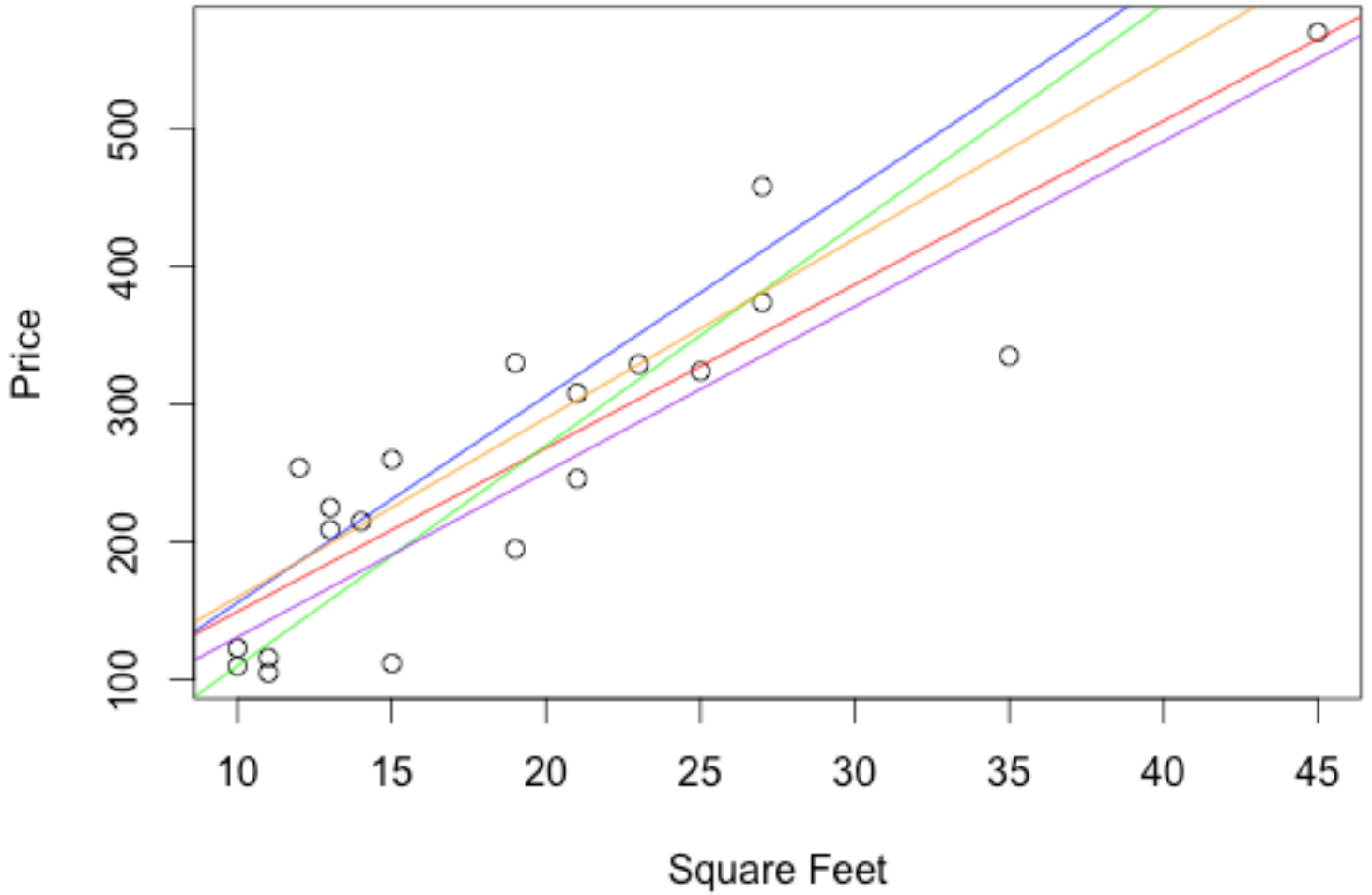
The Least-Squares Line

Keegan Korthauer
Department of Statistics
UW Madison

Recall the Housing Data



Which line "fits" the data best?

