Stat 206: Linear Models

Lecture 2

Sept. 30, 2019

Simple Linear Regression Model

n **cases** (trials/subjects): Y_i – the value of the response variable in the ith case; X_i – the value of the predictor variable in the ith case.

- **Model equation**:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad i = 1, \ldots, n. \tag{1}$$

- **Model assumptions**:
  - ϵ_is are uncorrelated, zero-mean, equal-variance random variables:

$$E(\epsilon_i) = 0, \ \ Var(\epsilon_i) = \sigma^2, \ \ i = 1, \ldots, n$$

$$\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0, \ \ 1 \leq i \neq j \leq n.$$

- **Unknown parameters**:
  - β_0 – regression intercept; β_1 – regression slope
  - σ^2: error variance

Given X_is, the distributions of the responses Y_is have the following properties:

- The response Y_i is the sum of two terms:
  - The mean of Y_i:
    $$E(Y_i) = \beta_0 + \beta_1 X_i,$$
    which is non-random.
  - The random error ϵ_i, which has zero-mean.
- ϵ_is have constant variance $\Longrightarrow$ Y_is have the same constant variance (regardless of the values of X_i):
  $$\mathrm{Var}(Y_i) = \sigma^2, \quad i = 1, \cdots, n.$$
- ϵ_is are uncorrelated $\Longrightarrow$ Y_is are uncorrelated:
  $$\mathrm{Cov}(Y_i, Y_j) = 0, \quad 1 \leq i \neq j \leq n.$$

In summary, the simple linear regression model says that the responses Y_i are

- random variables
- whose means are linear in X_i
- whose variances are a constant.
- Moreover, two responses Y_i and Y_j $(i \neq j)$ are uncorrelated.

Are the distributions of the responses Y_i fully specified by this model?

Regression Function

$$y = \beta_0 + \beta_1 x$$

- A straight line.
- β_1 is the slope of the regression line: the change in $E(Y)$ per unit change of X.
- β_0 is the intercept of the regression line: the value of $E(Y)$ when $X = 0$.

We will study how to model and fit the regression function from data.

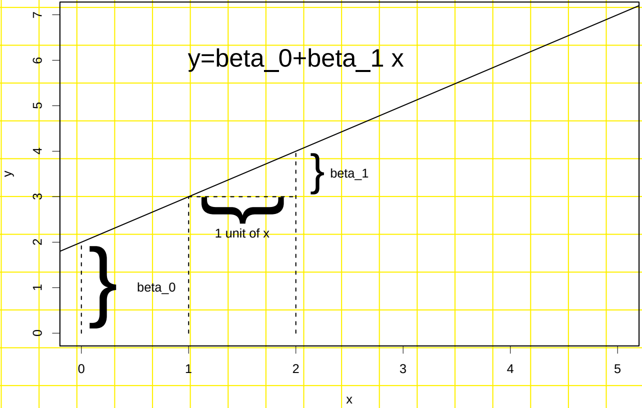


Figure: Regression line: $y = \beta_0 + \beta_1 x$

Least Squares Principle

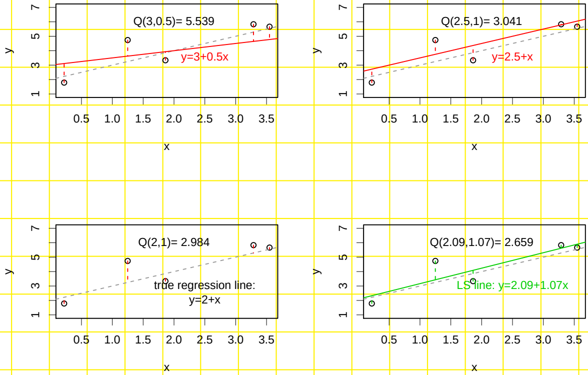
For a given line: $y = b_0 + b_1 x$, the *sum of squared vertical deviations* of the observations $\{(X_i, Y_i)\}_{i=1}^{n}$ from the corresponding points on the line is:

$$Q(b_0, b_1) = \sum_{i=1}^{n} (Y_i - (b_0 + b_1 X_i))^2 .$$

- $(X_i, b_0 + b_1 X_i)$ is the point on the line with the same x-coordinate as the ith observation point (X_i, Y_i).
- The *least squares (LS) principle* is to fit the observed data by minimizing the sum of squared vertical deviations.

