MAT 2377 (Summer 2009) - June 23, 2009
Introduction to Simple Linear Regression
Sections 11.1-11.3

§11 Simple Linear Regression

§11.1-11.2 Regression Model

Introduction: We would like to analyze the relationship between two variables. Regression is the study of the relationship between a dependent variable Y and an independent variable X.

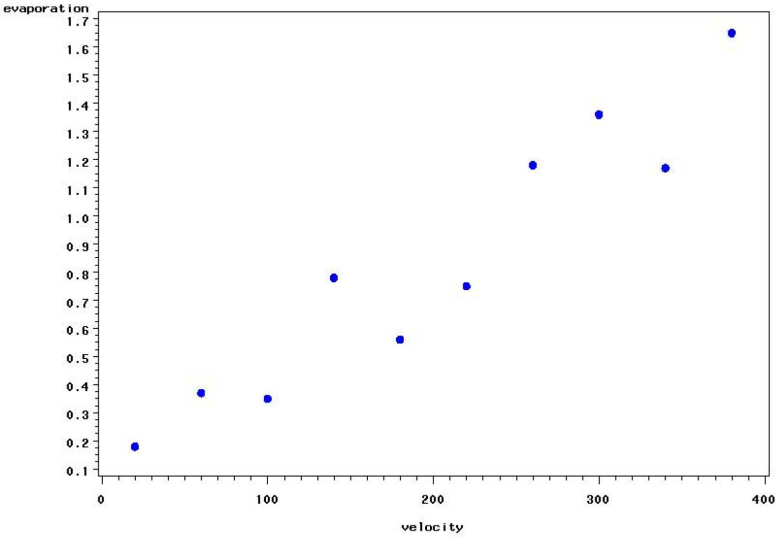
Example 0: In a chemical process the amount of the product is related to the process-operating temperature. Regression analysis can be used to predict the amount at a given temperature level!

Terminology:
Y which is the **dependent variable** is also called the **response variable** and X which is the **independent variable** will be called a **predictor variable**.

Example 1: The following are measurements of the air velocity and evaporation coefficient of burning fuel in an impulse engine:

air velocity	evaporation coefficient	air velocity	evaporation coefficient
cm/sec	mm^2/sec	cm/sec	mm^2/sec
x	y	x	y
20	0.18	220	0.75
60	0.37	220	0.75
100	0.35	260	1.18
140	0.78	300	1.36
180	0.56	340	1.17
		380	1.65

Here is a scatter diagram of y versus x.



Question: Does there appear to be a linear trend?

We will suppose that there exists a linear statistical relationship between the response Y and the predictor X. We can represent such a relationship with a **simple linear regression** model.

Simple Linear Regression Model is

$$Y = \beta_0 + \beta_1 \, x + \epsilon,$$

where β_0 and β_1 are unknown constants (or: regression coefficients), x is a value taken by the predictor X and ϵ is **random error**.

We will assume that ϵ is a random variable with mean 0 and variance σ^2. That is,

$$E(\epsilon) = 0 \qquad \text{and} \qquad V(\epsilon) = \sigma^2$$

Interpretation of the model:

Given a value x of the predictor variable X, Y is a random variable with mean

$$\mu_{Y|x} = E[Y|x] = \beta_0 + \beta_1 \, x.$$

Terminology:

$$\mu_{Y|x} = \beta_0 + \beta_1 \, x$$

is called the regression line with **intercept** β_0 and **slope** β_1. So, it is a line of $mean$ values!

Variation: Given a value x of the predictor, the variance of Y is

$$V(Y|x) = V(\epsilon) = \sigma^2. \qquad (WHY?)$$

Note: σ^2 is called the variance of the random error.

Terminology: β_0, β_1 and σ^2 are called parameters of the simple linear regression model.

In many real-life problems $\beta_0, \beta_1, \sigma^2$ are NOT known, so we must estimate them from sample data!

Estimation of the parameters:

Sample : We select a random sample of n paired observations:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n).$$

Assuming that the simple linear regression is appropriate, then we can express the observations as follows:

$$y_i = \beta_0 + \beta_1\, x_i + \epsilon_i, \qquad i = 1, 2, \ldots, n,$$

where ϵ_i represents the ith error (or deviation from the regression line).

We would like to find the line that "best" fits the data. We will use the sum of the squared deviations from the line, that is

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1\, x_i)\right]^2 ,$$

as a measure of distance from the line.

Least-Squares Estimation: This method of estimation consists of minimizing L with respect to β_0 and β_1, by solving

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\hat{\beta}_0,\hat{\beta}_1} = 0 \qquad \text{and} \qquad \left.\frac{\partial L}{\partial \beta_1}\right|_{\hat{\beta}_0,\hat{\beta}_1} = 0.$$

Simplifying these two equations give a system of linear equations called the **normal equations:**

$$\begin{aligned} n\,\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} y_i\, x_i \end{aligned}$$

The solutions of the normal equations are called the **least-squares estimates**.

