Simple Linear Regression

Lecture 2

Alexandra Chronopoulou

ILLINOIS

COLLEGE OF LIBERAL ARTS & SCIENCES

Department of Statistics
101 Illini Hall, MC-374
725 S. Wright St.
Champaign, IL 61820-5710

Learning objectives

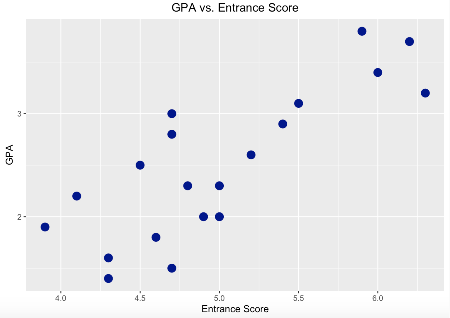
In this lecture we will:

- introduce the Simple Linear Regression Model (SLR).
- use the least-squares approach to estimate the model parameters.
- discuss goodness-of-fit and regression through the origin
- use R to fit a SLR model to the data.

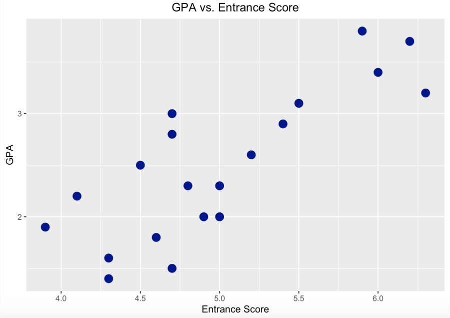
Simple Linear Regression

University Admissions Example

The director of admissions of a small college administered a newly designed entrance test to *20* students selected *at random* from the freshman class in a study to determine whether a student's grade point average (GPA) at the end of the freshman year (Y) can be predicted from the entrance test score (X).



University Admissions Example

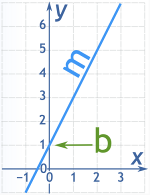


Based on the scatterplot:

- What conclusions do you draw?
- Which variable depends on the other?
- How can we initially describe the data? Is there a trend?

Equation of a Straight Line

– Mathematically, a straight line is defined as follows:



$$y = \underbrace{m}_{slope} x + \underbrace{b}_{intercept}$$

– The notation we use in regression is typically:

$$y = \underbrace{\beta_0}_{intercept} + \underbrace{\beta_1}_{slope} x$$

- One Response **Y**
- One Predictor **X**
- The data come in pairs:

$$\begin{array}{cc} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{array}$$

(n denotes the *total number of observations*)

Regression Model

- A *Regression Model* is a statistical relationship.
  - The Y's (*dependent variable*) tend to vary with the (*independent variable*) X in a systematic (*linear*) fashion.
  - There is a "scattering" of points around the statistical relationship.
- From the regression model describing the relation, we have that:
  - There is a probability distribution of Y for every level of X: *the probability of Y happening at that level of X*.
  - Each probability distribution of Y has a *mean* or "*center*".
  - The means of all distributions vary in some systematic fashion.

$\Rightarrow$ *Y is a RANDOM VARIABLE that has a distribution for every level of the independent variable.*

Simple Linear Regression (SLR) Model

Simple Linear Regression Model

$$\underset{\substack{\text{dep. variable} \\ \text{in } i\text{th trial}}}{\underset{\uparrow}{y_i}} = \beta_0 + \underset{\substack{\text{known constant,} \\ \text{has specific} \\ \text{value in } i\text{th trial}}}{\underset{\uparrow}{\beta_1 x_i}} \underset{\substack{\text{random} \\ \text{error}}}{\underset{\uparrow}{+\varepsilon_i}}$$

where the **intercept** β_0, the **slope** β_1, and the **error variance** σ^2 are the *model parameters.*

Model Assumptions

The **errors** $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are assumed to

- have **mean zero**: $\mathbb{E}(\varepsilon_i) = 0$
- be **uncorrelated**: $\mathrm{Cov}(\varepsilon_\mathrm{i}, \varepsilon_\mathrm{j}) = 0, \ \mathrm{i} \neq \mathrm{j},$
- be **homoscedastic**: $\mathrm{Var}(\varepsilon_\mathrm{i}) = \sigma^2$ does not depend on i.

Interpretation of β_1, β_0

- β_1 is the change in the mean of the probability distribution function of y *per unit change in* x.

