Statistics 191: Introduction to Applied Statistics

Review

Jonathan Taylor
Department of Statistics
Stanford University

February 22, 2010

Outline

- What is a regression model?
- Descriptive statistics – numerical
- Descriptive statistics – graphical
- Inference about a population mean
- Difference between two population means

What is course about?

- It is a course on applied statistics.
- Hands-on: we use R, an open-source statistics software environment.
- We will start out with a review of introductory statistics to see `R` in action.
- Main topic is "(linear) regression models": these are the *bread and butter* of applied statistics.

What is a "regression" model?

A regression model is a model of the relationships between some *covariates (predictors)* and an *outcome*. Specifically, regression is a model of the *average* outcome *given* the covariates.

Heights of couples

- To study height of the wife in a couple, based on the husband's height and her parents height: `Wife` is the outcome, and the covariates are `Husband`, `Mother`, `Father`.
- A mathematical model, using only `Husband`'s height:

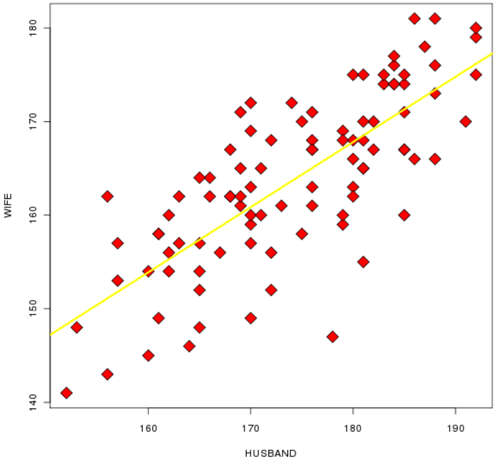
$$\texttt{Wife} = f(\texttt{Husband}) + \varepsilon$$

  where f gives the average height of the wife of a man of height `Husband` and ε is "error": not *every* man of height of `Husband` marries a woman of height $f(\texttt{Husband})$.
- A statistical question: is there *any* relationship between covariates and outcomes – is f just a constant?
- Here is some
  http://stats191.stanford.edu/review.htmldata using only

Heights data

Statistics 191: Introduction to Applied Statistics

Jonathan Taylor
Department of Statistics
Stanford University



Heights data

Linear regression models

- We might model the data as

$$\texttt{Wife} = \beta_0 + \beta_1 \texttt{Husband} + \varepsilon.$$

- This model is *linear* in `Husband`, it is a *simple linear regression model*.
- Another model:

$$\texttt{Wife} = \beta_0 + \beta_1 \texttt{Husband} + \beta_2 \texttt{Mother} + \beta_3 \texttt{Father} + \varepsilon.$$

- Also linear (in `Husband`, `Mother`, `Father`).
- Which model is better? We need a tool to compare models ...

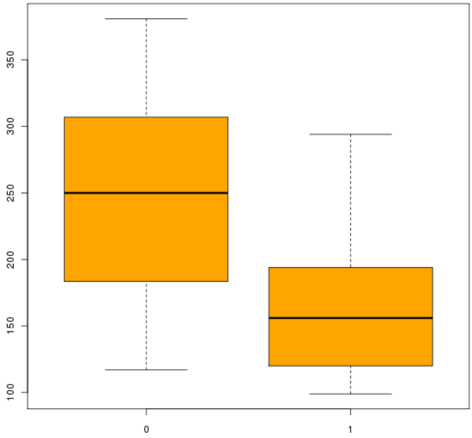
Right-to-work example

Jonathan Taylor
Department of Statistics
Stanford University

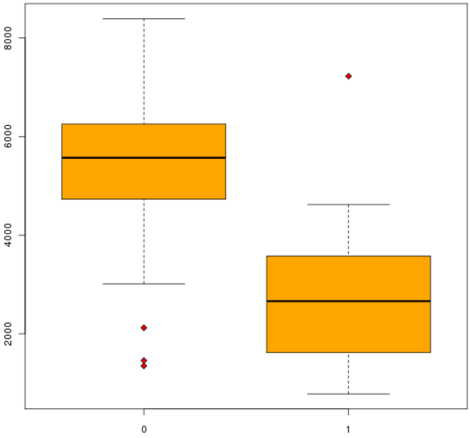
Data description

- Income: income for a four-person family
- COL: cost of living for a four-person family
- PD: Population density
- URate: rate of unionization in 1978
- Pop: Population
- Taxes: Property taxes in 1972
- RTWL: right-to-work indicator

