Stat 102

Introduction to Business Statistics

Class 6

Richard P. Waterman

Wharton

Table of contents I

1. Today's module
2. Last time
3. A motivating example
4. Lines
5. Fitting lines to data
   - Why fit?
   - The least squares criterion
6. Interpretation
7. Summary
8. Next time

Today's module

Topics to be covered in this module:

- Last time
- Global and local approaches to prediction
- Line definition
- The least squares fit to data
- The residuals
- Interpreting the slope and intercept of a regression
- Summary
- Next time

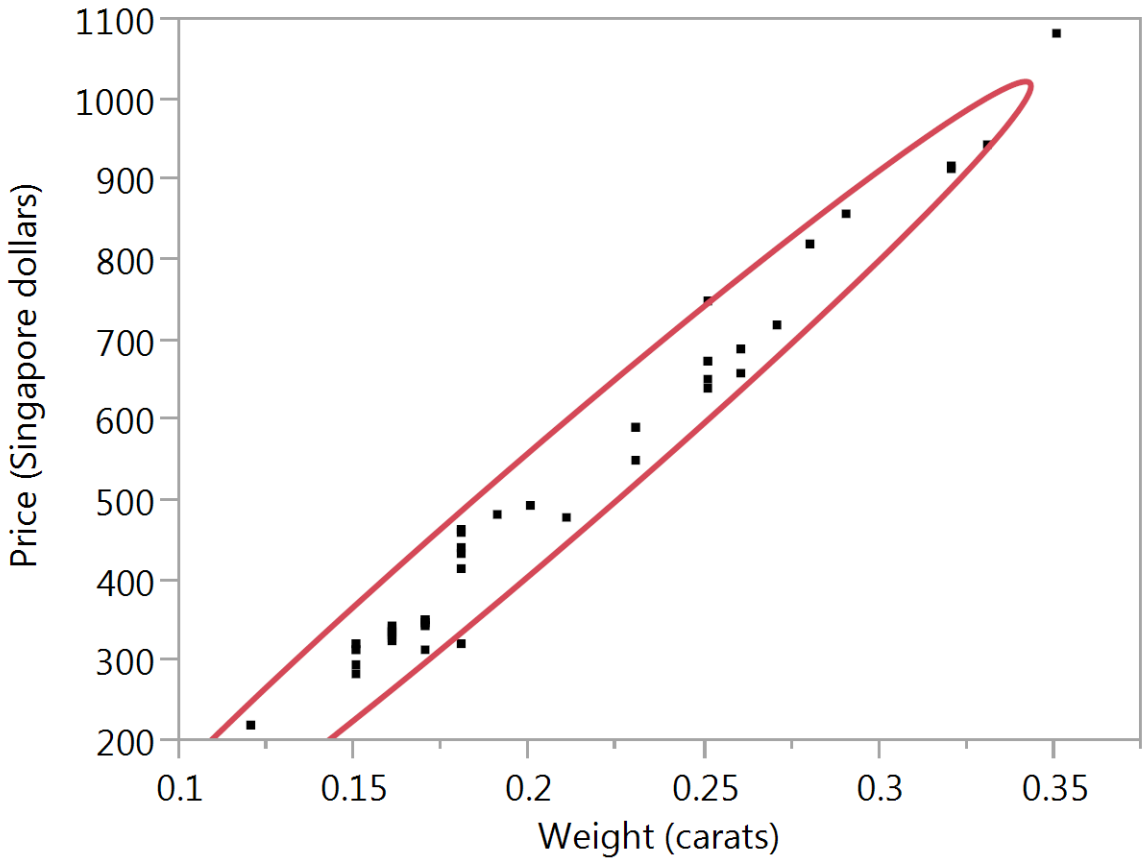
Last time

Main points:

- Association between categorical variables
- The χ^2 test for independence.
- The χ^2 goodness-of-fit test.

Motivating example

- Given the diamonds data set, what are reasonable approaches to estimating the price of a diamond that weighs 0.25 carats?
- What about a diamond that weighs 0.3 carats?



Two approaches

1. Be local: use only information in the neighborhood of the prediction.
2. Be global: assert a universal truth (like linearity) and exploit that assertion. As lines have constant slope, if we believe in linearity, then we can use all the data to estimate that common slope.

