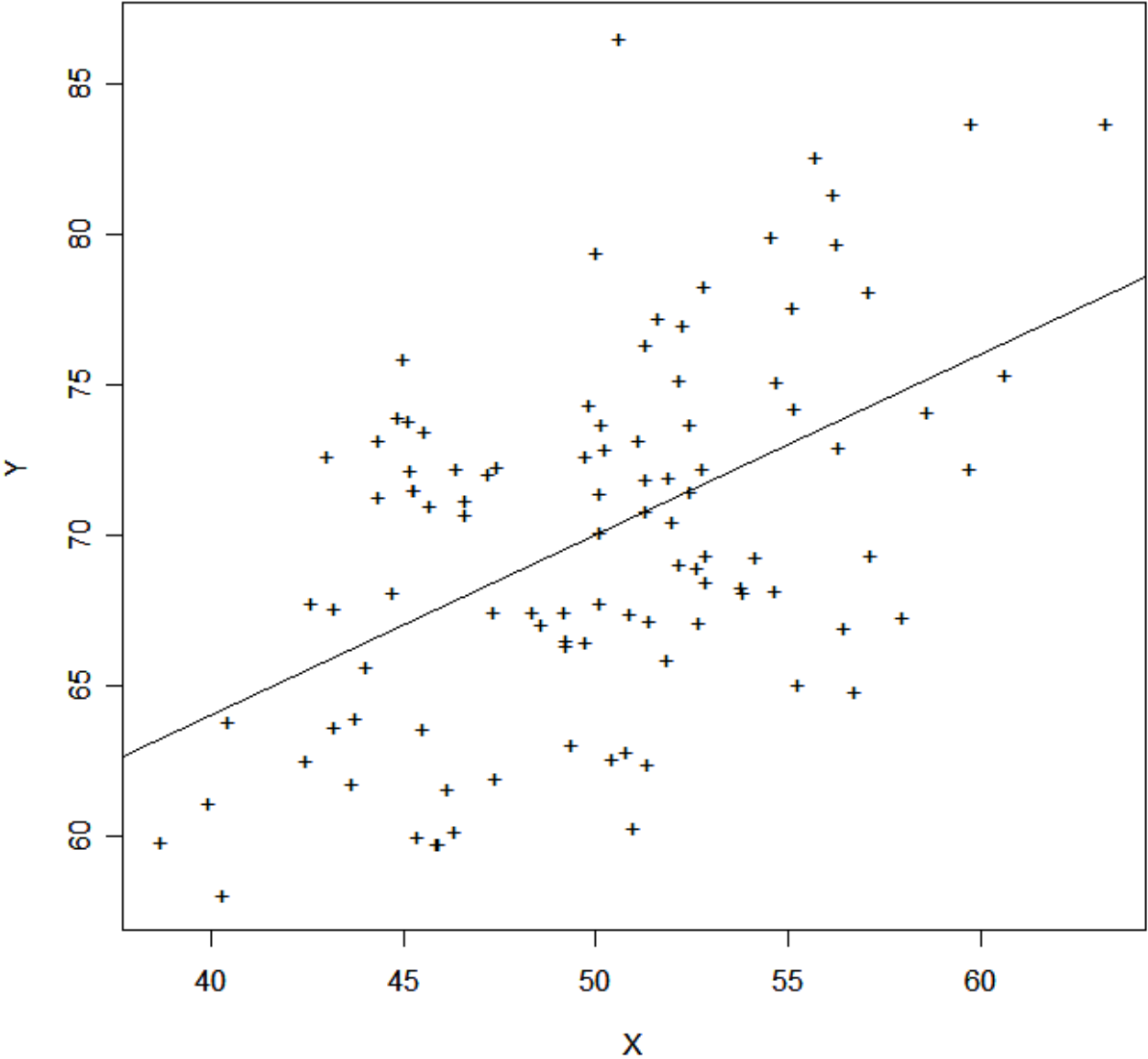
All of these things are special cases of linear regression.

Simple (One X, one Y variable) regression was “discovered” by numerous folks. One of the first to properly lay it out was _____ . Hint: we've talked about him before.

