Announcements

1. Turn in HW #2
2. Lab due Monday

Agenda

1. Bivariate Relationships
2. Simple Linear Regression
3. Residuals

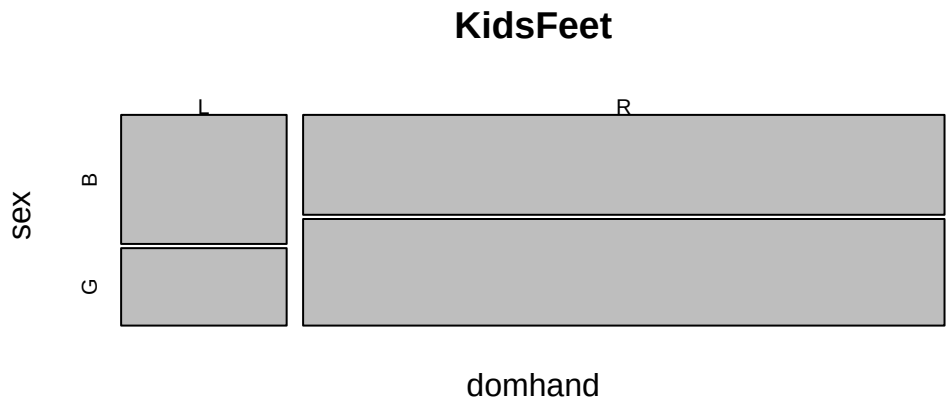
Bivariate Relationships

- Response variable (aka dependent variable): the variable that you are trying to understand
- Explanatory variable (aka independent variable, aka predictor): the variable that you think might be related to the response variable

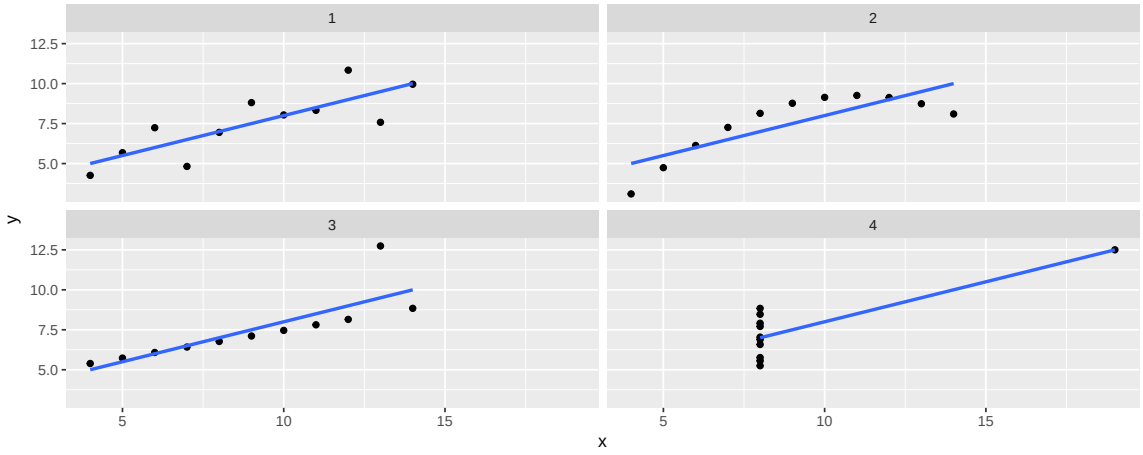
1. Visualize: Put response variable on y-axis and explanatory variable on x-axis
   - Two numerical variables: scatterplot [`qplot()`]
     - Overall patterns and deviations from those patterns
     - Form (e.g. linear, quadratic, etc.), direction (positive or negative), and strength (how much scatter?)
     - Outliers
   - Two categorical variables: mosaic plot [`mosaicplot()`]
   - Numerical response and a categorical explanatory variable:
     - Side-by-side box plots [`geom = "boxplot"`]
     - Multiple density plots [`geom = "density"` with `color` aesthetic or facets]
   - Multivariate relationships:
     - For a third variable that is categorical, use the `color` aesthetic or facets
     - For a third variable that is numerical, consider using the `cuts` option, or 3d effects!
2. Numerical Summary: Correlation (r)—a numerical measure of direction and strength of a *linear* relationship!

```
library(mosaic)
qplot(data = KidsFeet, y = length, x = width)
qplot(data = KidsFeet, y = length, x = width, color = sex)
qplot(data = KidsFeet, y = length, x = sex, geom = "boxplot")
qplot(data = KidsFeet, x = length, color = sex, geom = "density")
qplot(data = KidsFeet, x = length, facets = ~sex, geom = "density")
```

```
mosaicplot(domhand ~ sex, data = KidsFeet)
```



Correlation The (Pearson Product-Moment) correlation coefficient [`cor()`] is a measure of the strength and direction of the *linear* relationship between two numerical variables. It is usually denoted r and is measured on the scale of $[-1, 1]$.



Note that correlation only measures the strength of a *linear* relationship. In each of the four very different (Anscombe) data sets shown above, the correlation coefficient is the same (up to three digits)! Get a feel for the value of the correlation coefficient in different scatterplots.

Correlation Problem

1. An article reported that there was a 0.42 correlation between alcohol consumption and income among adults with a four-year college degree. Is it reasonable to conclude that increasing one's alcohol consumption will increase one's income? Explain why or why not.

Simple linear regression Linear regression can help us understand changes in a numerical response variable in terms of a numerical explanatory variable.

A simple linear regression model for y in terms of x takes the form

$$y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i\,, \text{ for } i = 1, \ldots, n$$

- β_0 is the *intercept* and β_1 is the *slope* coefficient. The ϵ_i's are the *errors*, or *noise*.
- There is only one regression line that fits the data best using a least squares criteria. That is, the *ordinary least squares* regression line is unique.
- The true values of the unknown parameters β_0 and β_1 are estimated by b_0 and b_1 (or if you prefer, $\hat{\beta}_0$ and $\hat{\beta}_1$)
- The *fitted values*, or *predicted values* are given by

$$\hat{y}_i = b_0 + b_1 \cdot x_i$$

- The model almost never fits perfectly, but what is left over is captured by the *residuals* ($e_i = y_i - \hat{y}_i$)

Example: RailTrail The Pioneer Valley Planning Commission (PVPC) collected data north of Chestnut Street in Florence, MA for ninety days from April 5, 2005 to November 15, 2005. Data collectors set up a laser sensor, with breaks in the laser beam recording when a rail-trail user passed the data collection station. The data is captured in the `RailTrail` data set.

```
library(mosaic)
data(RailTrail)
```

1. In R, create a scatterplot [`qplot()`] for the response: *volume* in terms of explanatory: *avgtemp*.

2. Describe the shape, direction, and strength of the relationship. Guess the correlation, r.

3. Compute the correlation coefficient [`cor()`]. Hint: use the same syntax that favstats() uses, that is, `cor(`$y \sim x$`, data = TheData)`. Is r what you expected?

4. Fit the linear regression model using `lm()`. Again, use the `favstats()` syntax.

5. Interpret the coefficients for the `Intercept` (b_0) and `avgtemp` (b_1) terms.