Simple Linear Model

- Our goal is to find the line that best describes the linear relationship between two variables:
  - **Independent** (or **explanatory**) variable x_i
  - **Dependent** (or **response**) variable y_i

- The simple linear regression model assumes:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

  - β_0: Intercept coefficient
  - β_1: Slope coefficient
  - ε_i: Random error

- Given paired observations (x_i, y_i), how do we estimate the regression coefficients β_0 and β_1?

Line of "Best Fit"

- β_0 and β_1 have some true underlying value (that we do not know – think of them as parameters)

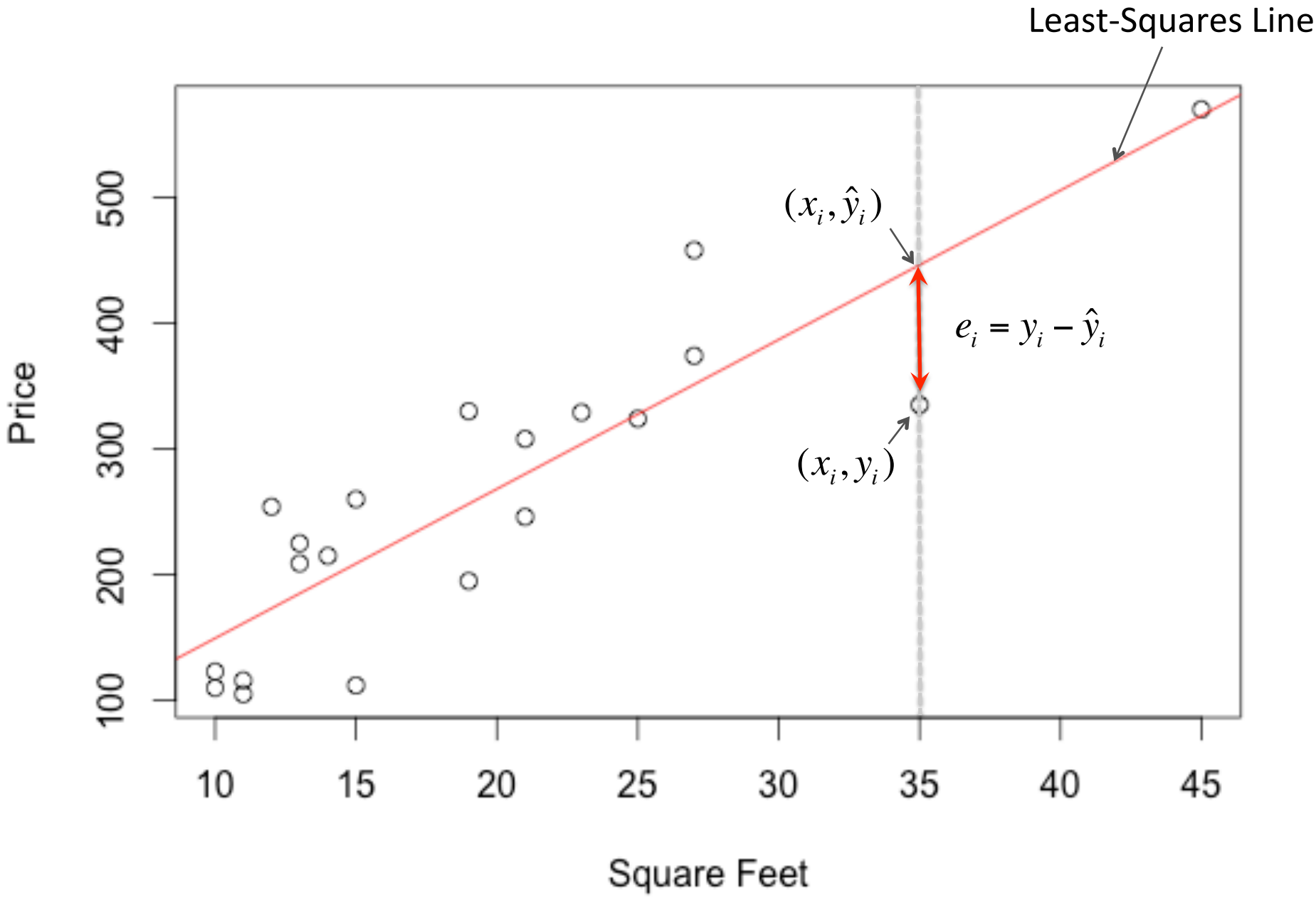
- We estimate them to find the line of "best fit":