- β_0 is the intercept, when $x = 0$. It is the mean of the probability distribution function of y (at $x = 0$) – this is the only time it has meaning. Otherwise β_0 has no particular meaning.

Least-Squares

Method of Least Squares

Model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \ldots, n$

Goal

- Estimate β_0, and β_1 and gain the *estimated* regression line.
- Consider the responses y_i and the *expected responses* $\mathbb{E}(y_i)$. We would like to minimize the difference between what we have (y_i) and what we expect $(\mathbb{E}(y_i))$, i.e.

$$\min\left[y_i - \mathbb{E}(y_i)\right] \Leftrightarrow \min\left[y_i - (\beta_0 + \beta_1 x_i)\right]$$

- Find estimates of β_0, β_1 to minimize this quantity.
- The best line will be closest to the actual data points.

Method of Least Squares

- The data we have are $(x_i, y_i)_i$.
- The quantity we want to minimize is

$$y_i - (\beta_0 + \beta_1 x_i)$$

  which can be positive or negative. So ...

- We minimize the *Residual Sum of Squares (RSS)* instead

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

  to obtain $(\hat{\beta}_0, \hat{\beta}_1)$.

$\Rightarrow$ Method of Least Squares!

Method of Least Squares

- To find the solution and obtain $(\hat{\beta}_0, \hat{\beta}_1)$, we have

$$\frac{\partial \mathrm{RSS}}{\partial \beta_0} = 0 \quad \Leftrightarrow -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$
$$\frac{\partial \mathrm{RSS}}{\partial \beta_1} = 0 \quad \Leftrightarrow -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)\, x_i = 0$$

- Re-arrange the equations to obtain

$$\beta_0 n + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$
$$\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

- Then, solve the system with respect to β_0, β_1.

Least Square Estimators

LS Estimators

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Alternative Representation of $\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} := \frac{S_{xy}}{S_{xx}} = r_{xy}\sqrt{\frac{S_{yy}}{S_{xx}}}$$

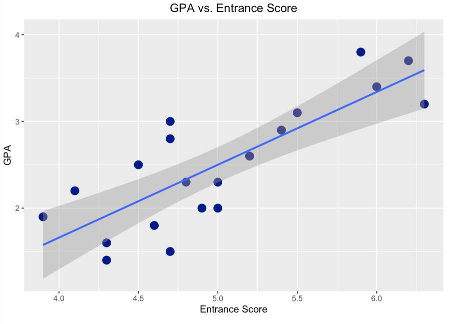
where $S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y})$, $S_{xx} = \sum_i (x_i - \bar{x})^2$, $S_{yy} = \sum_i (y_i - \bar{y})^2$

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \text{ the sample correlation}$$

University Admissions Example (Revisited)

In our previous example, the fitted regression line is:

$$(\text{GPA}) = 3.0539 + 0.7785 \cdot (\text{Entrance Score})$$



How should we interpret the line?

University Admissions Example (Revisited)

- Suppose you only have the following information available:

	Mean	Variance
Entrance Score	2.5	0.52
GPA	5	0.48

$$Corr(\text{GPA}, \text{Entrance Score}) = 0.81$$

- If you knew your entrance score was 4.3, could you *guess* your GPA?

Sample Correlation & Linear Regression

- The "unit-free, location/scale invariant" version of the *GPA* (Y) and the "unit-free, location/scale invariant" version of the *Entrance Score* (X) have the following relationship

$$\frac{y - \mu_y}{\sigma_y} \approx r_{xy} \frac{x - \mu_x}{\sigma_x}$$

- If we now take the *sample equivalent* of this expression, we get

$$\frac{y - \bar{y}}{\sqrt{S_{yy}}} \approx r_{xy} \frac{x - \bar{x}}{\sqrt{S_{xx}}}$$

- Re-arranging the terms

$$y \approx \underbrace{\left(\bar{y} - r_{xy}\sqrt{\frac{S_{yy}}{S_{xx}}}\,\bar{x}\right)}_{=\hat{\beta}_0} + \underbrace{\left(r_{xy}\sqrt{\frac{S_{yy}}{S_{xx}}}\right)}_{=\hat{\beta}_1} x$$

Fitted Values & Residuals

- Given $\hat{\beta}_0$, $\hat{\beta}_1$, the LS estimates of the regression coefficients, we call

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

  the fitted value (or predicted value) at x_i, or the prediction of y_i.