The least-squares estimate of the slope and intercept are (respectively)

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\, \overline{x},$$

and so the **fitted** or **estimated regression line** is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1\, x.$$

Notation :

$$\overline{x} = \sum_{i=1}^{n} \frac{x_i}{n} \qquad \text{and} \qquad \overline{y} = \sum_{i=1}^{n} \frac{y_i}{n}$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \left(\sum_{i=1}^{n} x_i^2\right) - n\,\overline{x}^2 = \left(\sum_{i=1}^{n} x_i^2\right) - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

$$S_{xy} = \sum_{i=1}^{n} y_i\,(x_i - \overline{x}) = \sum_{i=1}^{n} (y_i - \overline{y})\,(x_i - \overline{x}) = \left(\sum_{i=1}^{n} x_i\, y_i\right) - n\,\overline{x}\,\overline{y} =$$

$$= \left(\sum_{i=1}^{n} x_i\, y_i\right) - \frac{\left(\sum_{i=1}^{n} x_i\right)\,\left(\sum_{i=1}^{n} y_i\right)}{n}$$

Example 2: Consider the data from Example 1. The $n = 11$ observations yielded
$\sum x_i = 2000, \quad \sum y_i = 8.35,$
$\sum x_i^2 = 532,000.0, \quad \sum y_i^2 = 9.1097$, and
$\sum x_i\, y_i = 2175.4.$

Suppose that the simple linear regression model is appropriate.

a) Determine the estimated regression line.

b) Estimate the mean evaporation coefficient when the air velocity is $x = 140.$

Sol: a) since $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1\, x$, we compute first $\beta_1 = \frac{S_{xy}}{S_{xx}}$ as follows:

$S_{xy} = (\sum_{i=1}^{11} x_i y_i) - n\bar{x}\bar{y} = 2175.4 - 11 \times \frac{2000}{11}\frac{8.35}{11} = 2175.4 - 1518.18 = 657.22$;

$S_{xx} = (\sum_{i=1}^{11} x_i^2) - n\bar{x}^2 = 532000.0 - 11 \times (\frac{2000}{11})^2 = 532000.0 - 363636.36 = 168363.64.$

Hence $\hat{\beta}_1 = \frac{657.22}{168363.64} = 0.0039 (\cong 0.004)$. To get the other parameter we compute:

$\hat{\beta}_0 = \overline{y} - \widehat{\beta_1}\, \overline{x} = \frac{8.35}{11} - 0.0039 \times \frac{2000}{11} = 0.05.$ We get $\hat{y} = 0.05 + 0.0039\, x.$

b) Just plug in $x = 140$ (see page 3 Interpretation of the model) and get $\hat{y} = 0.05 + 0.0039 \times 140 = 0.596.$ SEE THE PICTURE on page 2 and compare!!

NOTE: These ESTIMATES are subject to error!

Estimating the variance of the random error.

Let (x_i, y_i) be the ith pair of observed values in the sample.

We denote the evaluation of the estimated regression line at $x = x_i$, as

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

Note: $\hat{y}_i$ is called the ith fitted value.

The difference between y_i and $\hat{y}_i$ is called the ith **residual**.

Notation:

$$e_i = y_i - \hat{y}_i.$$

Consider the sum of the squared residuals, that is

$$SS_E = \sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

which is often called the **error sum of squares**.

It can be shown that $E(SS_E) = (n-2)\,\sigma^2$, which implies that

$$\hat{\sigma}^2 = \frac{SS_E}{n-2}$$

is unbiased for estimating σ^2.

Note: It is not necessary to compute each residual since there exist an alternate computational formula for SS_E.

Computational formula for SS_E :

$$SS_E = SS_T - \hat{\beta}_1 \, S_{xy}.$$

where

$$SS_T = S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2 = \left(\sum_{i=1}^{n} y_i^2 \right) - n \, \overline{y}^2 = \left(\sum_{i=1}^{n} y_i^2 \right) - \frac{\left(\sum_{i=1}^{n} y_i \right)^2}{n}.$$

Remark: We sometimes call SS_y the total variation, since it measures the variation among the responses $y_1, \ldots, y_n$.

§11.3 **Properties of the least-squares estimators:**
When the values of x are fixed, $\widehat{\beta_0}$ and $\widehat{\beta_1}$ depend on the observed y's.
The least-squares estimators for the slope and the intercept are respectively

$$\widehat{\beta_1} = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \overline{x}) \, Y_i \quad \text{and} \quad \widehat{\beta_0} = \overline{Y} - \widehat{\beta_1} \, \overline{x}.$$

Remark: Both estimators are linear combinations of the independent random variables

$$Y_1, \ldots, Y_n.$$

Thus, we can compute their expectation and variation.