Right-to-work vs. cost of living



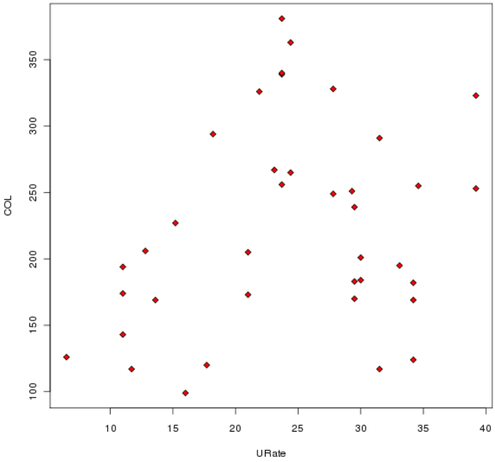
Right-to-work vs. income



Unionization vs. cost of living

Statistics 191: Introduction to Applied Statistics

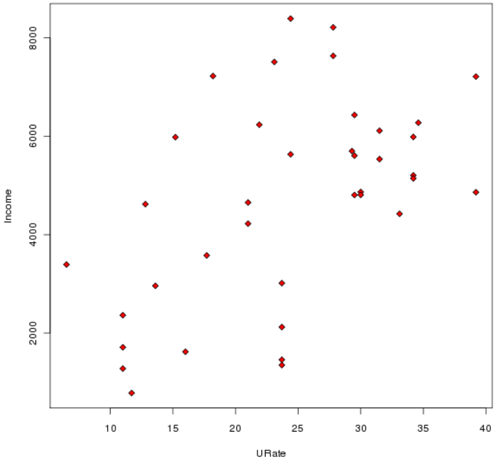
Jonathan Taylor
Department of Statistics
Stanford University



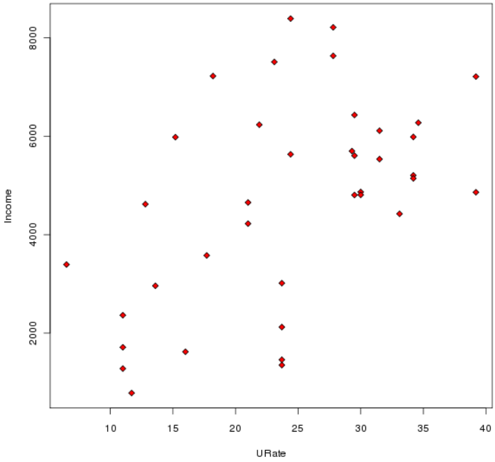
Unionization vs. income

Statistics 191: Introduction to Applied Statistics

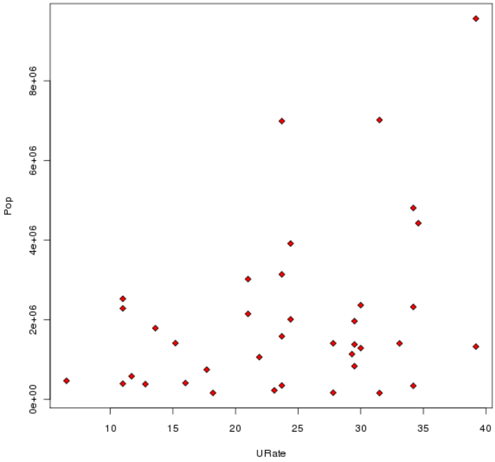
Jonathan Taylor
Department of Statistics
Stanford University



