STAT 513 fa 2020 Lec 07 slides

Simple linear regression

Karl B. Gregory

University of South Carolina

These slides are an instructional aid; their sole purpose is to display, during the lecture, definitions, plots, results, etc. which take too much time to write by hand on the blackboard. They are not intended to explain or expound on any material.

Regression model

For data pairs $(Y_1, x_1), \dots, (Y_n, x_n)$, suppose

$$Y_i = f(x_i) + \varepsilon_i$$

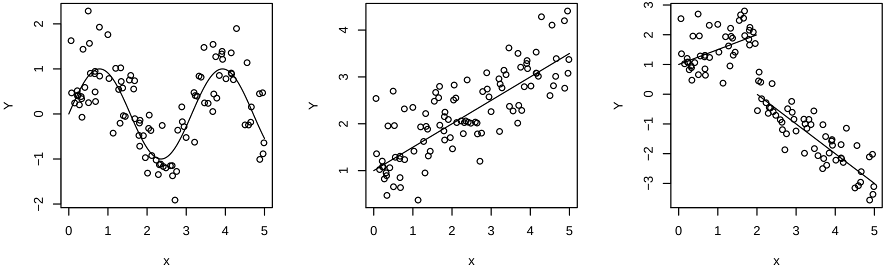
for $i = 1, \dots, n$, where

- $x_1, \dots, x_n$ are fixed real numbers
- $Y_1, \dots, Y_n$ are independent random variables
- $f : \mathbb{R} \to \mathbb{R}$ is an unknown function
- $\varepsilon_1, \dots, \varepsilon_n$ are iid rvs called *errors* with
  - $\mathbb{E}\varepsilon_i = 0$
  - $\operatorname{Var} \varepsilon_i = \sigma^2$

  for $i = 1, \dots, n$.

Goal: Estimate the unknown function f and the error variance σ^2.

We observe a function plus noise:



Simple linear regression model

For data pairs $(Y_1, x_1), \dots, (Y_n, x_n)$, suppose

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

for $i = 1, \dots, n$, where

- $x_1, \dots, x_n$ are fixed real numbers
- $Y_1, \dots, Y_n$ are independent random variables
- β_0 and β_1 are unknown constants
- $\varepsilon_1, \dots, \varepsilon_n$ are iid errors with
  - $\mathbb{E}\varepsilon_i = 0$
  - $\text{Var}\,\varepsilon_i = \sigma^2$

  for $i = 1, \dots, n$.

Goal: Estimate the unknown constants β_0 and β_1 and the error variance σ^2.

Topics:

1. Estimation of β_0, β_1, and σ^2 as well as of $\beta_0 + \beta_1 x_{\text{new}}$.
2. Inference about β_0 and β_1, e.g. testing H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$. Also confidence intervals for $\beta_0 + \beta_1 x_{\text{new}}$.
3. Prediction of Y_{new} of a "new" obs. $(x_{\text{new}}, Y_{\text{new}})$ with a *prediction interval*.
4. Likelihood approach under Normal errors.

Least-squares estimators of simple linear regression coefficients

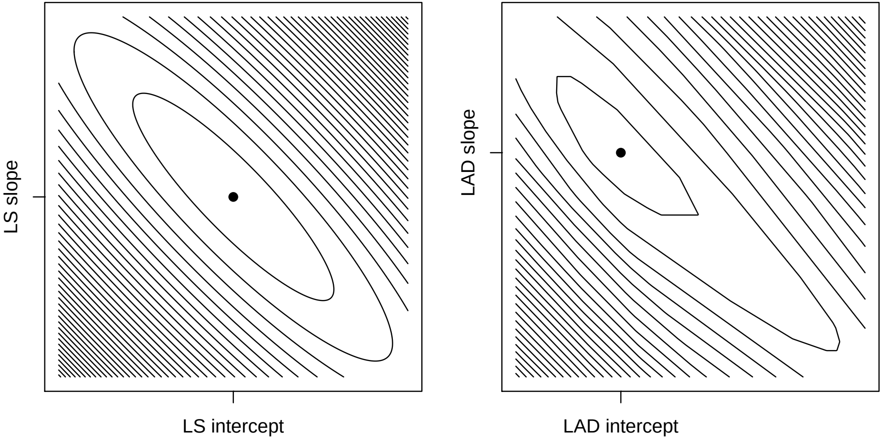
Provided $\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 > 0$, the function

$$Q_n(\beta_0, \beta_1) := \sum_{i=1}^{n}[Y_i - (\beta_0 + \beta_1 x_i)]^2$$

is (uniquely) minimized at

$$\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{x}_n$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}$$

Exercise: Derive this result.


