Statistics 191: Introduction to Applied Statistics

Simple linear regression

Jonathan Taylor
Department of Statistics
Stanford University

February 22, 2010

Outline

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

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.

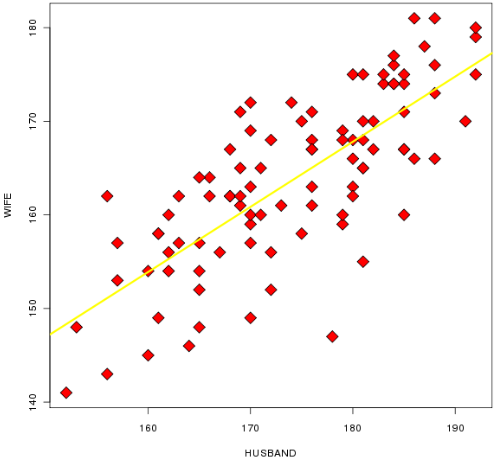
Mathematical formulation

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

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

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

Height data



Regression models

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:

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

  for some unknown parameter vector (β_0, β_1).
- Could also be a sum (linear combination) of *known* functions of `Husband`:

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

Specifying the (statistical) model

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

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

- 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

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

Jonathan Taylor
Department of Statistics
Stanford University

Why Least Squares?

- With least squares, the minimizers have explicit formulae – not so important with today's computer power – especially when L is convex.
- Resulting formulae are *linear* in the outcome Y. This is important for inferential reasons. For only predictive power, this is also not so important.
- If assumptions are correct, then this is "maximum likelihood" estimation.
- Some statistical theory tells us the "maximum likelihood" estimators are generally pretty good estimators.

Least squares

Alternative definition of (sample / population) mean

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

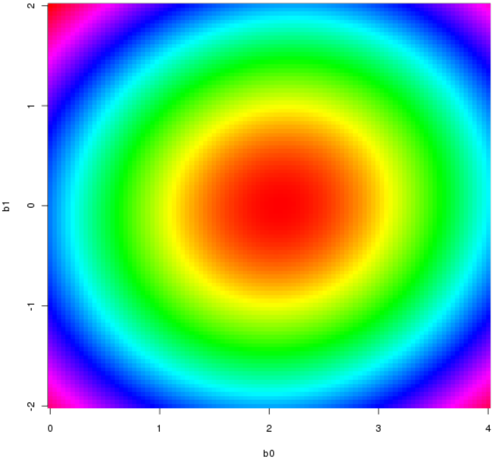
$$SSE(\mu) = \sum_{i=1}^{n} (Y_i - \mu)^2 \qquad \left(\text{population:} \quad = \mathbb{E}((Y - \mu)^2)\right).$$

Alternative definition of (sample / population) median

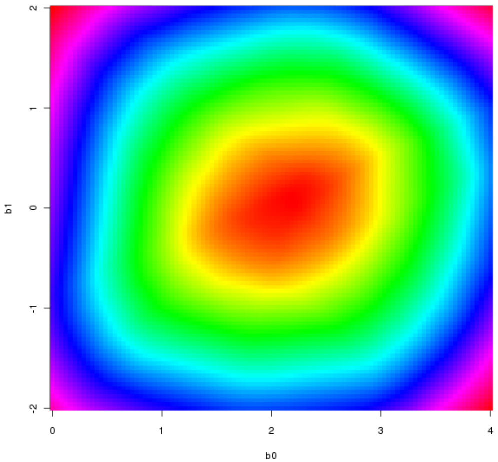
The median of a sample $(Y_1, \dots, Y_n)$ (or population Y) is any number that minimizes