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- We define the line of "best fit" to be the **Least Squares Line:**
  - The line that minimizes the sum of the squared **vertical distances of each point from the line**

Known as **residuals**

The Residual



The Least-Squares Line and Residuals

- The least squares line is the line that minimizes the **error sum of squares (SSE):**

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

- Points **above** the least squares line have **positive** residuals
- Points **below** the least squares line have **negative** residuals

- The closer the residuals are to zero, the better the least squares line fits the data
  - A residual of zero means the data point lies right on the line of best fit

Errors vs. Residuals

- NOT the same thing!
- **Residual (e_i)**: difference between observed value y_i and fitted value $\hat{y}_i$
  - Can calculate given observed data and estimated coefficients
- **Error (ε_i)**: difference between observed value y_i and the true value $\beta_0 + \beta_1 x_i$
  - Unknown since we do not know the true values of the coefficients
- So the sum of squared **errors** (SSE) is a misnomer, since it is actually the sum of squared **residuals**

Least-Squares Coefficients

- We want to solve for the values of the coefficients that minimize the following sum:

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- A little bit of calculus (try it yourself, or see the derivations on pages 533-534) shows that these quantities are:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

Often easier to compute by hand

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Interpreting Coefficients

- $\hat{\beta}_0$ is the predicted value of y when x is equal to zero
  - Often this is nonsensical – e.g. square footage of zero?

- $\hat{\beta}_1$ is the predicted increase in y for every one unit increase in x
  - Often of the most interest to us
  - In the housing example, interpret as the increase in price (thousands of dollars) for every additional hundred square feet

- We can use these estimates to predict the value of y for a new observation x (*as long as x is within the range of observed values*)

Housing Data Example

Let x = square feet, y = cost. Given the following summary data, compute the regression coefficients and predict the price of a home with 2500 square feet:

$$\bar{x} = 19.3,\ \bar{y} = 259.9$$

$$\sum_{i=1}^{n} x_i y_i = 119{,}156$$

$$\sum_{i=1}^{n} x_i^2 = 9036,\ \sum_{i=1}^{n} y_i^2 = 1{,}639{,}188$$

*I changed the example a bit so numbers are slightly different than when we were first presented with this example to calculate correlation

An Alternative Representation of the Line

Correlation Coefficient

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Least Squares Coefficient

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

We can derive the following relationship: $\hat{\beta}_1 = r\frac{s_y}{s_x}$

Further solving for the intercept and plugging the coefficients into the simple linear model, we get:

$$\hat{y}_i - \bar{y} = r\frac{s_y}{s_x}(x_i - \bar{x})$$

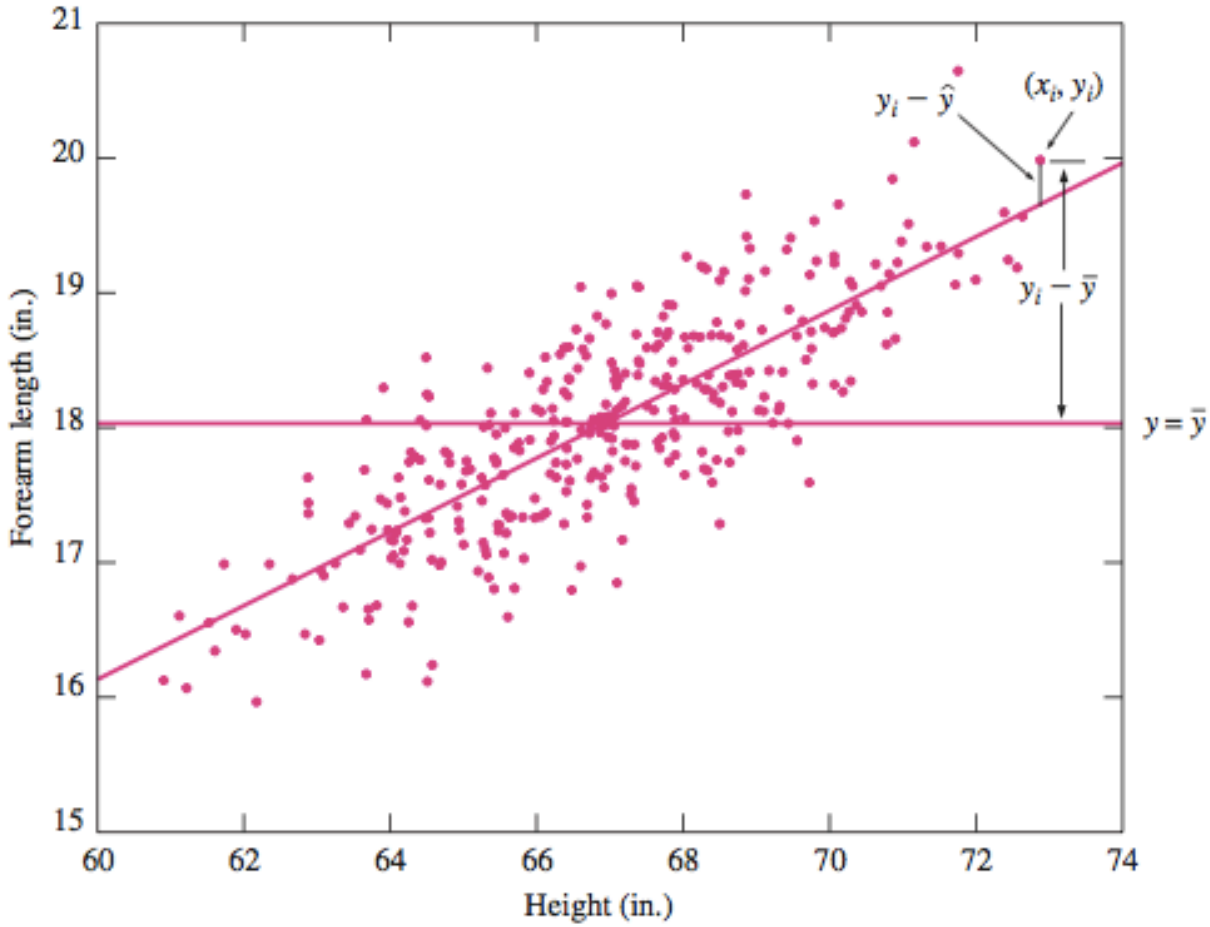


FIGURE 7.12 Heights and forearm lengths of men. The least-squares line and the horizontal line $y = \bar{y}$ are superimposed.

More Sums of Squares

The following quantities are useful for describing how well a linear regression model fits (also applicable in multiple regression):

- Error Sum of Squares $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$
- Total Sum of Squares $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$
- Regression Sum of Squares $SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$

Note that $SST = SSR + SSE$

Coefficient of Determination

- The correlation coefficient r can be thought of as a measure of how well the simple linear regression model fits

