STAT 203: Regression Models and ANOVA
Lecture 2: Simple Linear Regression

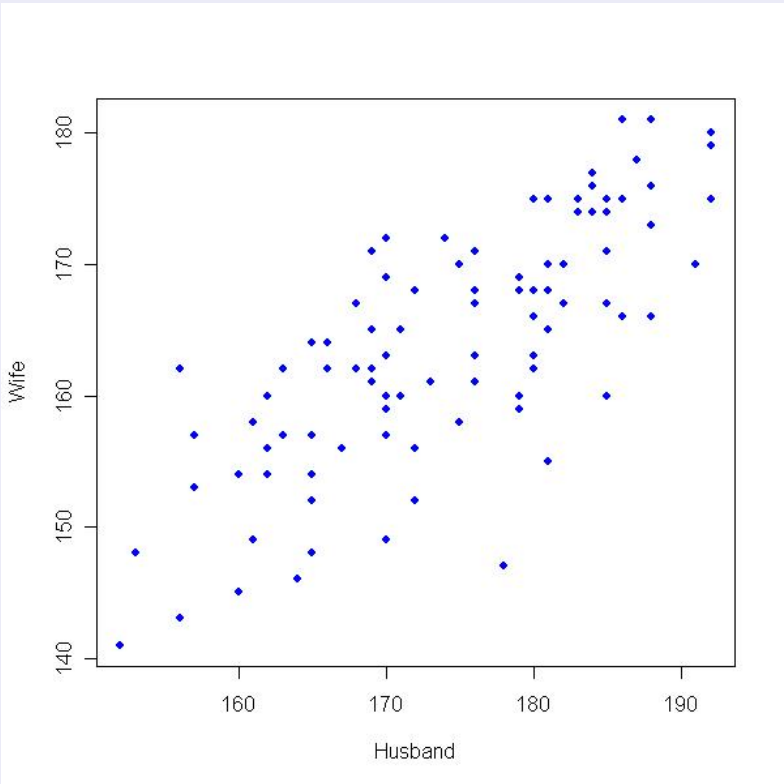
Nancy R. Zhang

Statistics 203, Stanford University

January 8, 2009

Last lecture...

Example: heights of husbands and wifes

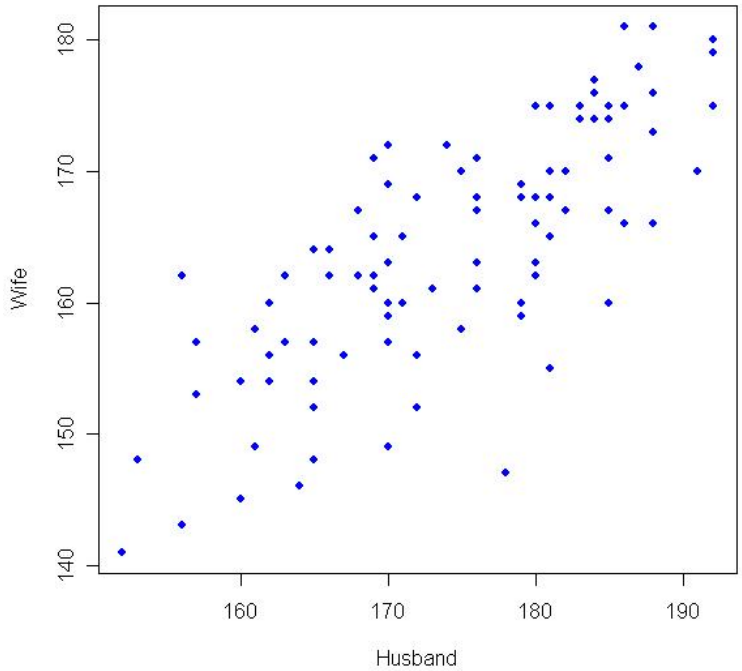


Do people of similar heights tend to marry each other?

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

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

Problem setup



We make n paired observations on two variables:

$$x_1, y_1$$

$$x_2, y_2$$

...

$$x_n, y_n$$

The objective is to test for a linear relationship between them.

$$y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{“predictable”}} + \underbrace{\varepsilon_i}_{\text{“random” error}} .$$

How to quantify a “good fit"

The least squares approach: choose β to minimize:

$$L(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

$L(\beta)$, sometimes called the loss function, is the sum of squared errors, or SSE.