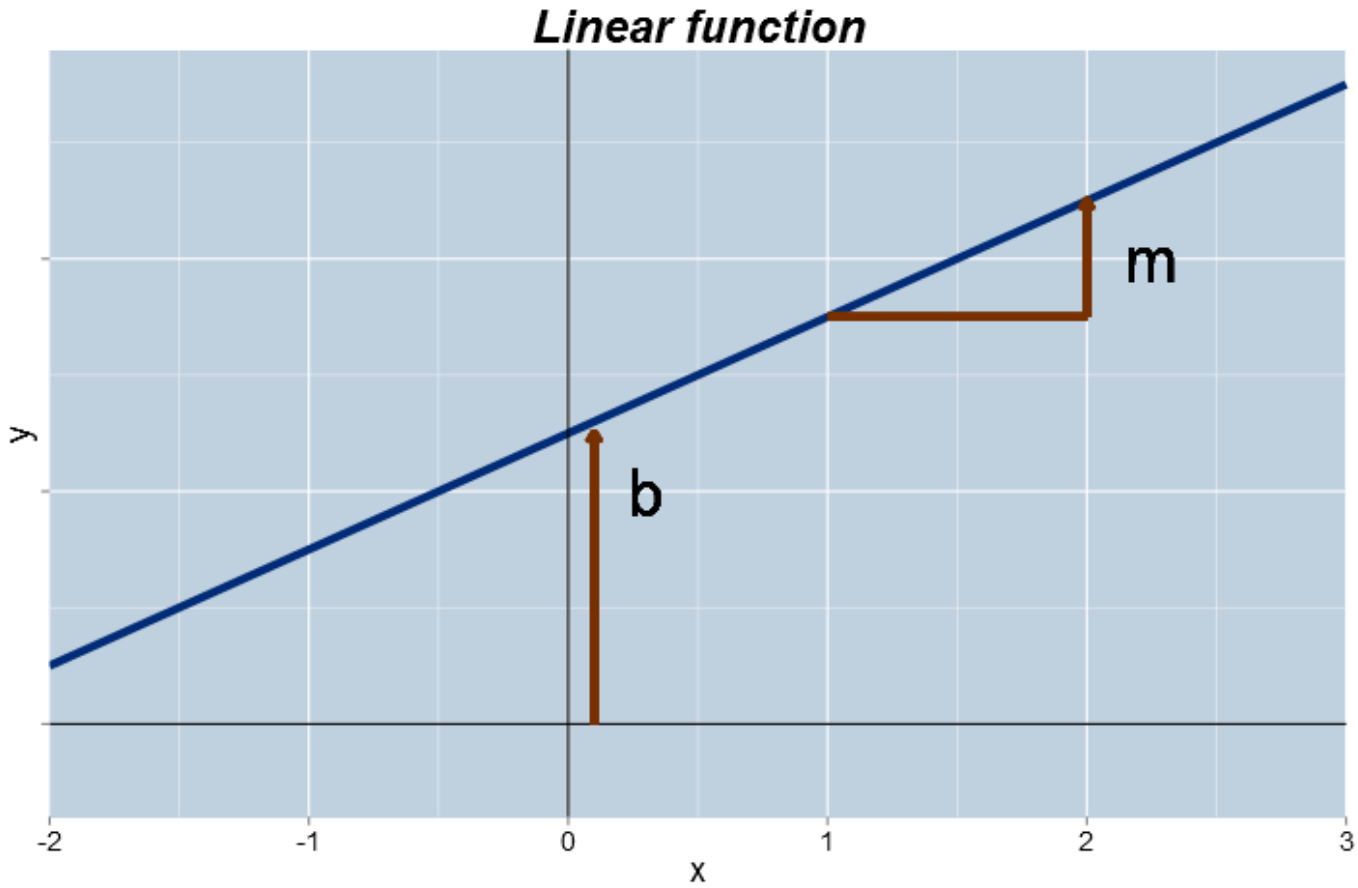
- Local approach:
  - Benefit: makes fewer assumptions.
  - Downsides: potentially leaves some information *on the table*. Doesn't work well in high dimensions as the neighborhoods are sparsely populated.
- Global approach:
  - Benefit: uses all of the data for more precision.
  - Downside: if the global assumption is wrong, then it's a fools paradise.

Stat 102 approach: be global, but think carefully and check the assumptions.

