

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University

Statistics 202: Introduction to Data Mining

Simple Linear Regression

Jonathan Taylor
Department of Statistics
Stanford University

November 20, 2009

Outline

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University

Simple Linear Regression

- Some definitions for regression models.
- Specifying the model.
- Fitting the model: least squares.
- Inference.
- What is a T -statistic?
- “Inference” for β_1 .
- Linear combinations of β_0, β_1 .

Reminder

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University

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*.

Mathematical formulation

For height of couples data: a mathematical model, using only Husband's height:

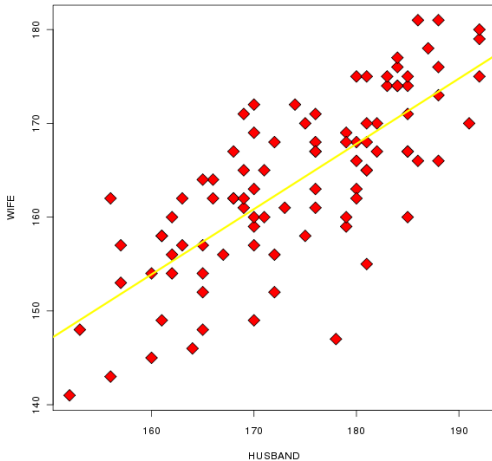
$$\text{Wife} = f(\text{Husband}) + \varepsilon$$

where f gives the average height of the wife of a man of height Husband and ε is the random error.

Height data

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University



Regression models

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University

Linear regression models

- A *linear* regression model says that the function f is a sum (linear combination) of functions of Husband.
- Simple linear regression model:

$$f(\text{Husband}) = \beta_0 + \beta_1 \cdot \text{Husband}$$

for some unknown parameter vector (β_0, β_1) .

- Could also be a sum (linear combination) of *known* functions of Husband:

$$f(\text{Husband}) = \beta_0 + \beta_1 \cdot \text{Husband} + \beta_2 \cdot \text{Husband}^2$$

Simple linear regression model

Specifying the (statistical) model

- *Simple linear* regression is the case when there is only one predictor:

$$f(\text{Husband}) = \beta_0 + \beta_1 \cdot \text{Husband}.$$

- Let Y_i be the height of the i -th wife in the sample, X_i be the height of the i -th husband.
- Model:

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{regression equation}} + \underbrace{\varepsilon_i}_{\text{error}}$$

where $\varepsilon_i \sim N(0, \sigma^2)$ are independent.

- This specifies a *distribution* for the Y 's given the X 's, i.e. it is a statistical model.

Fitting the model

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University

Least squares

- We will be using “least squares” regression. This measures the goodness of fit of a line by the sum of squared errors, *SSE*.
- Least squares regression chooses the line that minimizes

$$SSE(\beta_0, \beta_1) = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 \cdot X_i)^2.$$

- In principle, we might measure “goodness of fit” differently: why do we use least squares?

Least squares

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University

Alternative definition of (sample / population) mean

The mean of a sample (Y_1, \dots, Y_n) (or population Y) is the number that minimizes

$$SSE(\mu) = \sum_{i=1}^n (Y_i - \mu)^2 \quad (\text{population: } = \mathbb{E}((Y - \mu)^2)).$$

Alternative definition of (sample / population) median

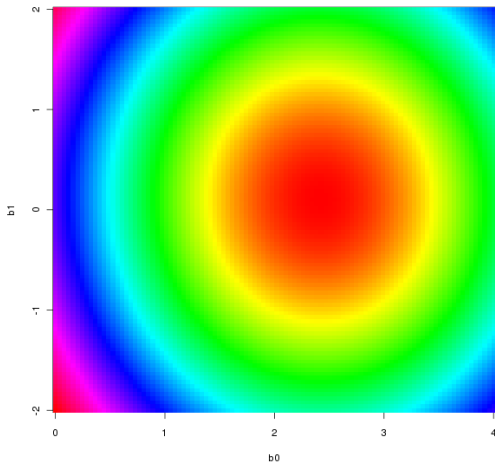
The median of a sample (Y_1, \dots, Y_n) (or population Y) is any number that minimizes

$$SAD(\mu) = \sum_{i=1}^n |Y_i - \mu| \quad (\text{population: } = \mathbb{E}(|Y - \mu|)).$$