Why use the least squares loss function?

Actually, in the early days of Statistics, regression analysis started off by using the absolute value norm:

$$L_1(\beta) = \sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 x_i|.$$

Doesn’t this seem more natural? There is no correct answer.

Squared Error or Absolute Value?

Alternative definition of (sample / population) mean

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

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

Alternative definition of (sample / population) median

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

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

In vector notation:

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \mathbf{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

$$L(\beta) = (y - \beta_0 \mathbf{1} - \beta_1 x)'(y - \beta_0 \mathbf{1} - \beta_1 x).$$

This is a convex function in β. Thus, it is minimized by solving $L'(\beta) = 0$.

Least Squares Solutions

$$L(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

or

$$L(\beta) = \|y - \beta_0 \mathbf{1} - \beta_1 x\|^2.$$

Solving $L'(\beta) = 0$, you get:

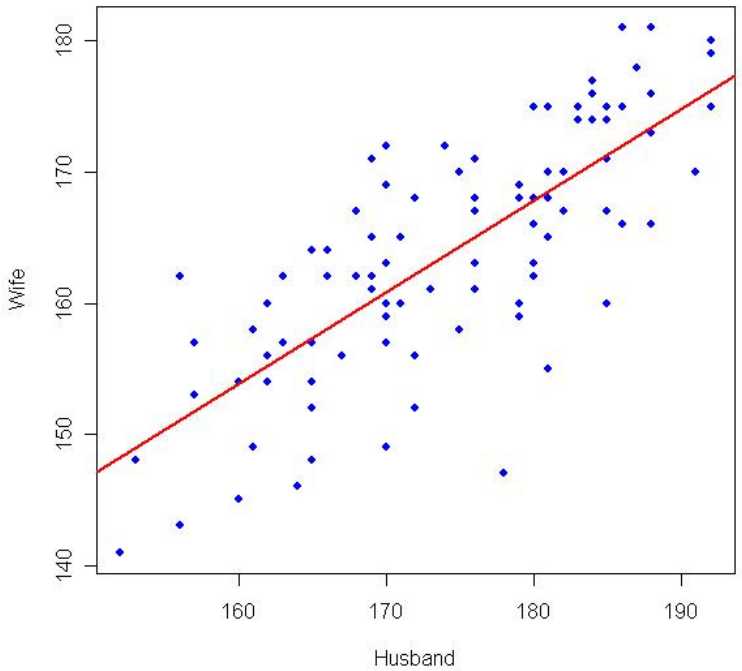
Regression line parameters: (β_0, β_1)

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

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}.$$

Prediction of y

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$, $\hat{Y}_i$ is the model fitted value for sample i.



Residuals

$$r_1 = y_1 - \hat{y}_1,$$

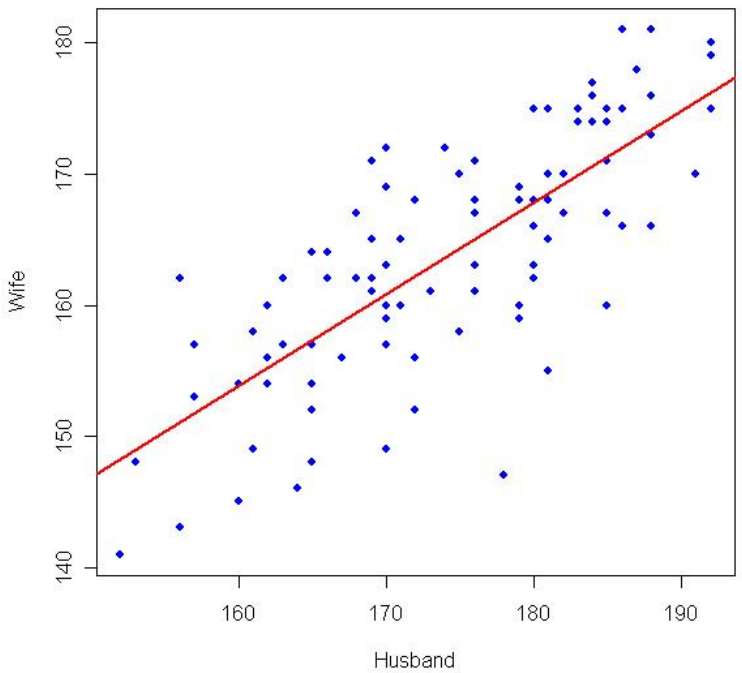
$$r_2 = y_2 - \hat{y}_2,$$

$$\ldots$$

$$r_n = y_n - \hat{y}_n,$$

Rule of Thumb

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$, $\hat{Y}_i$ is the model fitted value for sample i.