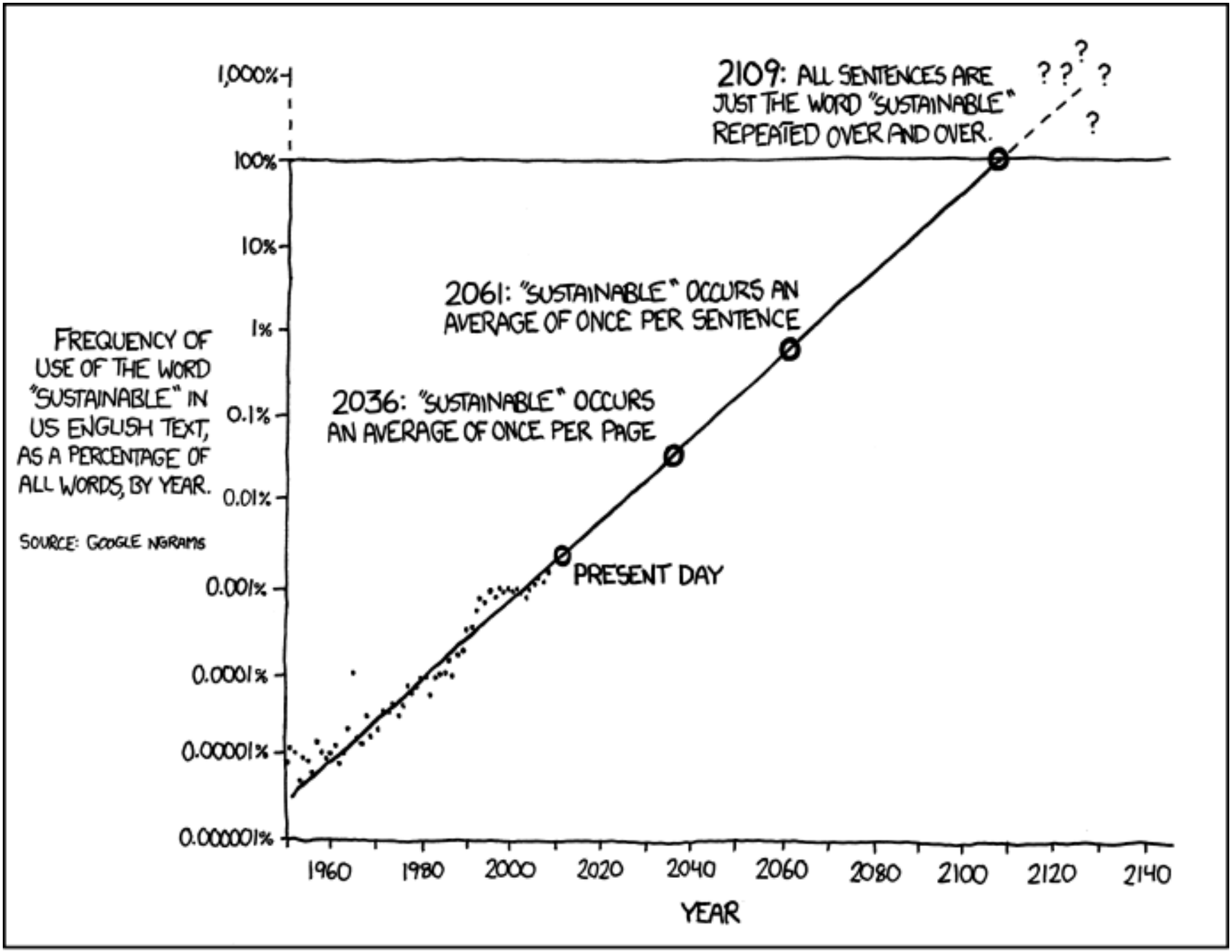
- The squared correlation coefficient, called the **coefficient of determination** can be interpreted as the proportion of total variance in the data that is explained by the regression model

$$r^2 = \frac{\text{Regression sum of squares}}{\text{Total sum of squares}} = \frac{SSR}{SST}$$

Warnings

- Never **extrapolate** (assume the linear regression model holds for x values outside of the observed data range)!

- Don't use the linear regression model when the relationship between x and y is **not linear**
  - Just like in the case of correlation, it only makes sense to describe a linear relationship