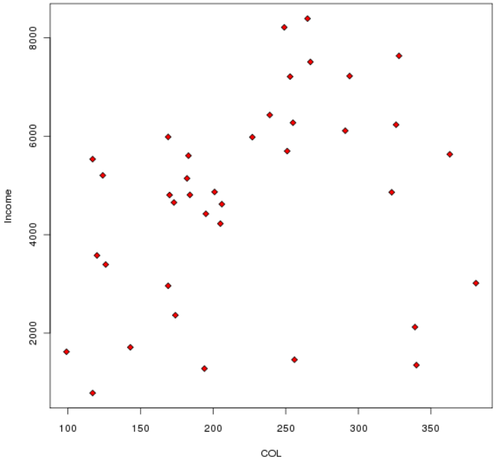
Unionization vs. income



Unionization vs. population



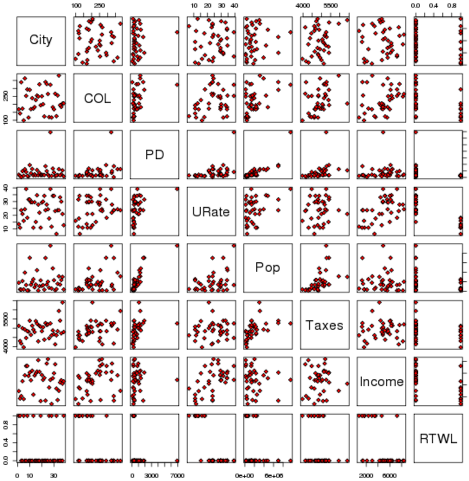
Cost-of-living vs. income



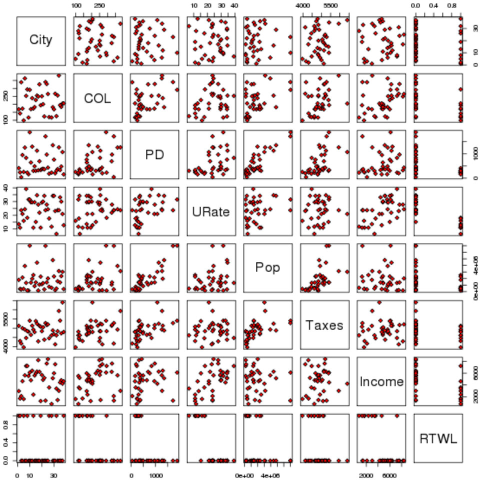
Full dataset

Statistics 191: Introduction to Applied Statistics

Jonathan Taylor
Department of Statistics
Stanford University



Without NYC



Right-to-work example

Building a model

Some of the main goals of this course:

- Build a statistical model describing the "effect of RTWL" on "COL"
- This model should recognize that other variables also affect "COL"
- What sort of "statistical confidence" do we have in our conclusion about "RTWL" and "COL"?
- Is the model adequate do describe this dataset?
- Are there other (simpler, more complicated) models?

Review

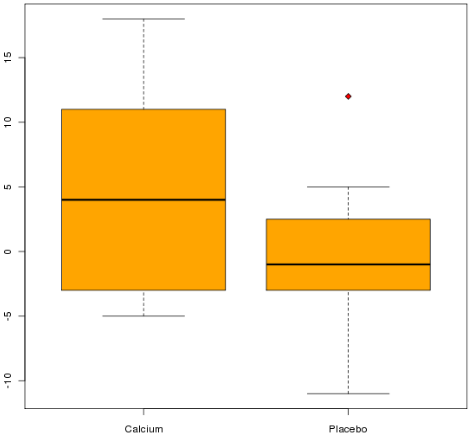
Effect of calcium on BP

- A study was conducted to study the effect of calcium supplements on blood pressure.
- More detailed data description can be found here.

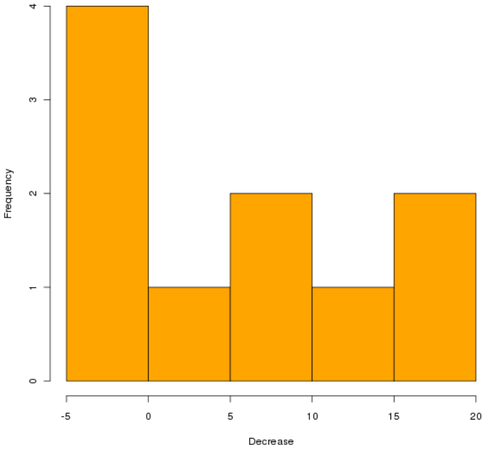
Questions

- What is the mean decrease in BP in the treated group? placebo group?
- What is the median decrease in BP in the treated group? placebo group?
- What is the standard deviation of decrease in BP in the treated group? placebo group?
- Is there a difference between the two groups? Did BP decrease more in the treated group?