LS slope
LS intercept
LAD slope
LAD intercept

Define some new quantities:

$$S_{xY} = \sum_{i=1}^{n}(x_i - \bar{x}_n)(Y_i - \bar{Y}_n), \quad S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x}_n)^2, \quad S_{YY} = \sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2,$$

$$r_{xY} = \frac{S_{xY}}{\sqrt{S_{xx}S_{YY}}}, \quad s_Y = \frac{S_{YY}}{n-1}, \quad s_X = \frac{S_{xx}}{n-1}.$$

Then we have the following simpler expressions for $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{S_{xY}}{S_{xx}} \quad \text{or} \quad \hat{\beta}_1 = r_{xY}\left(\frac{S_{YY}}{S_{xx}}\right)^{1/2} \quad \text{or} \quad \hat{\beta}_1 = r_{xY}(s_Y/s_x).$$

Exercise: Generate a toy data set in R and plot the least-squares line.

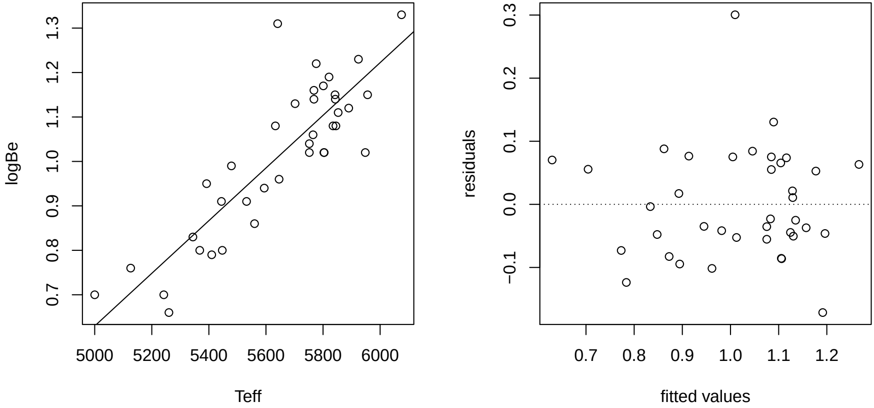
- The *fitted values* are

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad \text{for } i = 1, \dots, n.$$

- The *residuals* are

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i \quad \text{for } i = 1, \dots, n.$$

Log of beryllium abundance versus temperature of 38 stars with least-squares line.



Data from [1].

Some moments of the least-squares estimators

We have $\mathbb{E}\hat{\beta}_0 = \beta_0$ and $\mathbb{E}\hat{\beta}_1 = \beta_1$ as well as

$$\begin{aligned}
\text{Var}\,\hat{\beta}_0 &= (n^{-1} + \bar{x}_n^2 S_{xx}^{-1})\sigma^2 \\
\text{Var}\,\hat{\beta}_1 &= S_{xx}^{-1}\sigma^2 \\
\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) &= -\bar{x}_n S_{xx}^{-1}\sigma^2.
\end{aligned}$$

Exercise: Derive these, beginning by showing that $\hat{\beta}_0$ and $\hat{\beta}_1$ can be written as

$$\begin{aligned}
\hat{\beta}_0 &= \beta_0 + \frac{\bar{x}_n}{S_{xx}} \sum_{i=1}^{n} \left[\frac{S_{xx}}{n\bar{x}_n} - (x_i - \bar{x}_n) \right] \varepsilon_i \\
\hat{\beta}_1 &= \beta_1 + \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}_n)\varepsilon_i.
\end{aligned}$$

Unbiased estimator of σ^2

The estimator

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

is unbiased for σ^2.

The proof is omitted. The best way to prove this is with matrix algebra.

Mean and variance of estimated function at a point

We have

$$\mathbb{E}(\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}}) = \beta_0 + \beta_1 x_{\text{new}}$$
$$\text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}}) = \left[\frac{1}{n} + \frac{(x_{\text{new}} - \bar{x}_n)^2}{S_{xx}}\right]\sigma^2.$$

Exercise: Derive the above.

Sampling distribution results under Normal errors

If $\varepsilon_1, \dots, \varepsilon_n \overset{\text{ind}}{\sim} \text{Normal}(0, \sigma^2)$, then

$$\begin{aligned}
\hat{\beta}_0 &\sim \text{Normal}(\beta_0, (n^{-1} + \bar{x}_n^2 S_{xx}^{-1})\sigma^2) \\
\hat{\beta}_1 &\sim \text{Normal}(\beta_1, S_{xx}^{-1}\sigma^2) \\
\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}} &\sim \text{Normal}(\beta_0 + \beta_1 x_{\text{new}}, [n^{-1} + S_{xx}^{-1}(x_{\text{new}} - \bar{x}_n)^2]\sigma^2) \\
(n-2)\hat{\sigma}^2/\sigma^2 &\sim \chi^2_{n-2}.
\end{aligned}$$

