STAT2203 STAT7203

Probability models and data analysis

Lecture 9.2

Linear regression

Regression Analysis

In the two-sample problems discussed thus far, we only compared a parameter of population X with that of population Y. Even in the paired experiments, we never used information about X to infer something about Y. This is precisely the objective of **regression analysis**: to leverage the relationship between two (or more) variables and gain information about one of them by using the values of the other(s).

For example let X denote the high school GPA and Y be the college GPA. Typically, those with high (low) high school GPAs tend to also have high (low) college GPAs. Knowledge of a student's high school GPA is useful in predicting how well that person will do in college.

In regression studies, Y is called the **response variable**,and X is interchangeably referred to as the **covariate**, or the **independent variable**, or the **predictor**, or the **explanatory variable**.

Linear Probabilistic Model

The simplest deterministic mathematical relationship between two variables x and y is a linear relationship

$$y = \overbrace{\beta_0^\star}^{\text{intercept}} + \overbrace{\beta_1^\star}^{\text{slope}} x.$$

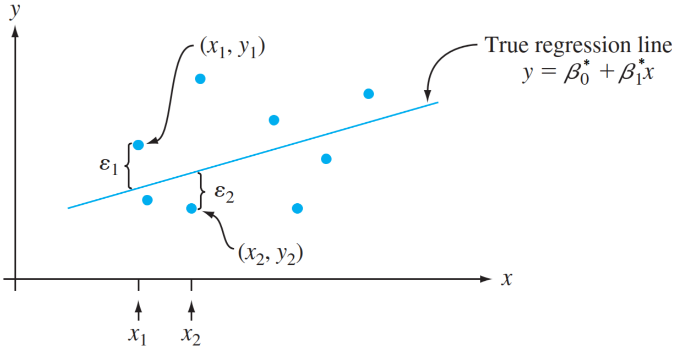
If the two variables are not deterministically related, then for a fixed value of x, there is uncertainty in the value of y. So we need to develop a linear probabilistic model for this relationship:

$$Y = \beta_0^\star + \beta_1^\star x + \mathcal{E},$$

where $\mathcal{E}$ is a random variable representing the random deviation.

Linear Probabilistic Model

Suppose, $\mathbb{E}(\mathcal{E}) = 0$, and $\text{Var}(\mathcal{E}) = \sigma^2$. Then, the points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ resulting from n independent observations, i.e., n independent realizations of $\mathcal{E}$ as $\varepsilon_1, \ldots, \varepsilon_n$, will then be scattered about the true regression line

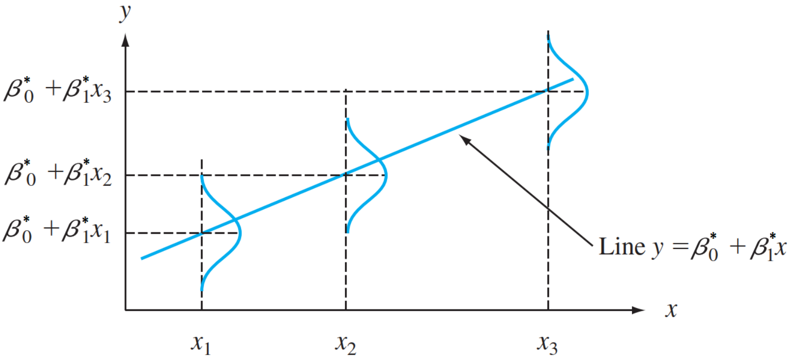


Simple Linear Regression Model

If $\mathcal{E} \sim \mathcal{N}(0, \sigma^2)$, then the probabilistic model is often called **simple linear regression model**.