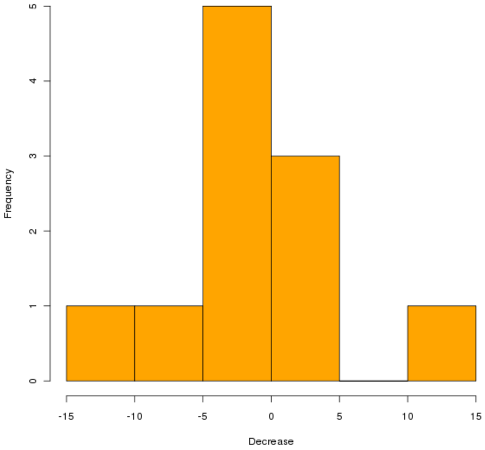
Boxplot



Histogram of Treated response



Histogram of Placebo response



Descriptive statistics – numerical

Jonathan Taylor
Department of Statistics
Stanford University

Mean of a sample

Given a sample of numbers $X = (X_1, \dots, X_n)$ the sample mean, $\overline{X}$ is

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Standard deviation of a sample

Given a sample of numbers $X = (X_1, \dots, X_n)$ the sample standard deviation S_X is

$$S_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

Descriptive statistics – numerical

Median of a sample

Given a sample of numbers $X = (X_1, \dots, X_n)$ the sample median is the "middle" of the sample: if n is even, it is the average of the middle two points. If n is odd, it is the midpoint.

Quantiles of a sample

Given a sample of numbers $X = (X_1, \dots, X_n)$ the q-th quantile is a point x_q in the data such that $q \cdot 100\%$ of the data lie to the left of x_q.
Example: the 0.5-quantile is the median: half of the data lie to the right of the median.

Inference about a population mean

Statistics 191: Introduction to Applied Statistics

Jonathan Taylor
Department of Statistics
Stanford University

Confidence interval

- If $(X_1, \dots, X_n)$ are independent, all having a normal distribution $N(\mu, \sigma^2)$, then a $(1-\alpha)$-confidence interval for μ is
$$\overline{X} \pm t_{n-1,1-\alpha/2} \cdot S_X/\sqrt{n}$$
- Where $t_{n-1,1-\alpha/2}$ is the $1-\frac{\alpha}{2}$ quantile of t_{n-1} random variable, defined by
$$\mathbb{P}(T_{n-1} \leq t_{n-1,1-\alpha/2}) = 1 - \frac{\alpha}{2}.$$

Inference about a population mean

Jonathan Taylor
Department of Statistics
Stanford University

Testing whether mean is 0

- Suppose we want a two-sided test of whether $\mu = 0$ based on a sample X, at level α.
- Compute

$$T = \frac{\overline{X}}{S_X / \sqrt{n}}.$$

- If $|T| > t_{n-1, 1-\alpha/2}$, then reject $H_0 : \mu = 0$.

Difference between means

BP example

- In our setting, we have two groups that we have reason to believe are different.
- We have two samples:
  1. $(X_1, \dots, X_{10})$ (Calcium)
  2. $(Z_1, \dots, Z_{11})$ (Placebo)
- Does treatment have an effect?
- We can answer this statistically by testing the null hypothesis $H_0 : \mu_X = \mu_Z$?

Difference between means

Statistics 191: Introduction to Applied Statistics

Jonathan Taylor
Department of Statistics
Stanford University

Testing $H_0 : \mu_X = \mu_Z$

- If variances are assumed equal, pooled t-test is appropriate

$$T = \frac{\overline{X} - \overline{Z}}{S_P\sqrt{\frac{1}{10} + \frac{1}{11}}}, \qquad S_P^2 = \frac{9 \cdot S_X^2 + 10 \cdot S_Z^2}{19}.$$

- For two-sided test at level α, reject if $|T| > t_{19,1-\alpha/2}$.

Our first regression model

Unified dataset

- Put two samples together:

$$Y = (X_1, \dots, X_{10}, Z_1, \dots, Z_{11}).$$

- Under the same assumptions as the pooled t-test:

$$Y_i \sim N(\mu_i, \sigma^2)$$
$$\mu_i = \begin{cases} \mu_X & 1 \leq i \leq 10 \\ \mu_Z & 11 \leq j \leq 21 \end{cases}$$

Our first regression model

t-test as regression model

- This is a (regression) model for the sample Y. The (qualitative) variable `Treatment` is called a "covariate" or "predictor".
- The decrease in BP is an outcome variable.
- We assume that the relationship between treatment and average decrease in BP is simple: it depends only on which group a subject is in.
- This relationship is "modelled" through the mean vector $\mu = (\mu_1, \dots, \mu_{21})$.