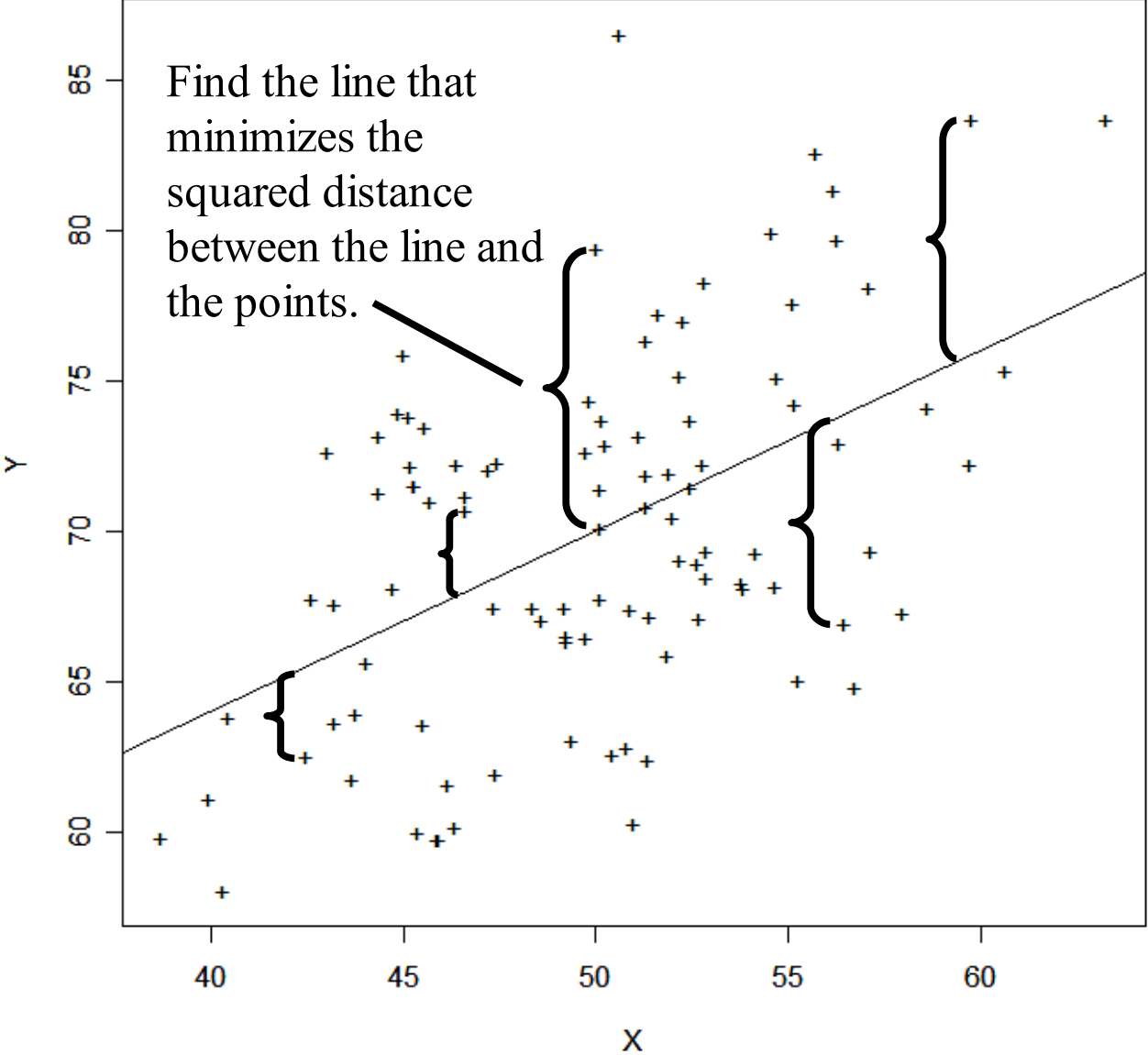
Regression involves finding a line of best fit through a group of points (scatterplot).