Lines from a math class

The usual definition:

$$y = mx + b.$$



Lines in the stat class

We will use the notation:

$$y = b_0 + b_1 x,$$

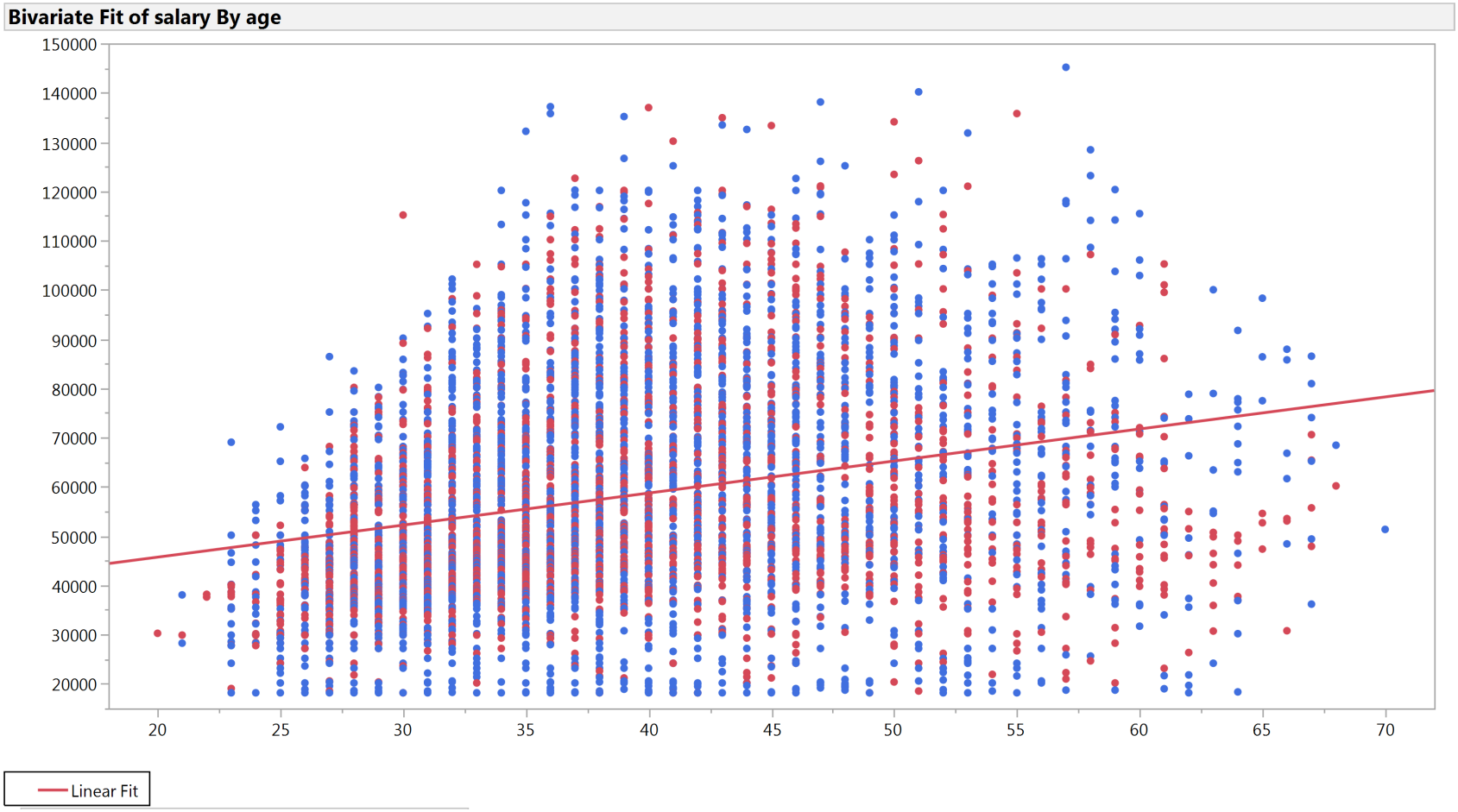
where

- y is the *response variable*
- x is the *predictor variable*
- b_0 is the intercept
- b_1 is the slope

The defining property of a line: the slope is constant.

Engaging the problem: does linearity make sense?

- What's the relationship between age and income?
- What does linearity imply?



Reasons to fit a line through data

Once we have an equation we can summarize and exploit fit:

- Graphically summarize.
- Interpolate.
- Forecast/extrapolate (with caution).
- Mathematically leverage the equation: derivatives and optimizations.

