STAT 203: Regression Models and ANOVA
Lecture 1: Introduction and Review

Nancy R. Zhang

Statistics 203, Stanford University

January 7, 2009

Topics outline:

The goal of this course is to equip you with tools for examining relationships among a given set of variables.

1. Linear regression and ANOVA models

$$y = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 + \cdots + \text{error}$$

2. Variable transformations, weighted least squares
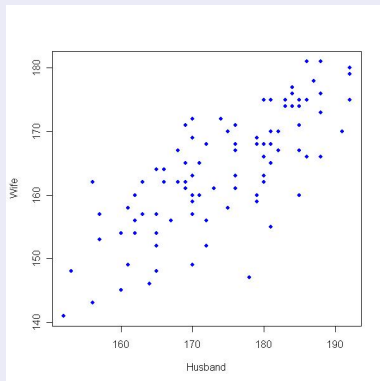   ... when the errors are not quite independent Gaussian.

3. Model building and variable selection
   ... Which variables to include in the model?

4. Non-linear models

Data analysis using R will be a substantial component.

Survey: Linear regression and ANOVA models

Example: heights of husbands and wifes

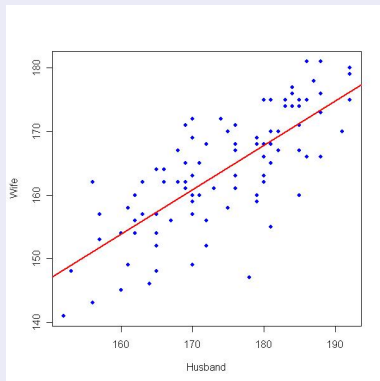Do people of similar heights tend to marry each other?

Plotted are the heights of a sample of newly married couples.

X: height of husband (cm)
Y: height of wife (cm)

Survey: Linear regression and ANOVA models

Example: heights of husbands and wifes

Simple linear model:

$$Y = \beta_0 + \beta_1 X + \text{Error}.$$

Assumptions:

1. Errors are independent.
2. Simple linear relationship.
3. No causal assumptions.

Survey: Linear regression and ANOVA models

Example: presidential elections

Goal: Predict which candidate wins.

Variables:

- Proportion democratic votes (V)
- Incumbent party? (I)
- Incumbent candidate? (D)
- Election held during war? (W)
- Economic growth during election year (G)
- Economic growth during last four years (P)
- Number of "good growth quarters" (N)

Year	Dem Votes
1916	0.5168
1920	0.3612
1924	0.4176
1928	0.4118
1932	0.5916
1936	0.6246
1940	0.5500
1944	0.5377
1948	0.5237
1952	0.4460
1956	0.4224
1960	0.5009
1964	0.6134
1968	0.4960
1972	0.3821
1976	0.5105
1980	0.4470
1984	0.4083
1988	0.4610
1992	0.5345
1996	0.5474

Note: some predictors are *quantitative*, others are *qualitative*.

Survey: Linear regression and ANOVA models

Example: presidential elections

`pairs` in R:

Variables:

- Proportion democratic votes (V)
- Incumbent party? (I)
- Incumbent candidate? (D)
- Election held during war? (W)
- Economic growth during election year (G)
- Economic growth during last four years (P)
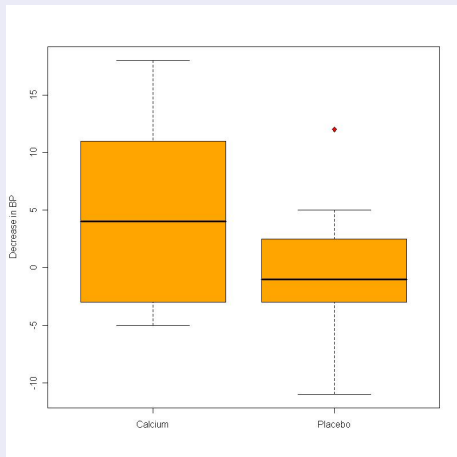- Number of "good growth quarters" (N)
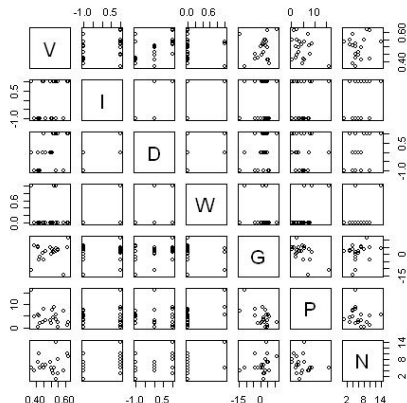
Which variables should be included in the model?