Best fit =

Scatterplot with Best Fit Line

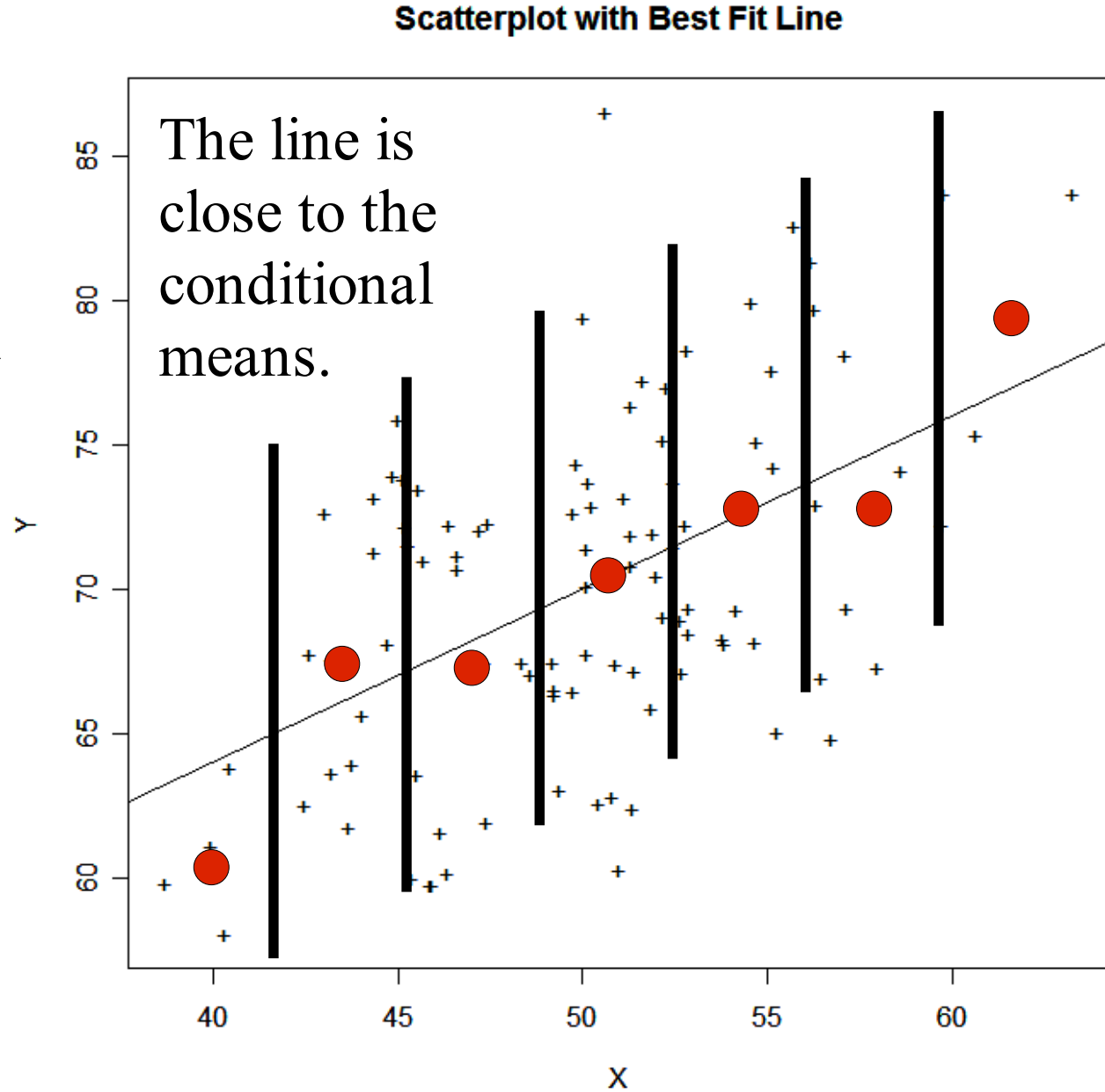


Scatterplot with Best Fit Line



Conditional mean: split up 1 variable (X) into smaller parts and find the mean on the second variable (Y).

It is the Y mean given that you are in a certain range (or just 1 value) of X .



Regression: Conceptual Review

We are going to find the best fit line through a group of points that minimizes the least _____ criteria.

Points on the line given the a score on the predictor (X) variable are _____ means.

According to our least squares criterion, how many lines can be the best?

Regression is useful for:

Regression: Equations

Basic model: $Y_p = a + bX_p + error_p$

where p is a person index

Y is the criterion variable

X is the predictor variable

a is the intercept

b is the slope

$error$ is an individual error term

Problem:

Prediction model: $\hat{Y}_p = a + bX_p$

where p is a person index

\hat{Y} is the predicted criterion value

X is a predictor value

a is the intercept

b is the slope

This is how to predict Y from an X value.

Regression: Slope and Intercept

When dealing with simple regression, we want to find the slope and the intercept.

Intercept: the predicted Y -value if the value if X is _____.

Sometimes the intercept is mathematically useful but practically meaningless. Examples:

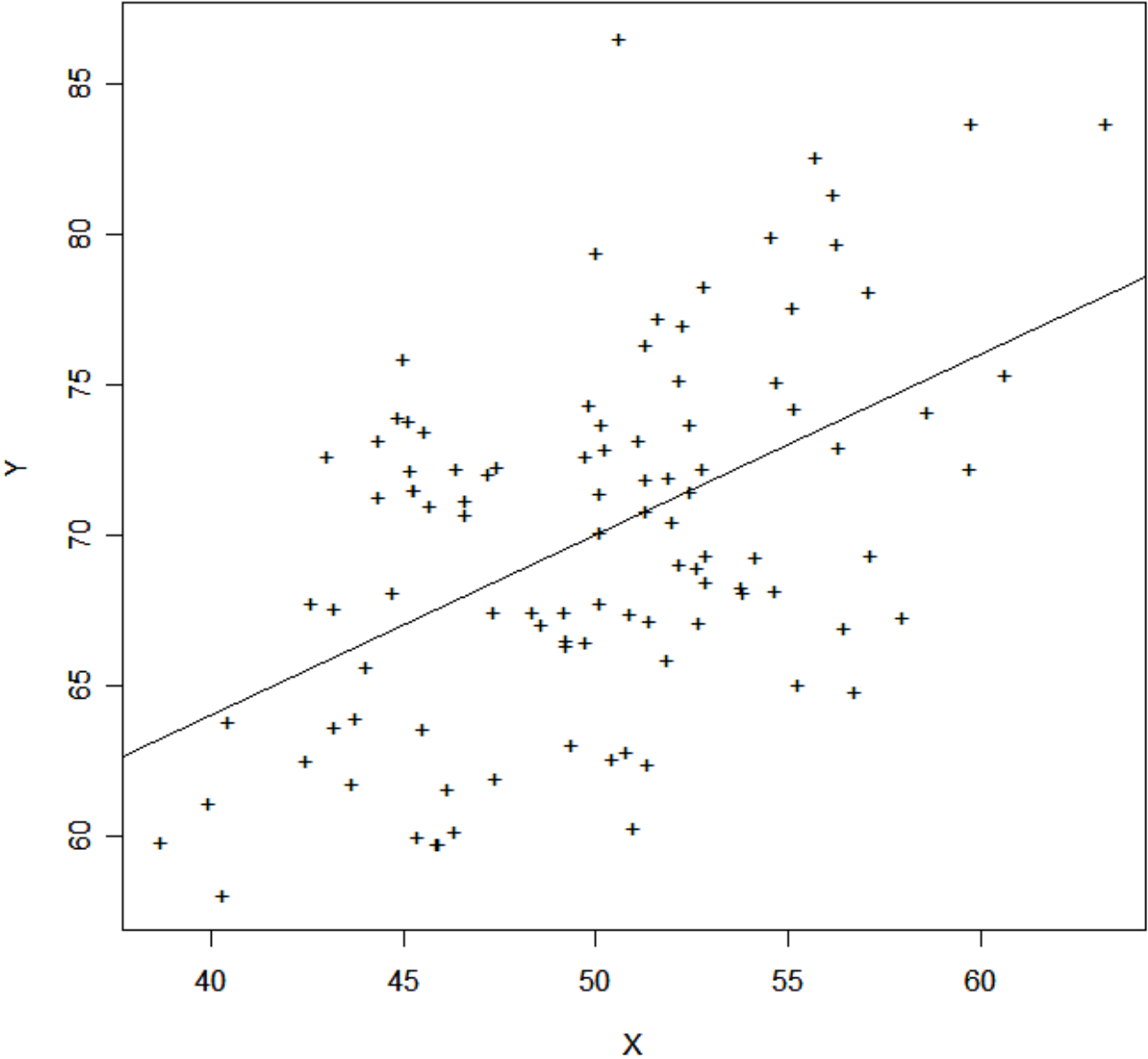
Slope: a value indicating how steep a line is.

Interpretation: Suppose you have a slope of 1.2.

If you have two people in your sample, one with an X score of 5 and another with an X of 6, we would expect that the second person's Y score would be 1.2 units higher on average.

Slopes can be negative, indicating an average decrease.