- The ith residual is the difference between y_i (observed value) and its prediction (fitted value):

$$r_i = y_i - \hat{y}_i$$

Residuals

Some Properties

1. $\sum_i r_i = 0$
2. $RSS = \sum_i r_i^2$ is a minimum
3. $\sum_i y_i = \sum_i \hat{y}_i$
4. $\sum_i x_i r_i = 0$
5. $\sum_i \hat{y}_i r_i = 0$
6. The regression line *always goes through the point* $(\bar{x}, \bar{y})$. (Why?)

The error variance is estimated by

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_i r_i^2$$

and the degrees of freedom (df) of the residuals is

$$(\text{sample size}) - (\#\text{ of parameters}) = n - 2$$

ILLINOIS

Goodness of Fit

Goodness of Fit: R-square

Total Variation Breakdown

The total variation in the response y, measured by the Total Sum of Squares (TSS), can be decomposed as follows:

$$\begin{aligned}\sum_i (y_i - \bar{y})^2 &= \sum_i (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 = \sum_i (r_i + \hat{y}_i - \bar{y})^2 \\ &= \sum_i r_i^2 + \sum_i (\hat{y}_i - \bar{y})^2 \\ TSS &= RSS + FSS\end{aligned}$$

the Residual Sum of Squares (RSS) and the Fitted value Sum of Squares (FSS).

Remark: The cross-product term vanishes due to orthogonality:

$$\sum_i r_i(\hat{y}_i - \bar{y}) = \hat{\beta}_0 \sum_i r_i + \hat{\beta}_1 \sum_i r_i x_i - \bar{y} \sum_i r_i = 0$$

ILLINOIS

Coefficient of Determination (R^2)

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = \frac{FSS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

- $0 \leq R^2 \leq 1$
- It measures the effect of X in reducing the variation in Y.
- The larger R^2 is, the more the total variation of y is reduced by reducing the independent variable x.
  - The closer R^2 is to 1, the greater the degree of *linear* association between X and Y
- We also have that $r_{xy} = \pm\sqrt{R^2}$, where the *sign* is the sign of the slope.

Misunderstandings

- R^2 measures the *relative reduction* of TSS
- Both R^2 and r_{xy} measure the degree of linear association (i.e. the actual relation may be curvilinear).
- In the 'University Admissions' Example, $R^2 = 0.6538$.
  What is your conclusion?

Affine Transformations

Affine Transformations of the Data

Suppose we have a SLR model of Y on X, i.e. $y_i = \beta_0 + \beta_1 x_i$.

- Rescale y_i by $\tilde{y}_i = ay_i + b$ and then regress $\tilde{y}_i$ on x_i.
  How would the LS estimates and R^2 be affected?

- Rescale x_i by $\tilde{x}_i = ax_i + b$ and then regress y_i on $\tilde{x}_i$.
  How would the LS estimates and R^2 be affected?

- Regress x on y instead, will the LS line be the same?
  How about R^2?

Regression Through the Origin

Birds Eggs Study

A study is conducted to understand the relationship between the height of a bird's egg and its weight. Based on the data collected, the following regression line was obtained:

$$\text{Height} = -1.774 + 1.444 \text{ Width}$$

- Is the intercept $\hat{\beta}_0 = -1.774$ meaningful here?
- Can we fit a model without an intercept? What does it change?

Reference: Kimber, H. (1995). The 'golden egg'. Teaching Statistics, 17(2), 34-7.

Regression Through the Origin

- Sometimes we want to fit a line with no intercept (a.k.a. regression through the origin):

$$y_i \approx \beta_1 x_i$$

- Using LS, we estimate β_1 by

$$\hat{\beta}_1 = \frac{\sum_i x_i y_i}{\sum_i x_i^2}$$

- The ordinary definition of R-square is *no longer meaningful*; A negative R-square is possible, since RSS may be larger than TSS.

Regression Through the Origin

In a model with no intercept, we have the following decomposition:

$$\sum_i y_i^2 = \sum_i (y_i - \hat{y}_i + \hat{y}_i)^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i \hat{y}_i{}^2$$

and a modified R-square

$$\tilde{R}^2 = 1 - \frac{RSS}{\sum_i y_i^2}$$

ILLINOIS