The simplest linear model

Calcium supplements on blood pressure

A study was conducted to study the effect of calcium supplements on blood pressure. Subjects were divided into two groups. One group received a calcium supplement while the other received a placebo.

Is there a significant difference between the two groups?

Review: Descriptive statistics

Mean of a sample

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

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

Review: Descriptive statistics

Median of a sample

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

Review: Inferences about a population mean

Confidence interval

- We observe measurements $(X_1, \ldots, X_n)$. Under the assumption that they are independent $N(\mu, \sigma^2)$, then the population mean μ can be estimated by the sample mean $\bar{X}$. A 95% confidence interval for μ can be computed:

$$\overline{X} \pm t_{n-1, 1-\alpha/2} \cdot S_X / \sqrt{n}$$

- Where $t_{n-1, 1-\alpha/2}$ is the $1 - \frac{\alpha}{2}$ quantile of a t_{n-1} random variable, defined by

$$P(T_{n-1} \leq t_{n-1, 1-\alpha/2}) = 1 - \frac{\alpha}{2}.$$

Review: Basic Distributions

Gaussian

$$\text{Density: } f_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

χ^2

If $X_1, X_2, \ldots, X_m$ are i.i.d. Gaussian, then $\sum_{i=1}^{m} X_i^2$ is distributed as a χ_m^2 random variable.

t

If $X \sim N(0, 1)$ is independent of $Z \sim \chi^2(m)$, then $X/\sqrt{Z/m}$ is distributed as a t_m random variable.

Inference about a population mean

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 $H_0 : \mu = 0$.

Comparing the means of two groups

In the Calcium-Blood Pressure example, we have two groups, Placebo and Treatment. We would like to test whether their means are equal.

- $(X_1, \ldots, X_{10})$ (Calcium)
- $(Z_1, \ldots, Z_{11})$ (Placebo)

In other words, does treatment have an effect?

We can answer this statistically by testing the null hypothesis

$$H_0 : \mu_X = \mu_Z?$$

Difference between means

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 $|T| > t_{19,1-\alpha/2}$.

T-test as a regression model

Combine the placebo and treatment samples:

$$Y_1, \ldots, Y_{10} \quad \text{Treated}$$

$$Y_{11}, \ldots, Y_{21} \quad \text{Placebo}$$

Now, let

$$X_i = \begin{cases} 0 & \textit{Placebo} \\ 1 & \textit{Treated} \end{cases}$$

Under the same assumptions as the pooled t-test:

$$Y_i \sim N(\mu_{X_i}, \sigma^2)$$

In other words,

$$Y_i = \mu_0 + (\mu_1 - \mu_0)X_i + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma^2)$.

T-test as a regression model

$$Y_i = \mu_0 + (\mu_1 - \mu_0)X_i + \epsilon_i,$$

- Y_i, the decrease in blood pressure, is the response variable.
- X_i is the treatment variable, sometimes called a "covariate".
- This model makes the important assumption that the error variance is equal and does not depend on the treatment. This is the same assumption as that made in the pooled *t*-test.

Next lecture: Estimation of μ_0, $\mu_1 - \mu_0$, and inferences for a simple linear regression model.

For the next week...

1. Material covered: Simple and multiple linear regression.
2. Review on your own:
   1. The basic distributions: Gaussian, T, χ^2, F.
   2. Basic matrix algebra.
   3. Basic concepts of probability and hypothesis testing.
3. Course work:
   1. 4-5 bi-weekly problem sets.
   2. 1 In-class midterm
   3. Take-home final.