LS line has the smallest sum of squared vertical deviations among all straight lines.

Figure: Illustration of LS principle



Which line has the smaller sum of squared vertical deviations, the LS line (a.k.a. the fitted regression line) or the true regression line?

Least Squares Estimators

LS estimators of β_0, β_1 are the pair of values b_0, b_1 that minimize the function $Q(\cdot, \cdot)$:

$$(\hat{\beta}_0, \hat{\beta}_1) = \mathrm{argmin}_{b_0, b_1} Q(b_0, b_1).$$

- LS estimators:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = r_{XY}\frac{s_Y}{s_X}, \qquad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} \quad (2)$$

- $\overline{X} = 1/n \sum_{i=1}^{n} X_i$, $\overline{Y} = 1/n \sum_{i=1}^{n} Y_i$ are the sample means.

Could you write down the formula for sample correlation r_{XY} and sample standard deviations s_Y, s_X?

- If X_is are all equal , then LS estimators do not exist! Though this is rare in practice.
- If the data are centered such that $\overline{X} = 0$, $\overline{Y} = 0$, then $\hat{\beta}_0 = 0$ and the LS line passes the origin $(0, 0)$. (Recall the “exam score” example.)

How to derive the LS Estimators?

The values of b_0, b_1 that minimize the function Q satisfy:

$$\frac{\partial Q(b_0, b_1)}{\partial b_0} = 0, \quad \frac{\partial Q(b_0, b_1)}{\partial b_1} = 0.$$

This leads to the **normal equations**:

$$\begin{aligned} nb_0 + b_1 \sum_{i=1}^{n} X_i &= \sum_{i=1}^{n} Y_i \\ b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 &= \sum_{i=1}^{n} X_i Y_i \end{aligned}$$

Can you solve these two equations with respect to b_0, b_1?

Fitted Values

- The **fitted regression line (LS line)**:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x = \overline{Y} + \hat{\beta}_1 (x - \overline{X}). \tag{3}$$

  - The fitted regression line passes through the point $(\overline{X}, \overline{Y})$, i.e., the *center of the data*.

- The **fitted value** for the *ith* case:

$$\widehat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = \overline{Y} + \hat{\beta}_1 (X_i - \overline{X}), \quad i = 1, \cdots n.$$

Residuals

Residuals are differences between the observed values Y_i and their respective fitted values $\widehat{Y}_i$:

$$\begin{aligned} e_i &= Y_i - \widehat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i), \quad i = 1, \cdots n. \\ &= (Y_i - \overline{Y}) - \hat{\beta}_1 (X_i - \overline{X}). \end{aligned}$$

- The residual e_i is an "estimator" of the respective error term: $\epsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$.
- Properties of residuals: (i) $\sum_{i=1}^n e_i = 0$; (ii) $\sum_{i=1}^n X_i e_i = 0$; (iii) $\sum_{i=1}^n \widehat{Y}_i e_i = 0$.
  What are geometric interpretation of these properties?

A Simulation Example

This is a simulated data set with $n = 5$ cases and

$$Y_i = 2 + X_i + \epsilon_i, \quad i = 1, \cdots, 5,$$

where ϵ_i are generated as i.i.d. $N(0, 1)$. *What is the true regression function and what is the true error variance σ^2?*

case i	X_i	Y_i	$X_i - \overline{X}$	$Y_i - \overline{Y}$	$(X_i - \overline{X})^2$	$(X_i - \overline{X})(Y_i - \overline{Y})$
1	1.86	3.34	-0.17	-0.94	0.03	0.16
2	0.22	1.79	-1.81	-2.48	3.29	4.50
3	3.55	5.66	1.52	1.39	2.30	2.11
4	3.29	5.83	1.26	1.56	1.58	1.96
5	1.25	4.74	-0.78	0.47	0.61	-0.36
Column Sum	10.17	21.36	0.00	0.00	7.81	8.37