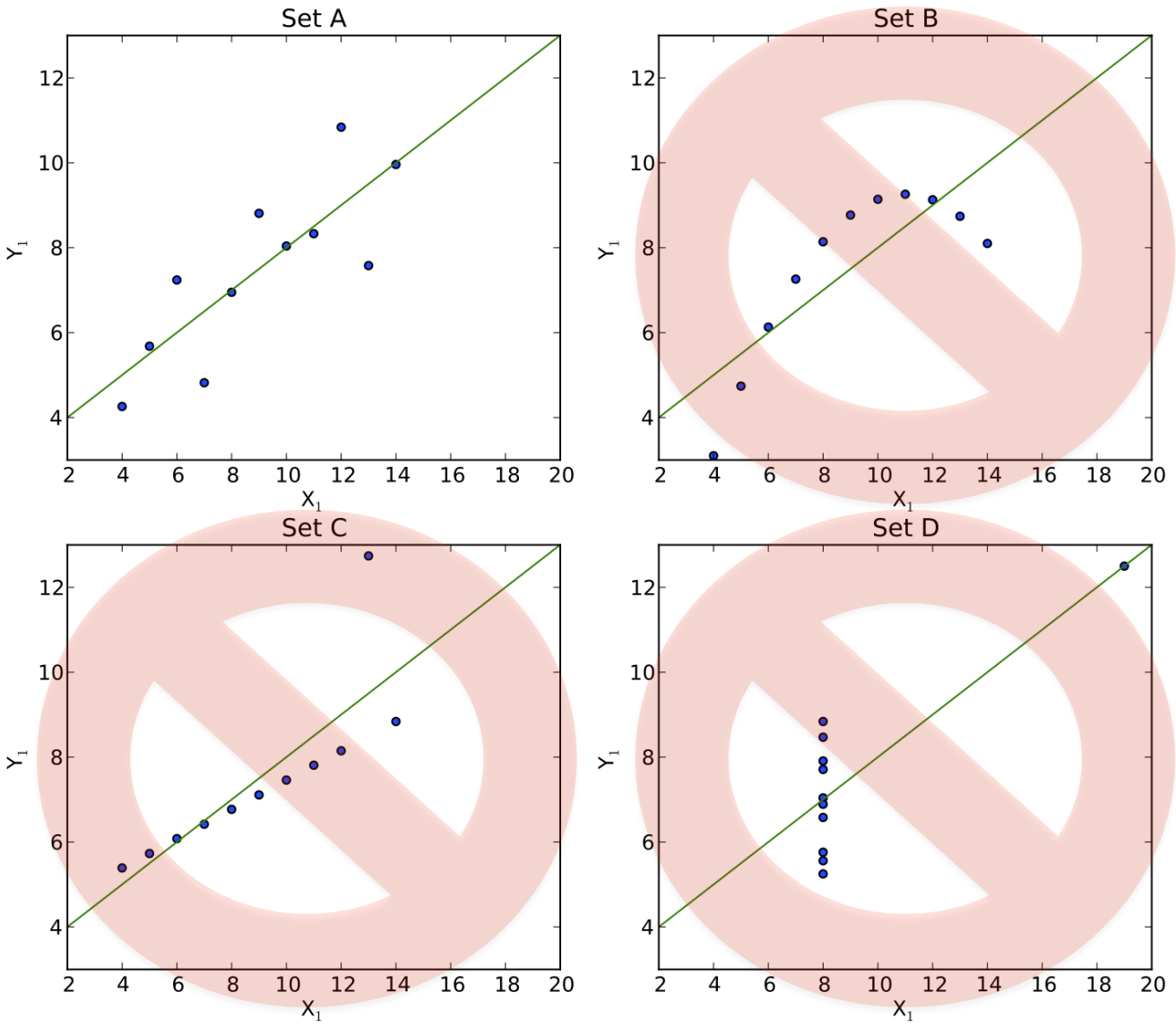
THE WORD "SUSTAINABLE" IS UNSUSTAINABLE.

Source: xkcd.com

Tip: **Always** Plot the Data!

	Set A		Set B		Set C		Set D	
	X	Y	X	Y	X	Y	X	Y
0	10	8.04	10	9.14	10	7.46	8	6.58
1	8	6.95	8	8.14	8	6.77	8	5.76
2	13	7.58	13	8.74	13	12.74	8	7.71
3	9	8.81	9	8.77	9	7.11	8	8.84
4	11	8.33	11	9.26	11	7.81	8	8.47
5	14	9.96	14	8.10	14	8.84	8	7.04
6	6	7.24	6	6.13	6	6.08	8	5.25
7	4	4.26	4	3.10	4	5.39	19	12.50
8	12	10.84	12	9.13	12	8.15	8	5.56
9	7	4.82	7	7.26	7	6.42	8	7.91
10	5	5.68	5	4.74	5	5.73	8	6.89
mean	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
std	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03
corr	0.82		0.82		0.82		0.82	
lin. reg.	$y = 3.00 + 0.500x$		$y = 3.00 + 0.500x$		$y = 3.00 + 0.500x$		$y = 3.00 + 0.500x$	

Tip: **Always** Plot the Data!



Notes

- When we fit a SLR model, we are estimating the regression coefficients
- In reality, we don't know their true value
- The estimates will change from experiment to experiment
  - They are **random**
  - We will compute their standard deviations so that we can find CIs and perform HTs on them

Next

- More on simple linear regression:
  - inference on coefficients
  - model assumptions
  - transformations

- HW 10 posted soon, Due Friday April 25th