Scatterplot with Best Fit Line



Equations for Calculating Slope and Intercept

Finding the slope:
$$b = \frac{cov_{XY}}{s_X^2} = r_{XY} \frac{s_Y}{s_X}$$

where b is the slope

cov_{XY} is the covariance between X and Y

s_X is the standard deviation of X

s_Y is the standard deviation of Y

r_{XY} is the correlation between X and Y

Finding the intercept:
$$a = \bar{Y} - b \bar{X}$$

where a is the intercept

b is the slope

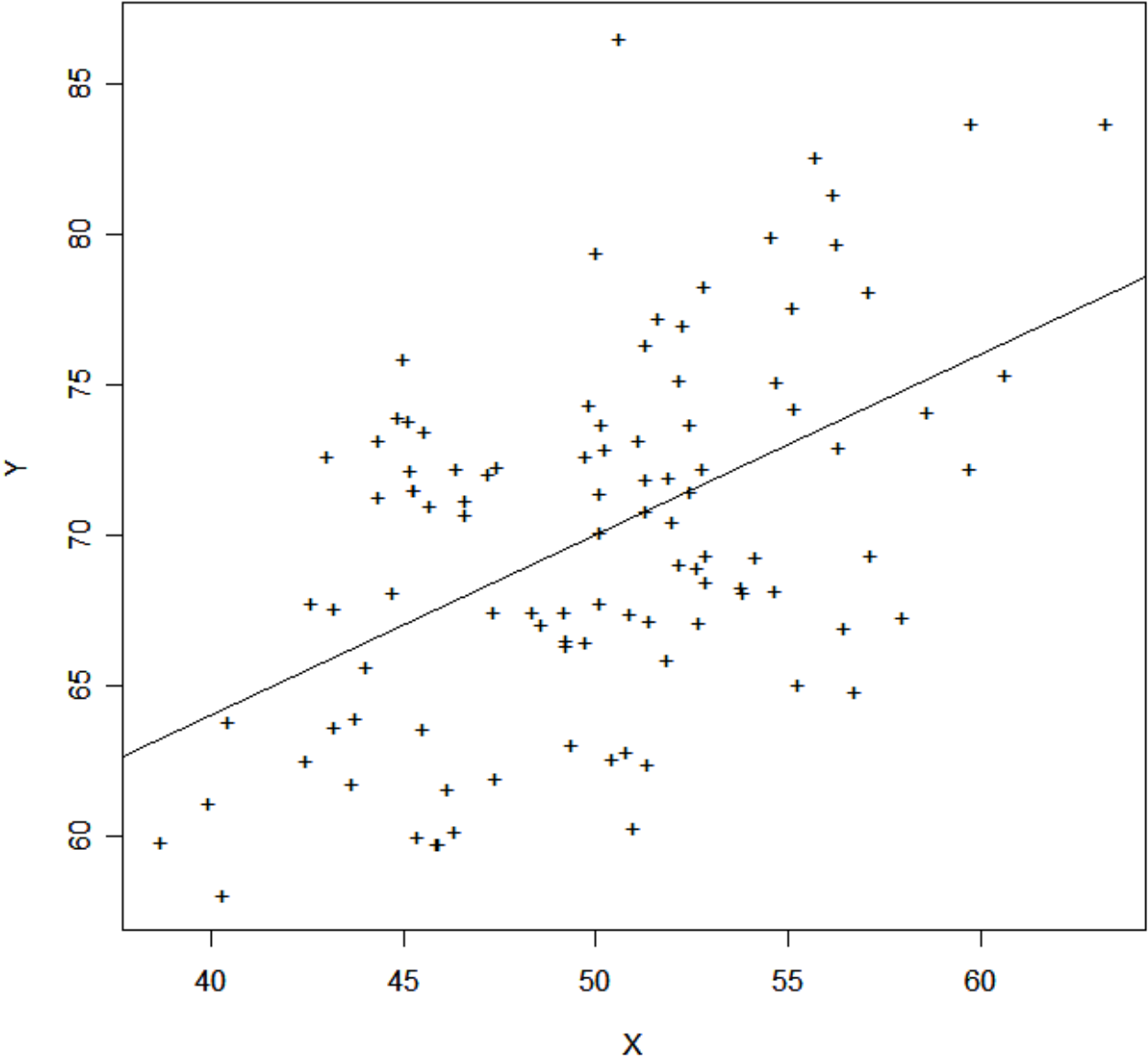
\bar{Y} is the mean of the criterion

\bar{X} is the mean of the predictor

Note that you must find b first.

Standard errors: H, 221. We'll let SPSS do that for us.

Scatterplot with Best Fit Line



Regression: Interpretation and Significance

Regression output from R from previous scatterplot. SPSS gives something similar. Suppose that X is a vocational performance test score and Y is pay in \$1,000.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	40.000	5.275	7.583	1.94e-11	***
data.5[, 1]	0.600	0.105	5.715	1.18e-07	***

Residual standard error: 5.223 on 98 degrees of freedom
Multiple R-Squared: 0.25, Adjusted R-squared: 0.2423
F-statistic: 32.67 on 1 and 98 DF, p-value: 1.180e-07

R-Squared:

F-statistic:

Significance of the coefficients:

Interpretation:

One prediction example:

Assumptions of the Model (Look Familiar?)

Multivariate normality:

Consequence:

Solution:

Independence of observations:

Consequence:

Solution:

Homoscedasticity:

Consequence:

Solution:

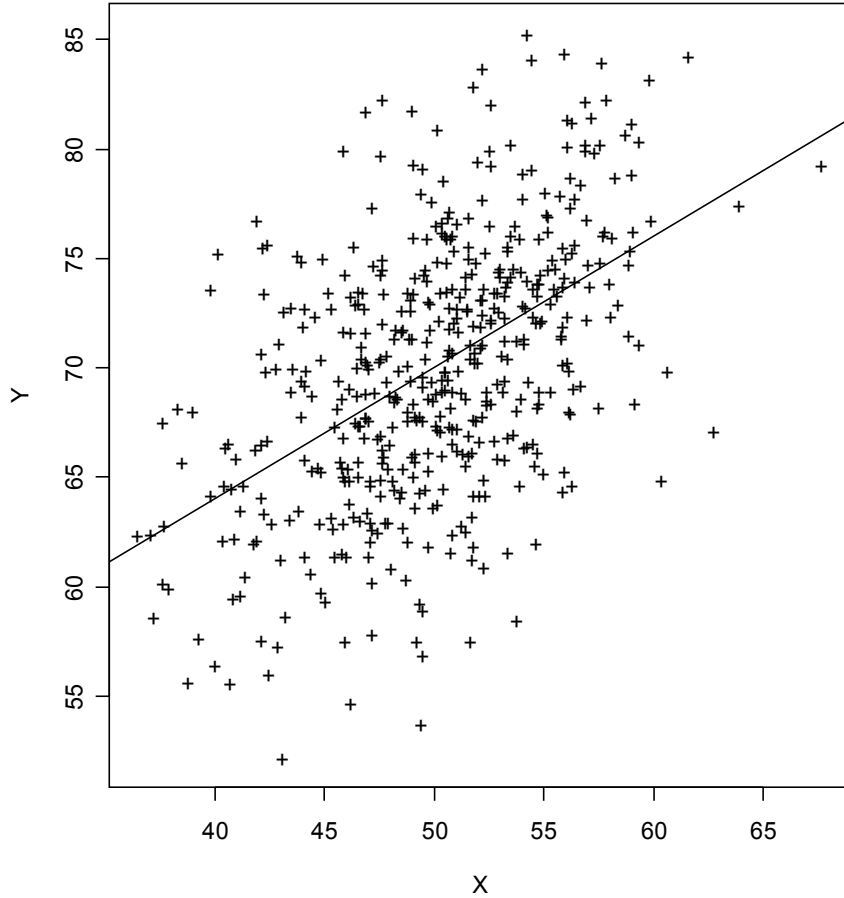
Linearity:

Consequence:

Solution:

What Multivariate Normality Looks Like

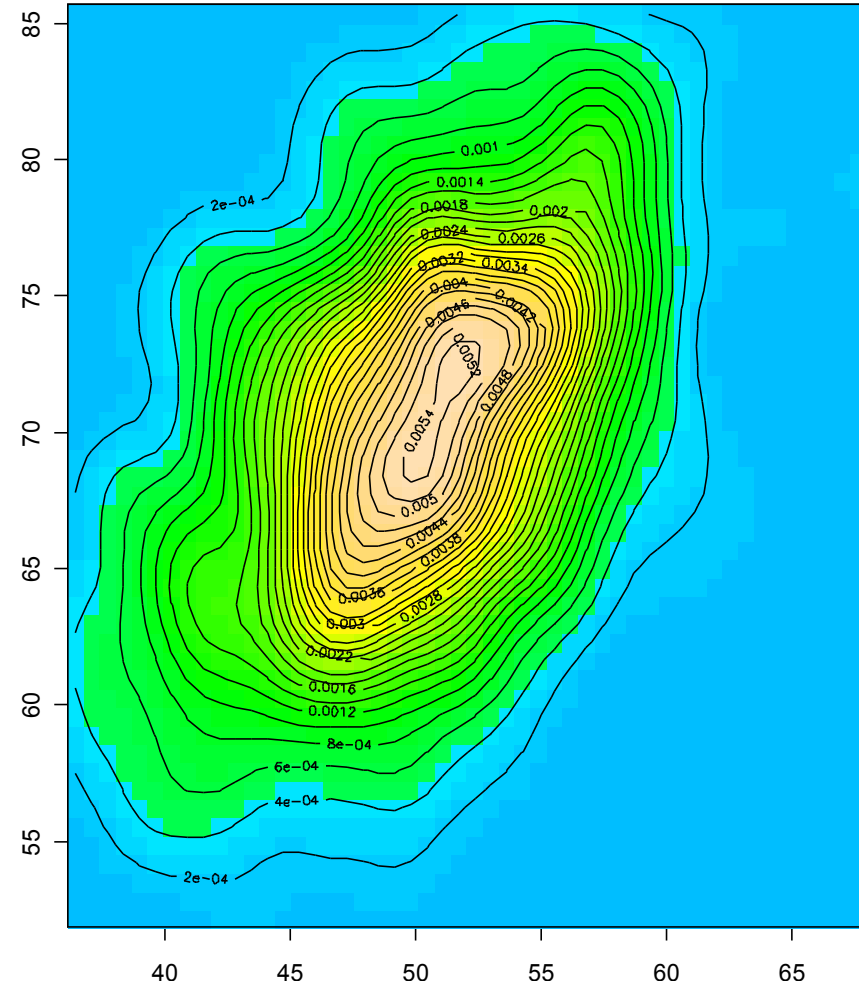
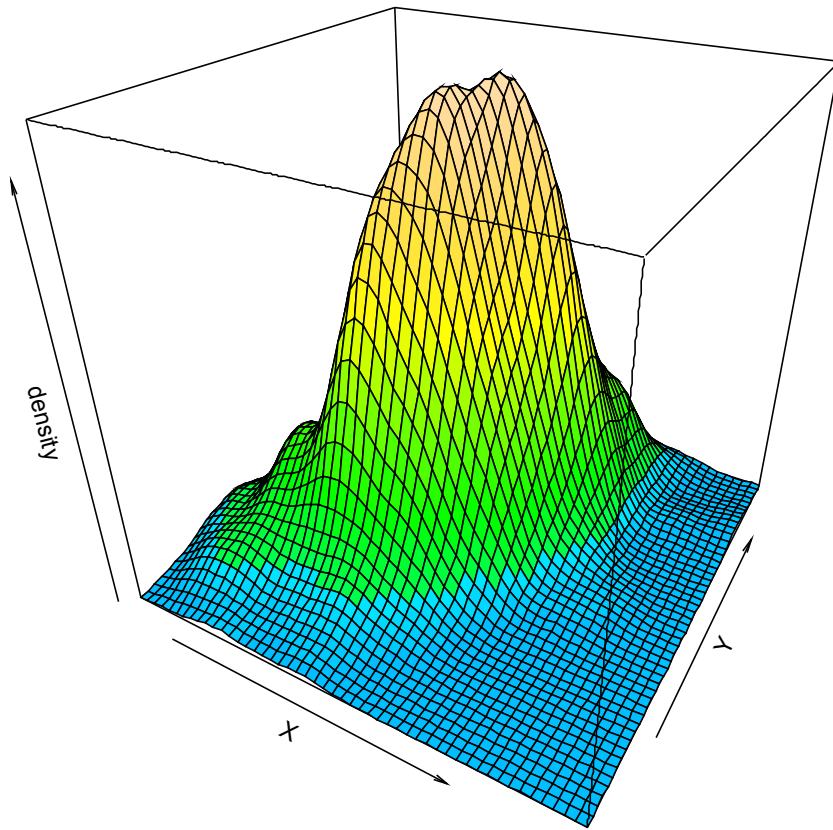
Scatterplot with Best Fit Line