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

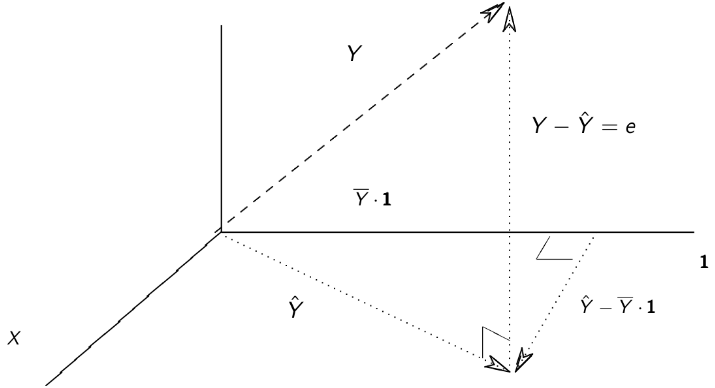
Least squares



Absolute deviation



Geometry of Least Squares



Least Squares Solutions

Regression line parameters: (β_0, β_1)

- $$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2} = \frac{\widehat{Cov}(X, Y)}{\widehat{Var}(X)}.$$
- $$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}.$$

Estimating variance: σ^2

- Strength of association between Y and X will depend on variability of errors ε, as in two sample t-test.
- $$\widehat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)^2 = \frac{SSE}{n-2} = MSE.$$

Least Squares

Predicting the mean

(Conditional) mean can be estimated for any given husband of height X as

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

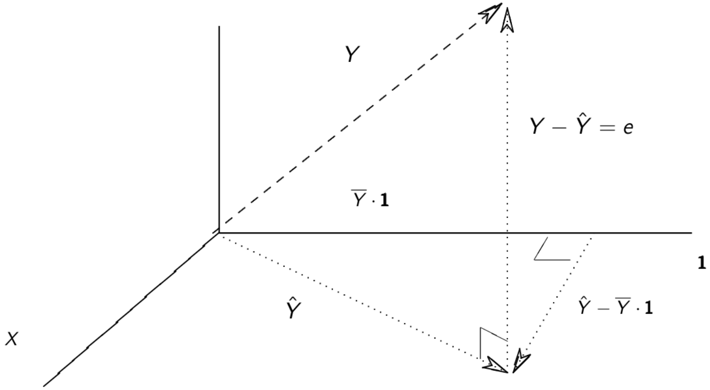
where $(\widehat{\beta}_0, \widehat{\beta}_1)$ are the minimizers of SSE.

Estimate of σ^2

- $$\widehat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \left(Y_i - \widehat{f}(X_i)\right)^2 = \frac{1}{n-2} \sum_{i=1}^n \left(Y_i - \widehat{Y}_i\right)^2.$$
- Why $n-2$? According to our statistical model

$$\frac{\widehat{\sigma}^2}{\sigma^2} \sim \frac{\chi^2_{n-2}}{n-2}.$$

Geometry of Least Squares



Inference

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:
  - *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, later 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 $\boldsymbol{Z} = (Z_1, \dots, Z_n)$, $\boldsymbol{W} = (W_1, \dots, W_m)$ 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

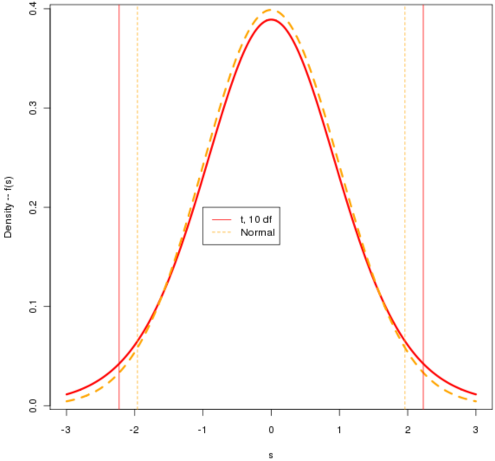
- Usually, we test a null hypothesis, H_0 based on some test statistic T whose distribution is fully known when H_0 is true.

t-statistics

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

t vs. Normal



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 can prove that the random variables

$$\overline{Z} \sim N(\mu, \sigma^2/n), \qquad \frac{S^2(Z)}{\sigma^2} \sim \frac{\chi^2_{n-1}}{n-1}$$

  are independent.
- Therefore

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

Confidence intervals

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 $\boldsymbol{Z}$ and $\boldsymbol{W}$ 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