Least squares

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University

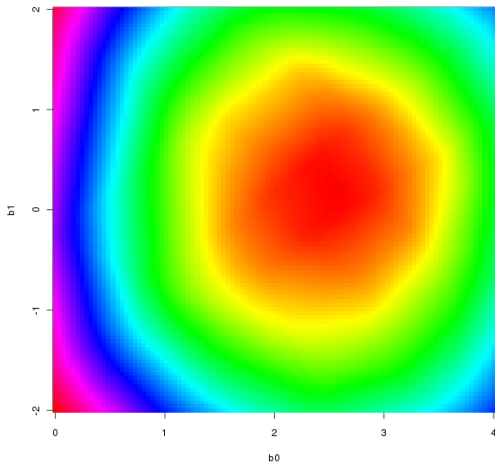


R code

Absolute deviation

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University



R code

Least Squares Solutions

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University

Regression line parameters: (β_0, β_1)

-

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\widehat{\text{Cov}}(X, Y)}{\widehat{\text{Var}}(X)}.$$

-

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

Estimating variance: σ^2

- Strength of association between Y and X will depend on variability of errors ε , as in two sample t -test.
-

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = \frac{SSE}{n-2} = MSE.$$

Least Squares

Predicting the mean

Mean can be estimated for any given husband of height X as

$$\hat{Y} = \hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1 \cdot X.$$

where $(\hat{\beta}_0, \hat{\beta}_1)$ are the minimizers of SSE.

Estimate of σ^2

-

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{f}(X_i))^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2.$$

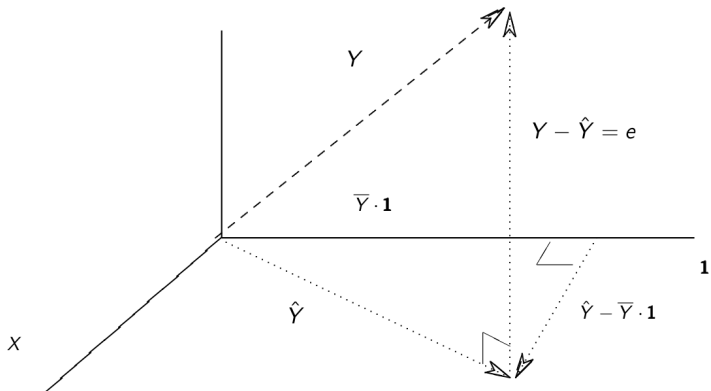
- Why $n - 2$? According to our statistical model

$$\frac{\hat{\sigma}^2}{\sigma^2} \sim \frac{\chi_{n-2}^2}{n-2}.$$

Geometry of Least Squares

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University



Inference

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University

What do we mean by inference?

- Generally: “learning something about the relationship between the sample (X_1, \dots, X_n) and (Y_1, \dots, Y_n) .”
- In the simple linear regression model, learning about β_0, β_1 :
 - learning: *confidence intervals, hypothesis tests.*

Tools for inference

- Most of the questions of “inference” in this course can be answered in terms of t -statistics or F -statistics.
- First we will talk about t -statistics, next class F -statistics.

Hypothesis tests

What is a (statistical) hypothesis?

Examples:

- One sample problem: given an independent sample (Z_1, \dots, Z_n) where $Z_i \sim N(\mu, \sigma^2)$, the *null hypothesis* $H_0 : \mu = 0$ says that in fact the population mean is 0.
- Two sample problem: given two independent samples $\mathbf{Z} = (Z_1, \dots, Z_n)$, $\mathbf{W} = (W_1, \dots, W_m)$ where $Z_i \sim N(\mu_1, \sigma^2)$ and $W_i \sim N(\mu_2, \sigma^2)$, the *null hypothesis* $H_0 : \mu_1 = \mu_2$ says that in fact the population mean of the two samples are identical.

Testing a hypothesis

- Usually, we test a null hypothesis, H_0 based on some test statistic T whose distribution is fully known when H_0 is true.

t -statistics

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University

What is a t -statistic?

- Start with $Z \sim N(0, 1)$ is standard normal and $X^2 \sim \chi^2_\nu$, independent of Z .