Moreover, the above gives

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}\sqrt{n^{-1} + \bar{x}_n^2 S_{xx}^{-1}}} \sim t_{n-2}, \qquad \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}\sqrt{S_{xx}^{-1}}} \sim t_{n-2},$$

$$\text{and} \quad \frac{\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}} - (\beta_0 + \beta_1 x_{\text{new}})}{\hat{\sigma}\sqrt{n^{-1} + S_{xx}^{-1}(x_{\text{new}} - \bar{x}_n)^2}} \sim t_{n-2}.$$

Confidence intervals

We may construct $(1-\alpha)100\%$ CIs for β_0, β_1, and $\beta_0 + \beta_1 x_{\text{new}}$ as

$$\hat{\beta}_0 \pm t_{n-2,\alpha/2}\hat{\sigma}\sqrt{n^{-1} + \bar{x}_n^2 S_{xx}^{-1}}$$

$$\hat{\beta}_1 \pm t_{n-2,\alpha/2}\hat{\sigma}\sqrt{S_{xx}^{-1}}$$

$$(\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}}) \pm t_{n-2,\alpha/2}\hat{\sigma}\sqrt{n^{-1} + S_{xx}^{-1}(x_{\text{new}} - \bar{x}_n)^2}$$

Exercise: Get the `beryllium data` and under $\alpha = 0.05$:

1. Build CIs for β_0 and β_1.
2. Build CIs for $\beta_0 + \beta_1 x_{\text{new}}$ across a range of x_{new} values and plot them.

Testing hypotheses about β_1

Consider testing hypotheses about β_1 with respect to a null value β_1^*, and define

$$T_{1,n} = \frac{\hat{\beta}_1 - \beta_1^*}{\hat{\sigma}\sqrt{S_{xx}^{-1}}}.$$

We have the following:

H_0	H_1	Reject H_0 at α iff	p-value
$\beta_1 \leq \beta_1^*$	$\beta_1 > \beta_1^*$	$T_{1,n} > t_{n-2,\alpha}$	$1 - F_{t_{n-2}}(T_{1,n})$
$\beta_1 \geq \beta_1^*$	$\beta_1 < \beta_1^*$	$T_{1,n} < -t_{n-2,\alpha}$	$F_{t_{n-2}}(T_{1,n})$
$\beta_1 = \beta_1^*$	$\beta_1 \neq \beta_1^*$	$\lvert T_{1,n}\rvert > t_{n-2,\alpha/2}$	$2[1 - F_{t_{n-2}}(\lvert T_{1,n}\rvert)]$

Exercise: Get the p-value for testing H_0: $\beta_1 = 0$ for the beryllium data.

Prediction interval for a new observation

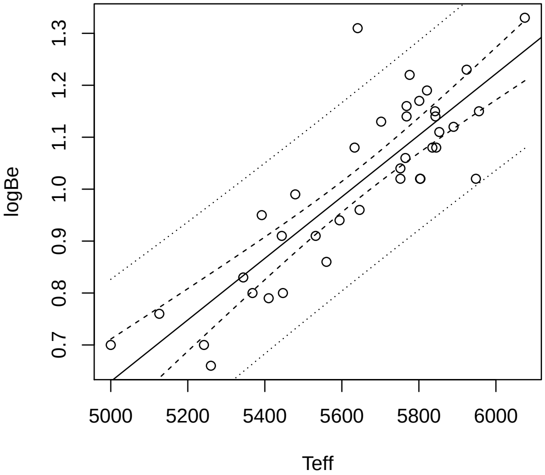
A $(1-\alpha)\times 100\%$ prediction interval for Y_{new} of a new obs. $(Y_{\text{new}}, x_{\text{new}})$ is given by

$$\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}} \pm t_{n-2,\alpha/2}\hat{\sigma}\sqrt{1 + n^{-1} + S_{xx}^{-1}(x_{\text{new}} - \bar{x}_n)^2}.$$

Exercise: Derive the above using the distribution of $\hat{\varepsilon}_{\text{new}} = Y_{\text{new}} - (\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}})$.

Exercise: With the Beryllium data, construct PIs over a range of x_{new} values.

CIs for the height of the regression function as well as PIs for new obs.



Data from [1].

Exercise: Let

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n,$$

where $\varepsilon_1, \dots, \varepsilon_n \overset{\text{ind}}{\sim} \text{Normal}(0, \sigma^2)$.

1. Give the likelihood function for β_0, β_1, and σ^2.
2. Give the log-likelihood function for β_0, β_1, and σ^2.
3. Show that the size-α LRT for H_0: $\beta_1 = 0$ vs H_1: $\beta_1 \neq 0$ is

$$\text{Reject } H_0 \text{ iff } S_{xx}^{1/2}|\hat{\beta}_1|/\hat{\sigma} > t_{n-2,\alpha/2}.$$

Nuno C Santos, G Israelian, RJ García López, M Mayor, R Rebolo, S Randich, A Ecuvillon, and C Domínguez Cerdeña.
Are beryllium abundances anomalous in stars with giant planets?
Astronomy & Astrophysics, 427(3):1085–1096, 2004.