$$\overline{X} = 10.17/5 = 2.03, \quad \overline{Y} = 21.36/5 = 4.27, \quad \sum_{i=1}^{5}(X_i - \overline{X})^2 = 7.81, \quad \sum_{i=1}^{5}(X_i - \overline{X})(Y_i - \overline{Y}) = 8.37.$$

$$\hat{\beta}_1 = 8.37/7.81 = 1.07, \quad \hat{\beta}_0 = 4.27 - 1.07 \times 2.03 = 2.09.$$

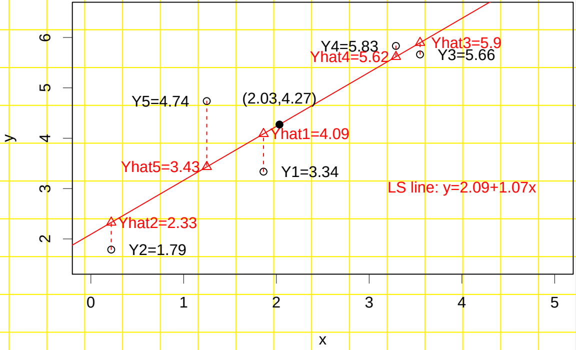
LS line: $y = 2.09 + 1.07x$.

Case i	X_i	Y_i	$\widehat{Y}_i$	e_i
1	1.86	3.34	4.09	-0.75
2	0.22	1.79	2.33	-0.54
3	3.55	5.66	5.90	-0.23
4	3.29	5.83	5.62	0.22
5	1.25	4.74	3.43	1.31

Example. $X_1 = 1.86$, $\widehat{Y}_1 = 2.09 + 1.07 \times 1.86 = 4.09$ and $e_1 = Y_1 - \widehat{Y}_1 = 3.34 - 4.09 = -0.75$.

Check the three properties of residuals.

Figure: LS line and fitted values



more example

Estimation of Error Variance by MSE

- Recall $\sigma^2 = \mathrm{Var}(\epsilon_i)$, so it is reasonable to estimate σ^2 by the "variance" of the residuals e_i.
- **Error sum of squares (SSE)**:

$$
\begin{aligned}
SSE := \sum_{i=1}^{n} e_i^2 &= \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 \\
&= \sum_{i=1}^{n} (Y_i - \overline{Y})^2 - \hat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \overline{X})^2.
\end{aligned}
$$

  - $E(SSE) = (n-2)\sigma^2$. *Could you derive this?*
  - The **degrees of freedom** of SSE is $n - 2$.
  - Two degrees of freedom are lost in estimating β_0, β_1.

- **Mean squared error (MSE)**:

$$s^2 = MSE = \frac{SSE}{n-2}, \quad E(MSE) = \sigma^2. \tag{4}$$

So MSE is an *unbiased estimator* of σ^2.

- *Do you know what does it mean to be an unbiased estiamtor?*
- *What are the similarities with and differences from the estimation of the variance of a single population based on an i.i.d. sample?*

Simulation Example (Cont'd)

$$SSE = (-0.75)^2 + (-0.54)^2 + (-0.23)^2 + 0.22^2 + 1.31^2 = 2.6715$$

and $n = 5$, so

$$MSE = \frac{2.6715}{5-2} = 0.8905.$$

What would be a reasonable estimator for σ? Would it be unbiased?