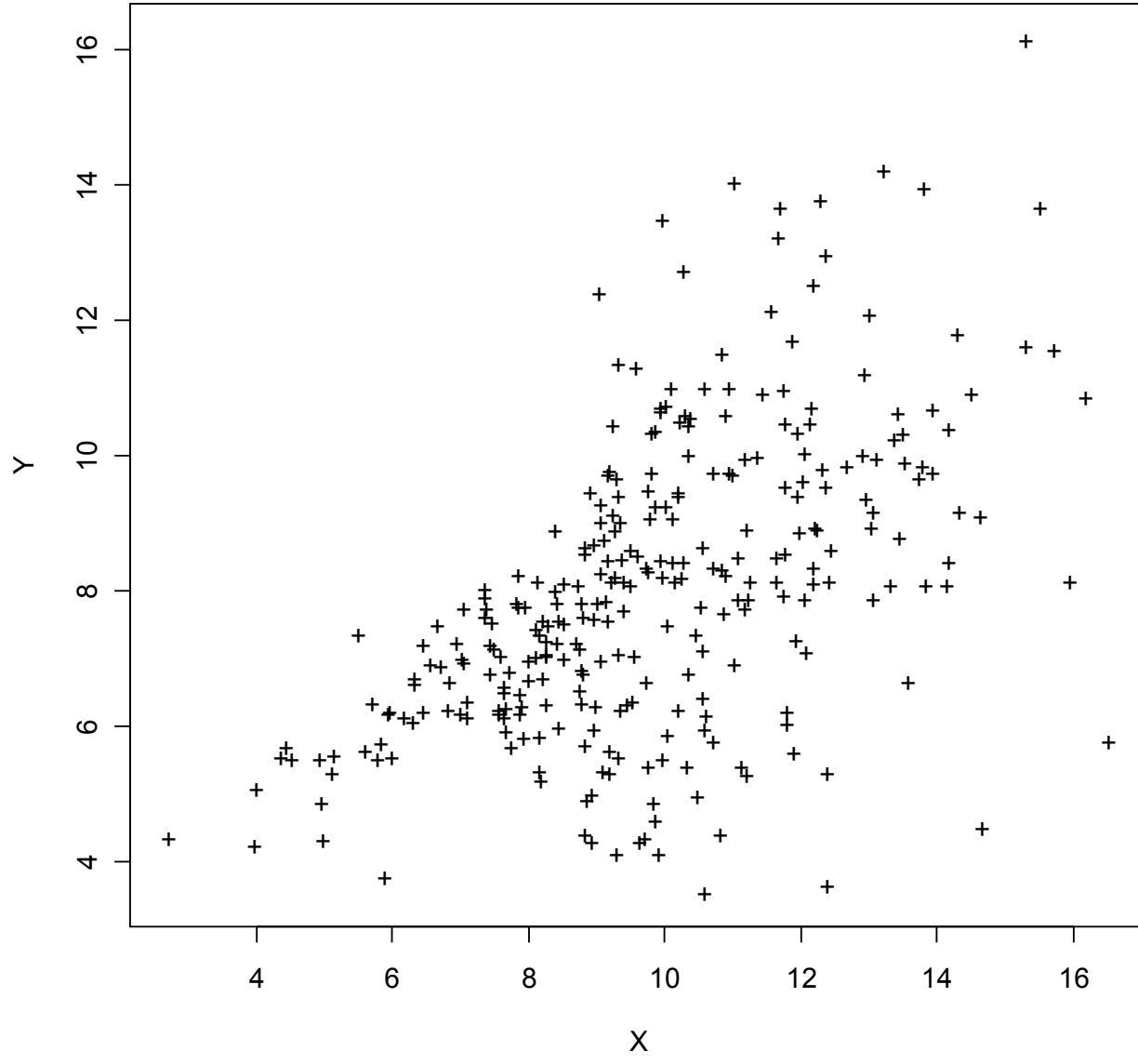
2 Dimensional Bivariate Density Plot
Darker Areas Indicate High Density