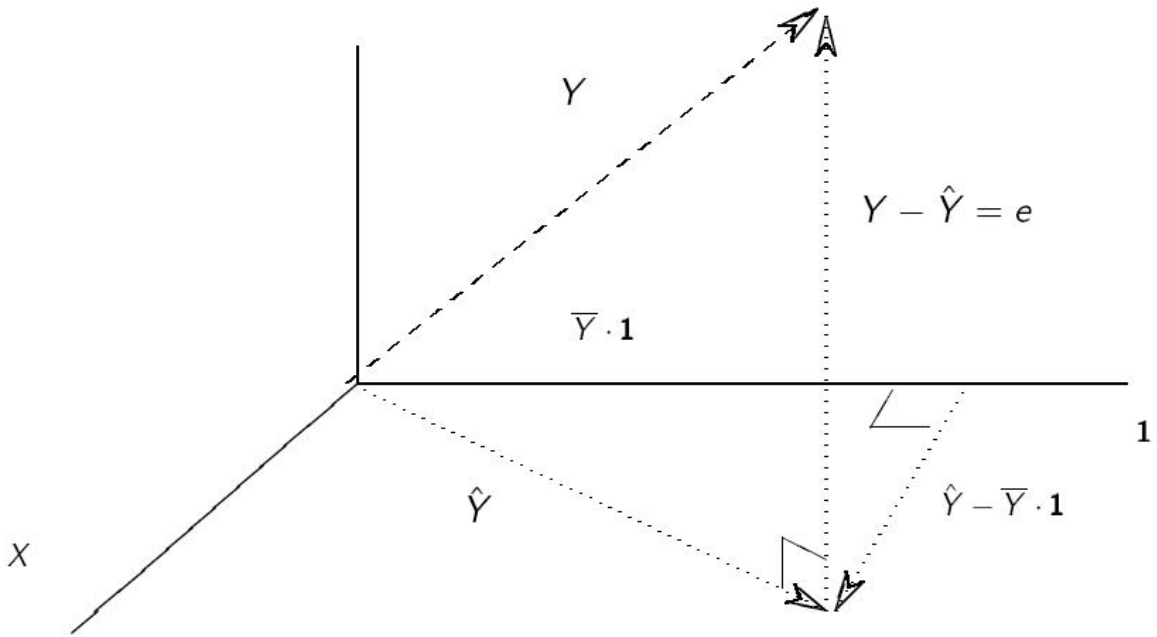
The *least squares* regression line is the line that goes through $(\bar{x}, \bar{y})$ and have slope

$$\frac{\langle x - \bar{x}, y - \bar{y} \rangle}{\{x\}^2}.$$

Note the asymmetry in x and y... why?

Geometry of least squares

$$L(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$



The squared-error loss function allows a simple geometric interpretation.

Good things about squared error loss

1. Least squares loss yields *analytical* solution.
2. The estimates $\hat{\beta}$ and $\hat{y}$ are *linear* in y.

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

$$\hat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 X$$

   We will see later how this makes inference easy.
3. For ϵ_i iid Gaussian, least squares yields maximum likelihood estimates.

What is maximum likelihood?

$$y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{“predictable”}} + \underbrace{\varepsilon_i}_{\text{“random” error}} .$$

If we assume that the errors are $N(0, \sigma^2)$, then

$$y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2).$$

Then, the entire data set has probability:

$$P_{\beta_0,\beta_1,\sigma^2}(y_1, \ldots, y_n | x_1, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}} \quad (1)$$

Now, find β_0, β_1 to maximize the probability of the observed data... you get the least squares regression estimates $\hat{\beta}_0$, $\hat{\beta}_1$.

Estimating Variance

Under the simple model $Y_1, Y_2, \ldots, Y_n$ are i.i.d. $N(\mu, \sigma^2)$, an estimate for σ^2 is

$$\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

The summation part, $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ has $\sigma^2\chi_2$ distribution with $n-1$ degrees of freedom.