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 $q = t_{n-1,(1-\alpha/2)}$

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

Inference in regression

Heights example

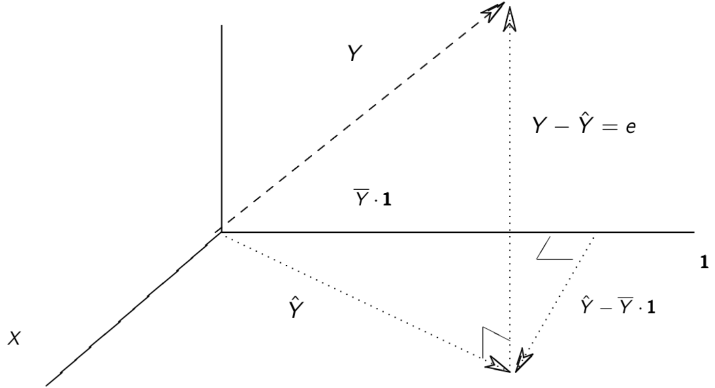
- Model:

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

  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



Simple linear regression: setup for inference

Geometry

- Let L be the subspace of $\mathbb{R}^n$ spanned $\mathbf{1} = (1, \dots, 1)$ and $\boldsymbol{X} = (X_1, \dots, X_n)$.
- Then,

$$\boldsymbol{Y} = P_L \boldsymbol{Y} + (\boldsymbol{Y} - P_L \boldsymbol{Y}) = \widehat{\boldsymbol{Y}} + \boldsymbol{e}$$

- In our model, if $\mu = \beta_0 \mathbf{1} + \beta_1 \boldsymbol{X}$ then

$$\widehat{\boldsymbol{Y}} = \mu + P_L \boldsymbol{\varepsilon}, \qquad \boldsymbol{e} = P_{L^\perp} \boldsymbol{Y} = P_{L^\perp} \boldsymbol{\varepsilon}$$

- Our assumption that ε_i's are independent $N(0, \sigma^2)$ tells us that (don't worry about proving this)
  - $\boldsymbol{e}$ and $\widehat{\boldsymbol{Y}}$ are independent
  - $\widehat{\sigma}^2 = \|\boldsymbol{e}\|^2/(n-2) \sim \sigma^2 \cdot \chi^2_{n-2}/(n-2)$.

Simple linear regression: distributions

Distribution of $\widehat{\beta}_1$

- Our assumptions tell us that

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

- Therefore,

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

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

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

Simple linear regression: testing

Statistics 191: Introduction to Applied Statistics

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 $H_0 : \beta_1 = \beta_1^0$

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

- 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 $\widehat{\theta} \sim N(\theta, \widetilde{\sigma}^2)$, and standard error $SE(\widehat{\theta})$ such that
$$\frac{\widehat{\theta} - \theta}{SE(\widehat{\theta})} \sim t_{\nu}.$$
- $(1-\alpha) \cdot 100\%$ confidence interval:
$$\widehat{\theta} \pm SE(\widehat{\theta}) \cdot t_{\nu, 1-\alpha/2}.$$
- Why? Expand absolute value as we did for the one-sample CI
$$1 - \alpha = \mathbb{P}\left(\left|\frac{\widehat{\theta} - \theta}{SE(\widehat{\theta})}\right| < t_{\nu, 1-\alpha/2}\right)$$

Confidence intervals for regression parameters

Jonathan Taylor
Department of Statistics
Stanford University

Interval for β_1

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

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

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

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

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

where

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

Forecasting (prediction) interval

Jonathan Taylor
Department of Statistics
Stanford University

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.
- $$SE(\widehat{\beta}_0 + \widehat{\beta}_1 X_{\text{new}} + \varepsilon_{\text{new}}) = \widehat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(\overline{X} - X_{\text{new}})^2}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}}.$$
- Prediction interval is
$$\widehat{\beta}_0 + \widehat{\beta}_1 X_{\text{new}} \pm t_{n-2,1-\alpha/2} \cdot SE(\widehat{\beta}_0 + \widehat{\beta}_1 X_{\text{new}} + \varepsilon_{\text{new}})$$