Least squares

The classical definition of the "best" line:

- Find the β_0 and β_1 that minimize

$$\sum_{i=1}^{n} \{y_i - (\beta_0 + \beta_1 x_i)\}^2.$$

  Call the minimizers b_0 and b_1.
- In English, the "best line" minimizes the sum of the squares of the vertical distances from the points to the line, and is called the *Least Squares Line*.
- Sometimes, we may fit a line on a transformed scale, then back-transform, which gives *best fitting* curves.

The least squares estimates of the slope and intercept

- Call x the **predictor** variable and y the **response** variable.
- The fitted values are written as $\hat{y}$, and using the line $\hat{y} = b_0 + b_1 x$.
- The difference between y and $\hat{y}$, $y - \hat{y}$ is called the **residual**.
- We often write the residual as e so that:

$$e = y - \hat{y}.$$

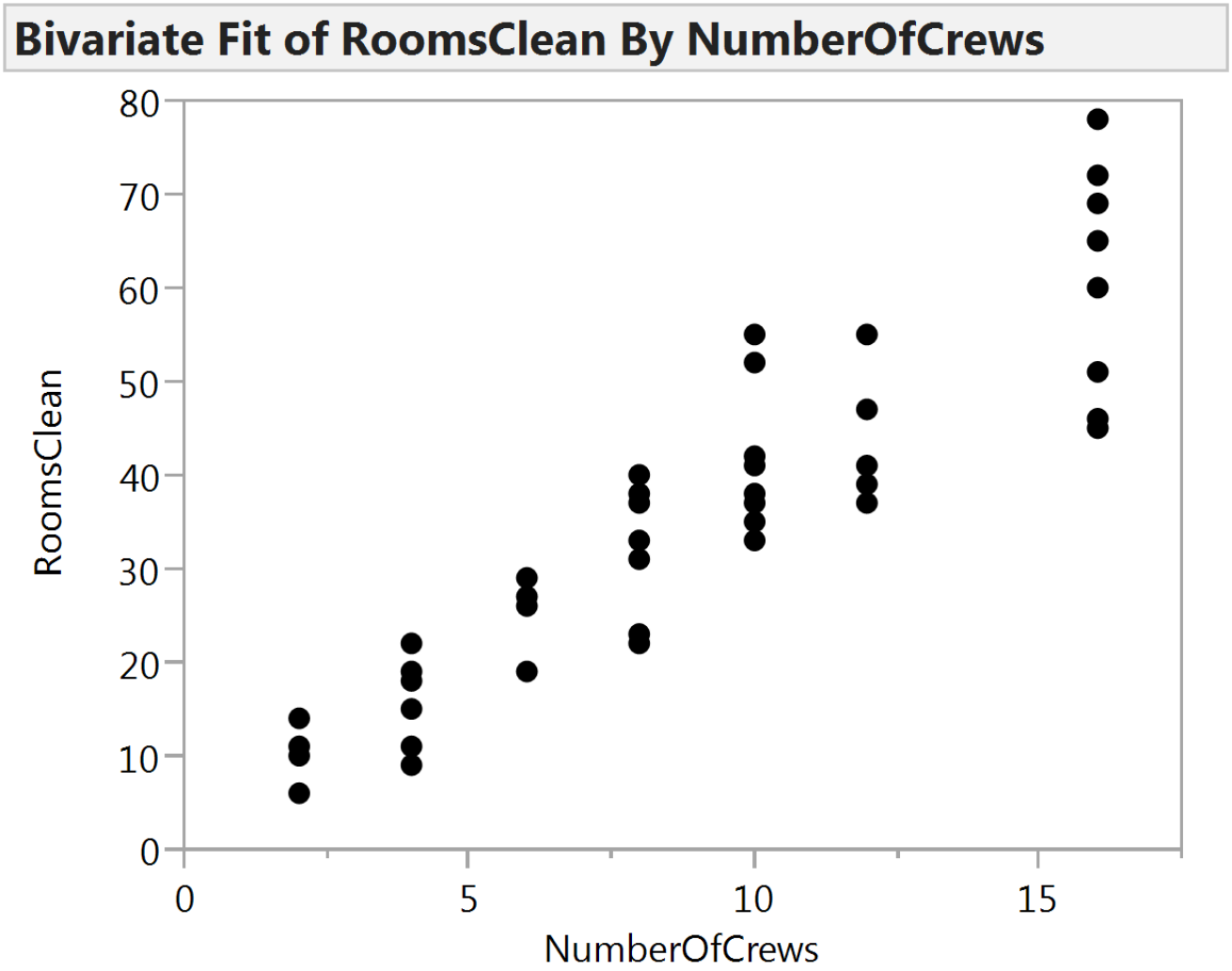
- The least squares estimates are given by:

$$b_1 = r\frac{s_y}{s_x} \qquad \text{and} \qquad b_0 = \bar{y} - b_1\bar{x},$$

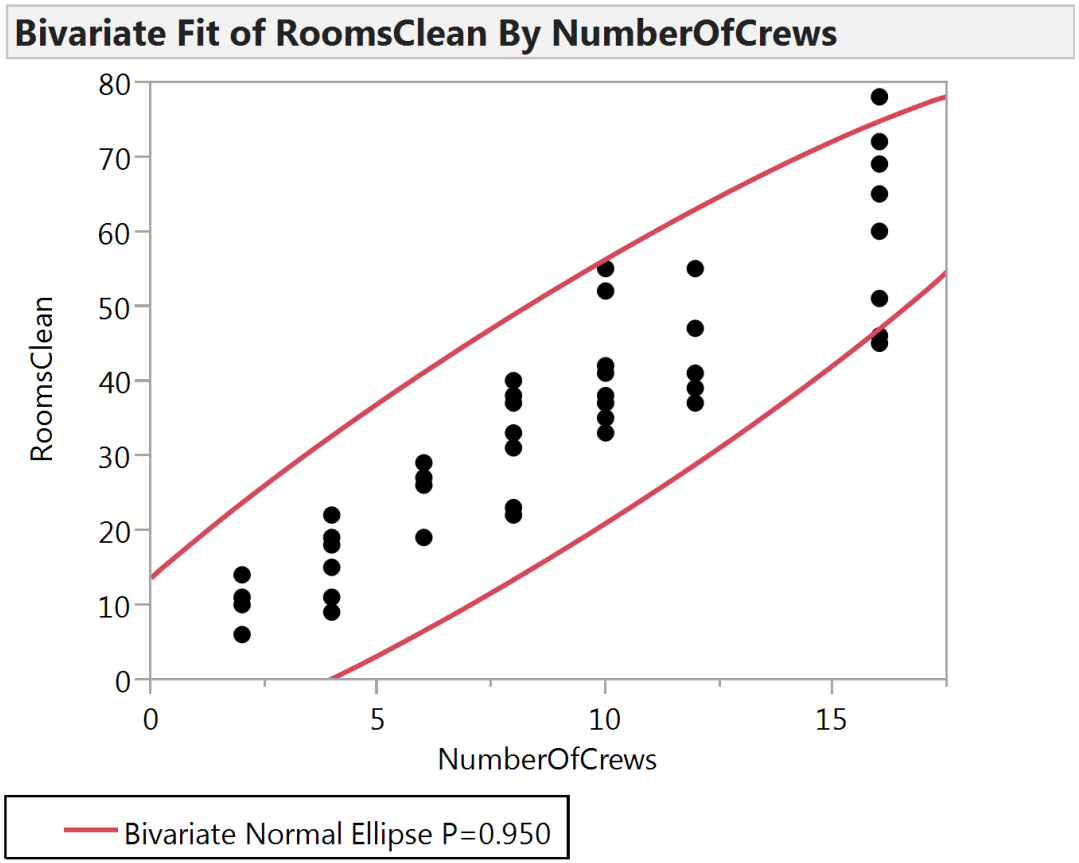
where r is the correlation between y and x.

The cleaning crews data set

How many crews does it take to clan 40 rooms?