- Compute

$$T = \frac{Z}{\sqrt{\frac{X^2}{\nu}}}.$$

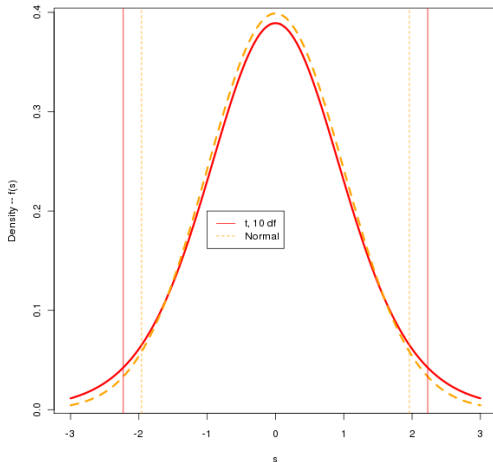
- Then, $T \sim t_\nu$ has a t -distribution with ν degrees of freedom.
- Generally, a t -statistic has the form

$$T = \frac{\text{parameter estimate} - \text{true parameter, i.e. } \hat{\beta}_1 - \beta_1}{\text{standard error of parameter, i.e. } SE(\hat{\beta}_1)}$$

t vs. Normal

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University



Example of a t -statistic

One sample t -test

- Given an independent sample (Z_1, \dots, Z_n) where $Z_i \sim N(\mu, \sigma^2)$ we can test $H_0 : \mu = 0$ using a T -statistic.
- We “know” that the random variables

$$\bar{Z} \sim N(\mu, \sigma^2/n), \quad \frac{S^2(Z)}{\sigma^2} \sim \frac{\chi_{n-1}^2}{n-1}$$

are independent.

- Therefore

$$\frac{\bar{Z} - \mu}{S(Z)/\sqrt{n}} = \frac{(\bar{Z} - \mu)/(\sigma/\sqrt{n})}{S(Z)/\sigma} \sim t_{n-1}.$$

Confidence intervals

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University

What is a confidence interval?

Examples:

- One sample problem: instead of deciding whether $\mu = 0$, we might want to come up with an (random) interval $[L, U]$ based on the sample Z such that the probability the true (nonrandom) μ is contained in $[L, U]$ equal to $1 - \alpha$, i.e. 95%.
- Two sample problem: find a (random) interval $[L, U]$ based on the samples \mathbf{Z} and \mathbf{W} such that the probability the true (nonrandom) $\mu_1 - \mu_2$ is contained in $[L, U]$ is equal to $1 - \alpha$, i.e. 95%.

Example of a confidence interval

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University

One sample: confidence interval for μ

- Given an independent sample (Z_1, \dots, Z_n) where $Z_i \sim N(\mu, \sigma^2)$ we can test construct a $(1 - \alpha) * 100\%$ using the numerator and denominator of the t -statistic...
- Let $q = t_{n-1, (1-\alpha)/2}$

$$\begin{aligned}1 - \alpha &= P\left(-q \leq \frac{\mu - \bar{Z}}{S(Z)/\sqrt{n}} \leq q\right) \\&= P\left(-q \cdot S(Z)/\sqrt{n} \leq \mu - \bar{Z} \leq q \cdot S(Z)/\sqrt{n}\right) \\&= P\left(\bar{Z} - q \cdot S(Z)/\sqrt{n} \leq \mu \leq \bar{Z} + q \cdot S(Z)/\sqrt{n}\right)\end{aligned}$$

Inference in regression

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University

Heights example

- Model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

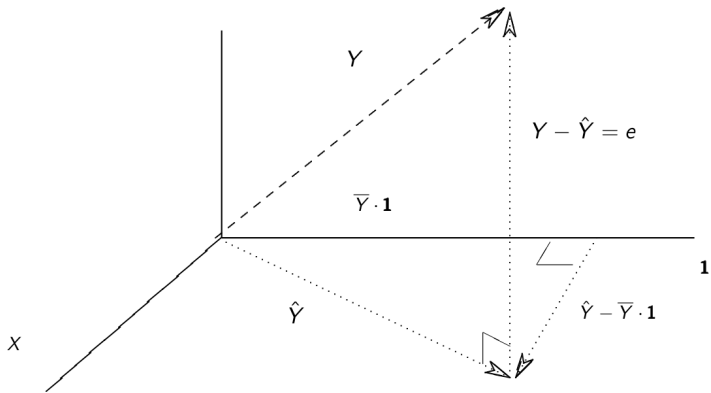
errors ε_i are independent $N(0, \sigma^2)$.

- In our “prototypical” data example, we might want to now if there really is a linear association between `Wife = Y` and `Husband = X`, *hypothesis test* of $H_0 : \beta_1 = 0$. This assumes the model above is correct, but that $\beta_1 = 0$.
- We might want to have a range of values that we can be fairly certain β_1 lies between: a *confidence interval* for β_1 .