Now, the regression model is

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$$

Estimate of σ^2

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

Note that the denominator is $n-2$ instead of $n-1$, why?

Inference on β

Regression line parameters: (β_0, β_1)

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

Inference:

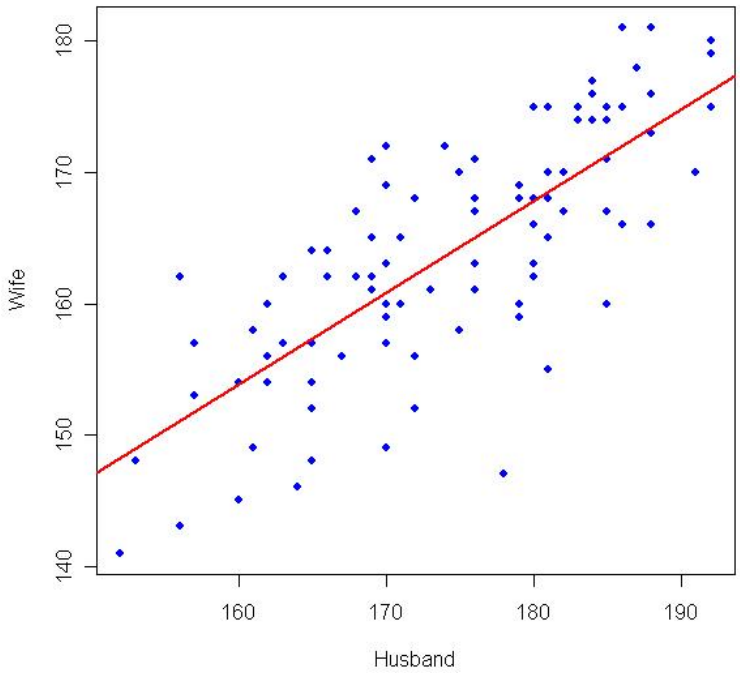
1. Hypothesis testing for β_1,

$$H_0 : \beta_1 = 0, \quad H_A : \beta_1 \neq 0.$$

2. Confidence intervals for β_1, i.e. find $[a, b]$ such that

$$P(\beta_1 \in [a, b]) = 1 - \alpha.$$

Review: Hypothesis Tests



We formulate the *null* hypothesis:

$$H_0 : \beta_1 = 0,$$

and calculate a test statistic based on $\hat{\beta}_1$, such that *assuming that H_0 is true*, the test statistic has a known distribution. If the statistic is too far in the tails of this distribution, we *reject* the null.

Distribution of $\hat{\beta}$

Regression line parameters: (β_0, β_1)

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

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}.$$

Important property: $\hat{\beta}_1$, $\hat{\beta}_0$ are *linear* functions of Y.

Linear combinations of Gaussian variables are Gaussian

If $X \sim N(\mu, \sigma^2)$, then $Y = aX + b \sim N(a\mu + b, a^2\sigma^2)$.

$\hat{\beta}_1$ is Gaussian with mean β_1, and variance $\frac{\sigma^2}{\|X\|^2}$.

$\hat{\beta}_1$ is Gaussian with mean β_1, and variance $\frac{\sigma^2}{\|X\|^2}$.

In the above formula, we do not know σ^2. This was the variance of the errors in the model

$$y_i = \beta_0 + \beta_1 X_i + \varepsilon_i.$$

However, from last lecture,

Estimate of σ^2

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

Thus, the standard errors of $\hat{\beta}_1$ is:

$s.e.(\hat{\beta}_1)$

$$s.e.(\hat{\beta}_1) = \frac{\hat{\sigma}}{\|X\|} = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

t-test

It can also be shown that $\hat{\beta}_1$ and $\hat{\sigma}^2$ are independent. Thus, under the *null* hypothesis $H_0 : \beta_1 = 0$,

$$\frac{\hat{\beta}_1}{s.e.(\beta_1)} \sim t_{n-2}.$$

t_{n-2} is the Student's t distribution with $(n-2)$ degrees of freedom. The level-α test for the two-sided alternative rejects when $\|\hat{\beta}_1\| > t_{(n-2,\alpha/2)}$.