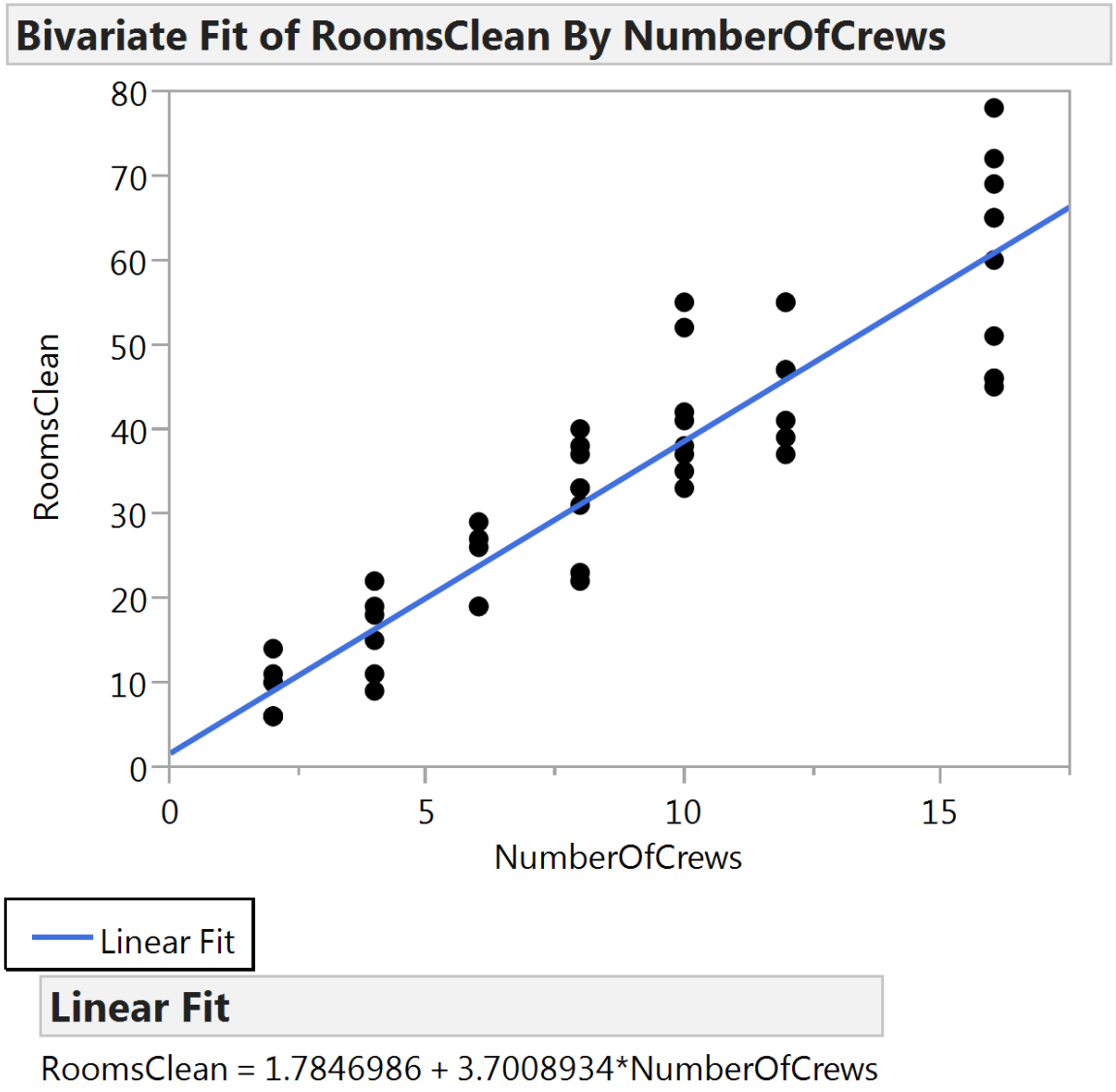
Adding in the correlation



Correlation

Variable	Mean	Std Dev	Correlation	Signif. Prob	Number
NumberOfCrew	8.679245	4.802938	0.925663	<.0001*	53
RoomsClean	33.90566	19.20263			

The least squares line



Interpreting the slope and intercept

- Don't forget the cardinal rule of data analysis: always, always plot your data.
- Our fitted model: $\hat{y} = b_0 + b_1 x$.
  - Intercept: b_0: the value of y, when x $= 0$. It has the units of y.
  - Slope: b_1: the change in y for every one unit change in x. Always understand the units on b_1. They are the units of y over the units of x.

Interpretations for the cleaning crews

- Slope: 3.7 additional rooms per additional crew.
- Intercept: more problematic. Sometimes the intercept does not have a clear interpretation. Sometimes it is an extrapolation outside the range of the data.
- It is possible to force a line through the intercept, but it is not usually considered *good practice*.
- We like our models to have as much flexibility as possible to fit well where the business operates. Forcing the line through the origin removes some of this flexibility. Further, it might just be a mistake to force it through the origin.

Module summary

1. Global and local approaches to prediction
2. Line definition
3. The least squares fit to data
4. The residuals
5. Interpreting the slope and intercept of a regression

Next time

- Residual analysis