Geometry of Least Squares

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University



Simple linear regression: distributions

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University

Distribution of $\hat{\beta}_1$

- Our assumptions tell us that

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)$$

- Therefore,

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma \sqrt{\frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2}}} \sim N(0, 1).$$

Standard error of $\hat{\beta}_1$

$$SE(\hat{\beta}_1) = \hat{\sigma} \sqrt{\frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2}} \quad \text{independent of } \hat{\beta}_1$$

Simple linear regression: testing

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University

t -test of $H_0 : \beta_1 = \beta_1^0$

- Suppose we want to test that β_1 is some pre-specified value, β_1^0 (this is often 0: i.e. is there a linear association)
- Under $H_0 : \beta_1 = \beta_1^0$

$$\frac{\hat{\beta}_1 - \beta_1^0}{\hat{\sigma} \sqrt{\frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2}}} = \frac{\hat{\beta}_1 - \beta_1^0}{\hat{\sigma} \cdot \sigma \sqrt{\frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2}}} \sim t_{n-2}.$$

- Reject $H_0 : \beta_1 = \beta_1^0$ if $|T| > t_{n-2, 1-\alpha/2}$.

Why reject for large $|T|$?

- Observing a large $|T|$ is unlikely if $\beta_1 = \beta_1^0$: reasonable to conclude that H_0 is false.
- Common to report p -value = $\mathbb{P}(T_{n-2} > |T|)$.

Confidence intervals based on t distribution

Generic setup

- Suppose we have a parameter estimate $\hat{\theta} \sim N(\theta, \tilde{\sigma}^2)$, and standard error $SE(\hat{\theta})$ such that

$$\frac{\hat{\theta} - \theta}{SE(\hat{\theta})} \sim t_{\nu}.$$

- $(1 - \alpha) \cdot 100\%$ confidence interval:

$$\hat{\theta} \pm SE(\hat{\theta}) \cdot t_{\nu, 1-\alpha/2}.$$

- Why? Expand absolute value as we did for the one-sample CI

$$1 - \alpha = \mathbb{P} \left(\left| \frac{\hat{\theta} - \theta}{SE(\hat{\theta})} \right| < t_{\nu, 1-\alpha/2} \right)$$

Confidence intervals for regression parameters

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University

Interval for β_1

- $(1 - \alpha) \cdot 100\%$ confidence interval:

$$\hat{\beta}_1 \pm SE(\hat{\beta}_1) \cdot t_{n-2, 1-\alpha/2}.$$

Interval for regression line $\beta_0 + \beta_1 \cdot X$

- $(1 - \alpha) \cdot 100\%$ confidence interval:

$$\hat{\beta}_0 + \hat{\beta}_1 X \pm SE(\hat{\beta}_0 + \hat{\beta}_1 X) \cdot t_{n-2, 1-\alpha/2}$$

where

$$SE(a_0 \hat{\beta}_0 + a_1 \hat{\beta}_1) = \hat{\sigma} \sqrt{\frac{a_0^2}{n} + \frac{(a_0 \bar{X} - a_1)^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}$$

Forecasting interval

Predicting a new observation

- Suppose we want an interval that will contain the height of the wife in a new couple sampled from the population where the husband has height X_{new} , i.e. an interval that will cover

$$Y_{\text{new}} = \beta_0 + \beta_1 X_{\text{new}} + \varepsilon_{\text{new}}$$

with a certain probability.

-

$$SE(\hat{\beta}_0 + \hat{\beta}_1 X_{\text{new}} + \varepsilon_{\text{new}}) = \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(\bar{X} - X_{\text{new}})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}.$$

- Prediction interval is

$$\hat{\beta}_0 + \hat{\beta}_1 X_{\text{new}} \pm t_{n-2, 1-\alpha/2} \cdot SE(\hat{\beta}_0 + \hat{\beta}_1 X_{\text{new}} + \varepsilon_{\text{new}})$$

Goodness of fit

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University

Sums of squares

$$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

$$SSR = \sum_{i=1}^n (\bar{Y} - \hat{Y}_i)^2 = \sum_{i=1}^n (\bar{Y} - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

$$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2 = SSE + SSR$$

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \widehat{Cor}(\mathbf{X}, \mathbf{Y})^2.$$

If R^2 is large: a lot of the variability in \mathbf{Y} is explained by \mathbf{X} .

Sums of Squares

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University