For a fixed x, we have

$$Y \sim \mathcal{N}(\beta_0^\star + \beta_1^\star x, \sigma^2).$$

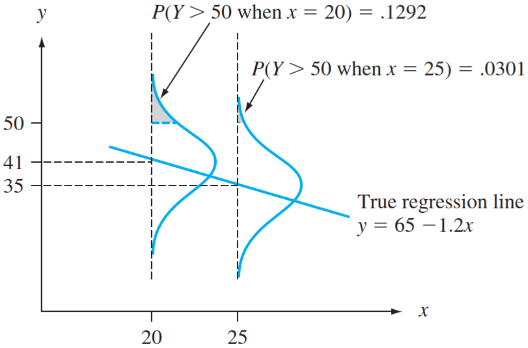


y
$\beta_0^* + \beta_1^* x_3$
$\beta_0^* + \beta_1^* x_2$
$\beta_0^* + \beta_1^* x_1$
Line $y = \beta_0^* + \beta_1^* x$
x
x_1
x_2
x_3

Example

Suppose $Y \sim \mathcal{N}(65 - 1.2x, 8^2)$. Then

$$\mathbb{P}(Y > 50 \text{ when } x = 20) = \mathbb{P}(Z > \frac{50 - 41}{8}) = 0.1292,$$
$$\mathbb{P}(Y > 50 \text{ when } x = 25) = \mathbb{P}(Z > \frac{50 - 35}{8}) = 0.0301.$$



$P(Y > 50$ when $x = 20) = .1292$
$P(Y > 50$ when $x = 25) = .0301$
y
50
41
35
True regression line
$y = 65 - 1.2x$
x
20
25

Estimation for Linear Regression

Obviously, we don't know $\beta_0^\star$, $\beta_1^\star$, or σ, but we can find a line that best "fits" the data, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

For a given β_0 and β_1, we get an estimate for y_i as

$$y_i(\beta_0, \beta_1) = \beta_0 + \beta_1 x_i.$$

Typically, $y_i(\beta_0, \beta_1) \neq y_i$, and so the goal is to find β_0 and β_1 such that the **residual error**, $e_i(\beta_0, \beta_1) = y_i(\beta_0, \beta_1) - y_i$, is small in some sense for all i.

One way is to find β_0 and β_1 such that the sum of the squared residual errors (SSE) is minimized:

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i(\beta_0, \beta_1) - y_i)^2.$$

Least Squares Regression

https://www.geogebra.org/m/xC6zq7Zv

Least Squares Regression: Estimates

Least Squares Regression: $\min_{\beta_0,\beta_1} \sum_{i=1}^{n}(\beta_0 + \beta_1 x_i - y_i)^2.$

The values for $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the least-squares criterion are called **least squares estimates** for $\beta_0^\star$ and $\beta_1^\star$, and are given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

The least squares estimate of σ^2 can be obtained by

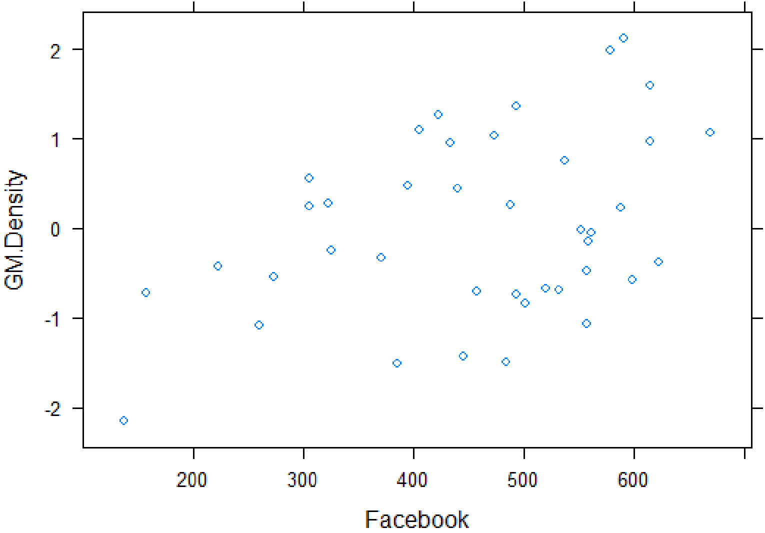
$$\hat{\sigma}^2 = \mathrm{MSE} = \frac{1}{n-2}\sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_i - y_i)^2 = \frac{\mathrm{SSE}}{n-2}.$$