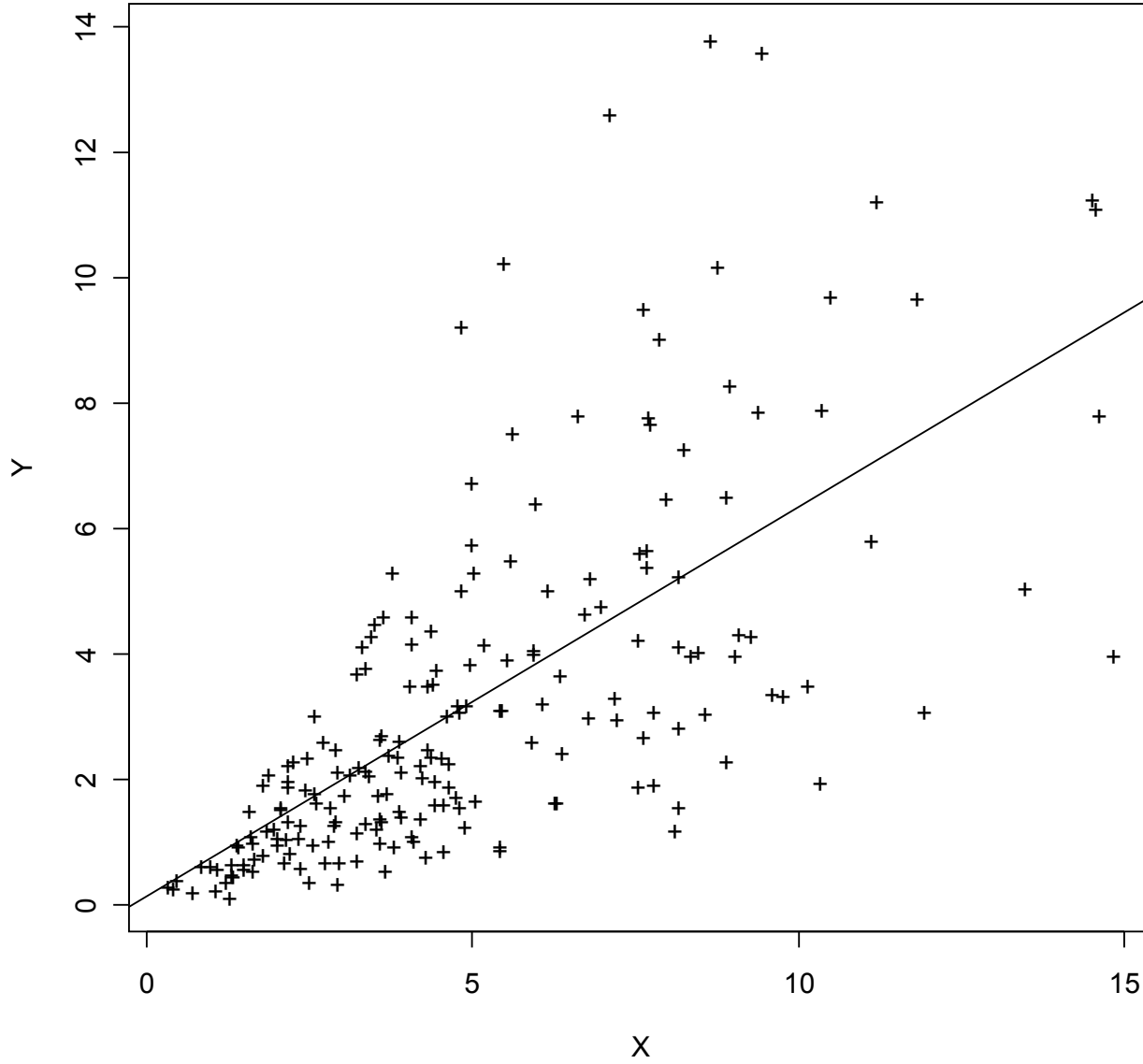
Another View of the Same Data



Scatterplot with Heteroscedasticity



Violation of Normality and Homoscedasticity



Problem Children

All of the “problem children” from correlation are true here.

Range Restriction

Heterogenous Subsamples

Outliers

Non-linear Effects

Add Heteroscedasticity to the list.

Correlation and Simple Regression

Correlation and simple regression are the same thing on different scales. Correlation is simply standardized simple regression.

If you take your raw scores, turn them into z -scores, and run a regression:

1. Intercept:

2. Slope:

3. Significance testing (p -values):

Regression: The Problem of Overfitting

The regression of Y on X fits everything (all characteristics of our data). This fitting includes bad stuff like _____ and _____. Your R^2 will be too large, and this can be a problem when we try to predict scores later on.

Solution: Cross Validation!

In cross validation, you take a very large sample (say 500 people). You develop the regression equation on 400 of the people and see how your coefficients work on the remaining 100 people. Your loss in R^2 is called _____.

Ways to minimize it: large sample size, reliable measures, representative sampling.

Multiple Regression

Multiple regression is like simple regression with more than one predictor.

Basic model: $Y = B_0 + B_1 X_1 + B_2 X_2 + \dots + B_k X_k + error$

Prediction model: $\hat{Y} = B_0 + B_1 X_1 + B_2 X_2 + \dots + B_k X_k$

The computation involved with this model goes beyond the scope of this class.

Take home concept: the 1 criterion and 1 predictor case can be generalized to using many predictors.