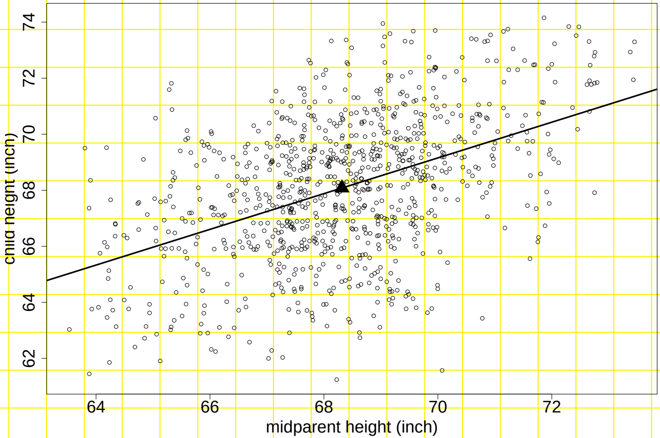
Notes: by Jensen's inequality, $\sqrt{MSE}$ would be an underestiamte for σ.

Heights

Summary statistics:
$n = 928$, $\overline{X} = 68.316$, $\overline{Y} = 68.082$, $\sum_i X_i^2 = 4334058$, $\sum_i Y_i^2 = 4307355$, $\sum_i X_i Y_i = 4318152$. Thus

$$\begin{aligned}
& \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y} \\
= \quad & 4318152 - 928 \times 68.316 \times 68.082 = 1936.738 \\
& \sum_{i=1}^{n} (X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - n(\overline{X})^2 \\
= \quad & 4334058 - 928 \times 68.316^2 = 3038.761. \\
& \hat{\beta}_1 = 1936.738/3038.761 = 0.637 \\
& \hat{\beta}_0 = 68.082 - 0.637 \times 68.316 = 24.54.
\end{aligned}$$

Figure: LS line of the heights data: $y = 24.54 + 0.637x$



```
Child Midparent
1 61.57220  70.07404
2 61.24382  68.22505
3 61.90968  65.12639
4 61.85769  64.23529
5 61.44986  63.88177
6 62.00005  67.02702
......
```

$X_1 = 70.07404$, $Y_1 = 61.57220$, $\widehat{Y}_1 = 24.54 + 0.637 \times 70.07404 = 69.17716$, $e_1 = 61.57220 - 69.17716 = -7.60496$.
$SSE = \sum_i e_i^2 = 4658.966$, $\quad n = 928$ so $MSE = \frac{4658.966}{928-2} = 5.031$.

back

Properties of LS Estimators

- **LS estimators are linear functions of the responses Y_is.**

$$\hat{\beta}_1 = \sum_{i=1}^{n} \frac{X_i - \overline{X}}{\sum_{j=1}^{n}(X_j - \overline{X})^2} Y_i = \sum_{i=1}^{n} k_i Y_i$$

$$\hat{\beta}_0 = \sum_{i=1}^{n} \left(\frac{1}{n} - \bar{X} k_i\right) Y_i.$$

- The fitted values $\widehat{Y_i}$ and the residuals e_i are also linear functions of the responses Y_is.

Can you write down their respective coefficients?

- **LS estimators are unbiased**: For **all** values of β_0, β_1,

$$E(\hat{\beta}_0) = \beta_0, \quad E(\hat{\beta}_1) = \beta_1.$$

Notes: Use the fact $E(Y_i) = \beta_0 + \beta_1 X_i, \quad i = 1, \cdots n.$

- Variances of $\hat{\beta}_0, \hat{\beta}_1$:

$$\begin{aligned} \sigma^2\{\hat{\beta}_0\} &= \sigma^2 \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2} \right] \\ \sigma^2\{\hat{\beta}_1\} &= \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2}. \end{aligned}$$

Notes: Use the fact that Y_is are uncorrelated.

Standard errors (SE) of the LS estimators.

- Replace σ^2 by MSE:

$$
\begin{aligned}
s^2\{\hat{\beta}_0\} &= MSE\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right], \\
s^2\{\hat{\beta}_1\} &= \frac{MSE}{\sum_{i=1}^n (X_i - \overline{X})^2}.
\end{aligned}
$$

- $s\{\hat{\beta}_0\}$ and $s\{\hat{\beta}_1\}$ are SE of $\hat{\beta}_0$ ad $\hat{\beta}_1$, respectively.

What are the implications?