Expectation:

$$E[\widehat{\beta_1}] = \beta_1 \qquad \text{and} \qquad E[\widehat{\beta_0}] = \beta_0$$

Variance:

$$V[\widehat{\beta_1}] = \frac{\sigma^2}{S_{xx}} \qquad \text{and} \qquad V[\widehat{\beta_0}] = \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} \right].$$

Remarks:

- The estimators $\widehat{\beta_1}$ and $\widehat{\beta_0}$ are unbiased estimators of β_1 and β_0, respectively.
- The standard deviation of the estimator (that we call standard error) allows us to measure the error in estimation:

  $$\sigma_{\widehat{\beta_1}} = \sqrt{V[\widehat{\beta_1}]} = \sqrt{\frac{\sigma^2}{S_{xx}}}$$

  and

  $$\sigma_{\widehat{\beta_0}} = \sqrt{V[\widehat{\beta_0}]} = \sqrt{\sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right]}.$$

- Since we do not know the true value of σ^2, we can estimate it with

  $$\hat{\sigma}^2 = \frac{SS_E}{n-2}.$$

Estimated standard errors:

$$\widehat{\sigma}_{\widehat{\beta_1}} = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}.$$

$$\widehat{\sigma}_{\widehat{\beta_0}} = \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right]}.$$

Recall:

$$\hat{\sigma}^2 = \frac{SS_E}{n-2} = \frac{S_{yy} - \hat{\beta}_1\, S_{xy}}{n-2}.$$

Exemple 3: Refer to Example 1 and Example 2.

(a) Compute the 2nd residual.
Recall: $x_2 = 60$ and $y_2 = 0.37$.

(b) Give a point estimate for σ^2.

(c) Give the estimated standard error for the estimation of the intercept and also for the estimation of the slope.

Sol: a) The 2nd residual is $e_2 = y_2 - \widehat{y_2} = 0.37 - (0.05 + 0.0039 \times 60) = 0.086$;

b) We compute: $\hat{\sigma}^2 = \frac{SS_E}{n-2} = \frac{SS_E}{11-2} = \frac{S_{yy} - \widehat{\beta_1} S_{xy}}{9} = \frac{(\sum_{i=1}^{11} y_i^2) - n\overline{y}^2 - \widehat{\beta_1} S_{xy}}{9} =$ $= \frac{9.1097 - 11 \times \frac{8.35^2}{11^2} - 0.0039 \times 657.22}{9} = 0.0231$; (see page 7 for the partial computations you need in this example!)

c) We compute: i) $\widehat{\sigma}_{\widehat{\beta_1}} = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = \sqrt{\frac{0.0231}{168363.64}} = 0.000370$,

ii) $\widehat{\sigma}_{\widehat{\beta_0}} = \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right]} = \sqrt{(0.0231)\left[\frac{1}{11} + 0.1963\right]} \cong 0.081453$.

Do 11-2 on page 399.

a) So they are asking for $\widehat{y}$. We have

$$i) \quad \widehat{\beta_1} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum_i y_i)}{n}}{(\sum_i x_i^2) - \frac{(\sum_i x_i)^2}{n}} =$$

$$= \frac{1083.67 - \frac{(1478)(12.75)}{20}}{(143215.8) - \frac{(1478)^2}{20}} = 0.00416$$

$ii) \quad \widehat{\beta_0} = \overline{y} - \widehat{\beta_1}\overline{x} = \frac{12.75}{20} - 0.00416 \times \frac{1478}{20} = 0.32999$, so $\widehat{y} = \widehat{\beta_0} + \widehat{\beta_1} x = 0.32999 + 0.00416x$.

Note that $S_{xy} = \sum_i x_i y_i - \frac{(\sum_i x_i)(\sum_i y_i)}{n} = 1083.67 - \frac{1478 \times 12.75}{20} = 141.445$.

Moreover, $\widehat{\sigma}^2 = \frac{SS_E}{n-2} = \frac{(\sum_i y_i^2) - \frac{(\sum_i y_i)^2}{20} - \widehat{\beta_1} S_{xy}}{18} = \frac{8.86 - \frac{12.75^2}{20} - 0.00416 \times 141.445}{18} = \frac{8.86 - 8.128125 - 0.5884112}{18} = 0.00797$, and do the graph....

b) $\widehat{y} = 0.32999 + 0.00416 \times 85 = 0.6836$;

c) $\widehat{y} = 0.32999 + 0.00416 \times 90 = 0.7044$;

d) THE SLOPE (think about the derivative...) is $\widehat{\beta_1} = 0.00416$. In other words: do you remember the approximation: $\frac{f(x+1) - f(x)}{x+1-x} \cong f'(x)$? What is the derivative of a linear function? The slope!