Example

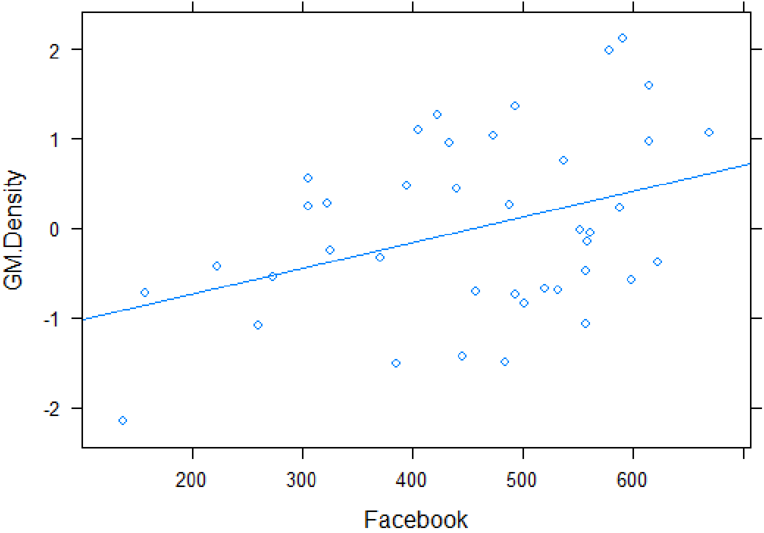
Researchers at University College London took a sample of 40 student volunteers and used MRI to measure grey matter (GM) density within three small volumes of the brain. They then looked at the association between these densities and the number of Facebook friends.

The results for one of the volumes, the left middle temporal gyrus (MTG), are shown in the following figure. The GM densities are given in standard units.

Example (continued)



Grey Matter = -1.311785 + 0.002886 Friends



Least Squares Regression: Estimators

Of course, $\hat{\beta}_0$ and $\hat{\beta}_1$ will change when we collect new samples. In other words, the values we obtained for them was just a realization of the underlying estimators[1] for $\beta_0^\star$ and $\beta_1^\star$ as

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

where

$$Y_i \sim \mathcal{N}(\beta_0^\star + \beta_1^\star x_i, \sigma^2), \quad \text{and} \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i.$$

What is the expectation and variance of these estimators?

[1] When lower case is used in statistics, both estimator and the estimate are denoted with the same lower case letter.

Distribution of the Estimators

Let's assume that the observations are modeled to be obtained *independently* of one another. That is,

$$Y_i \sim \mathcal{N}(\beta_0^\star + \beta_1^\star x_i, \sigma^2),$$

and

$$Y_1, Y_2, \ldots, Y_n \ \text{ are independent.}$$

As our estimators are linear combinations of independent normal random variables (Y_i), we know our estimators also have a normal distribution.

Unbiased Estimators

Recall an estimator is said to be unbiased if its expectation is equal to the quantity we wish to estimate.

$$\mathbb{E}\hat{\beta}_1 = \mathbb{E}\left(\frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) = \frac{\sum_{i=1}^n (x_i - \bar{x})(\mathbb{E}Y_i - \mathbb{E}\bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$= \frac{\sum_{i=1}^n (x_i - \bar{x})([\beta_0^\star + \beta_1^\star x_i] - [\beta_0^\star + \beta_1^\star \bar{x}])}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$= \frac{\beta_1^\star \sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta_1^\star.$$

You can also easily show that $\mathbb{E}\hat{\beta}_0 = \beta_0^\star$. It is also possible to show that $\mathbb{E}(\hat{\sigma}^2) = \sigma^2$, although showing this is somewhat harder.

Variance of Estimators

Using what we have learned about the variance of linear combinations of independent random variables, it is possible to show

$$\mathrm{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \right)$$

and

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \ .$$

Finally, we can also show that

$$\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x}\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Distribution of Estimators

So,

$$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0^\star, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)\right),$$
$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1^\star, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right).$$

Aside: In fact, one can show that, in the simple linear regression problem, the joint distribution of $(\hat{\beta}_0, \hat{\beta}_1)$ is the bivariate normal distribution for which the means, variances, and covariance are given as before.