Simple Linear Regression

Recall: In the previous lectures we have discussed various types of ANOVA in which the dependence of a quantitative response on the levels of one (or more) categorical factors is studied. The ANOVA model represents the response as a sum of terms representing the factor levels and their interactions, e.g.,

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

Now we will investigate scenarios in which the relationship between two (or more) random variables can be described by a linear equation.

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Statistical linear regression is applied to cases where two (usually quantitative) variables are related, but do not determine each other:

Example:

- Age and height of a child.
- The highschool GPA and college GPA of a randomly selected student at San Jose State.
-

Notation: Usually, these experiments are designed to study one variable as a (random) function of the other: $Y = Y(x)$. The X-variable is referred to as the INDEPENDENT VARIABLE, the PREDICTOR, or the EXPLANATORY VARIABLE. The Y variable is commonly called the DEPENDENT VARIABLE or the RESPONSE.

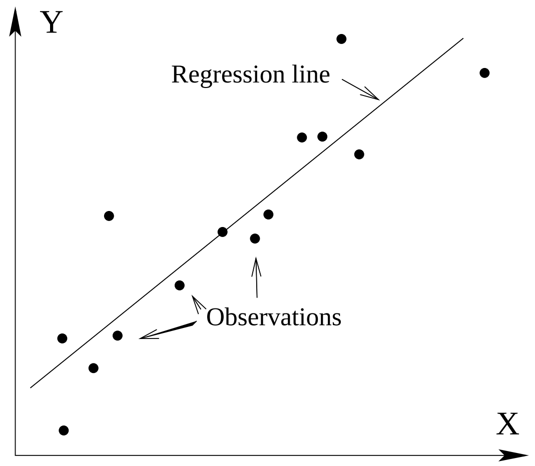
Graphical Display: The simplest graphical display method for two quantitative variables is a ________________________. It is convention to plot the independent variable x on the x-axis, and the dependent variable y on the y-axis.

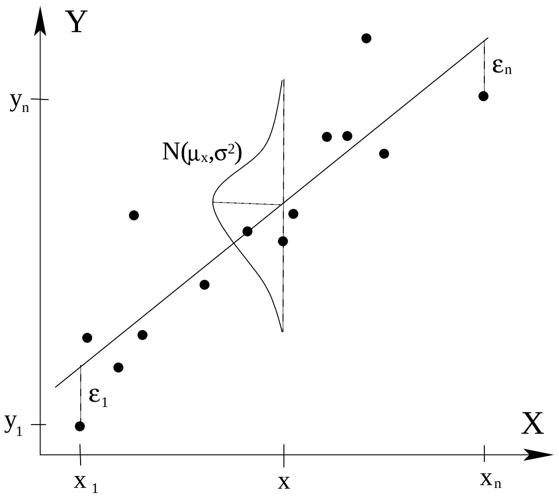
The Model: The simple linear regression model assumes that the observed variables X and Y have a linear relationship

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where the random variable ϵ respresents the error in the model, and we assume that the ϵ's are independent, normally distributed with mean zero and constant variance $V(\epsilon) = \sigma^2$.

If you draw a scatter-plot of the data and the points can be approximated by a line, then β_1 is the _______________ of the line and β_0 is the _______________.





Probabilistically, for a fixed level x of variable X, the response Y can be understood as observation(s) on a Normal random variable with mean

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

and variance σ^2. Note, that the variance of Y *does not* depend on the value of x.

Example: Suppose that in a certain chemical process the reaction time y (hr) is related to temperature x (°F) in the chamber in which the reaction takes place according to the simple linear regression model with equation

$$y = 5 - 0.01x, \qquad \text{and } \sigma = 0.075.$$

(a) What is the average reaction time if the process is carried out at 200°?

(b) What is the probability that the reaction time will exceed 150 minutes if the process is carried out at 200°?

(c) By how much do you expect the reaction time to change for every 1°F degree increase in temperature?

Estimating Model Parameters

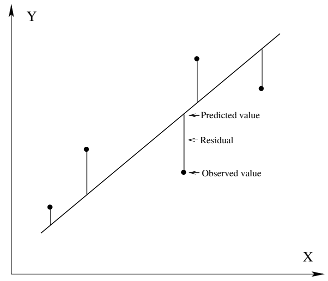
Assume that two observed variables X and Y are related according to the simple linear regression model

$$Y = \beta_0 + \beta_1 X + \epsilon$$

with $\epsilon \sim \text{Normal}(0, \sigma^2)$ uncorrelated.

In practice, the values of β_0, β_1 and σ^2 are generally unknown. Instead, sample data on X and Y are available: $(x_1, y_1), \ldots, (x_n, y_n)$. Our goal is to estimate the values of the model parameters from these observations.

Least Squares Regression Idea: We need a method to "fit" a line to the scatterplot of the observations. There are several ways this can be done. By far the most popular method is that of Least Squares. Consider the vertical distance between the observation and the regression line (Residual). This distance may be positive or negative, depending on whether the observation lies above or below the line. We will *fit* the regression line (i.e., choose the parameters β_0 and β_1), so that the sum of squares of these vertical distances is *minimized*.



Simple Calculus yields the parameter estimates of the model slope β_1 and intercept β_0:

$$b_1 = \hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$b_0 = \hat{\beta}_0 = \frac{\sum y_i - \hat{\beta}_1 \sum x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x}$$

Where computation formulas for S_{xy} and S_{xx} are

$$S_{xy} = \sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)/n, \qquad S_{xx} = \sum x_i^2 - \left(\sum x_i\right)^2/n$$

If you do not use software to analyze data for a regression problem, then it is efficient to first compute the data summaries $\sum x_i, \sum y_i, \sum x_i^2, \sum y_i^2$ and $\sum x_i y_i$ and to base the computation of the model parameters on these summary statistics.

Definition: Fitted or Predicted Values $\hat{y}_1, \ldots, \hat{y}_n$ for the response variable can be obtained by substituting the observed values of the dependent variable $x_1, \ldots, x_n$ into the regression equation.

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

The residuals ϵ_i can then be computed as $y_i - \hat{y}_i$.

Example: Italian (!!!) researchers studied the stretchability of mozarella cheese based on temperature. The measurements are x = temperature in °F and y = elongation (in %) of cheese just before failure:

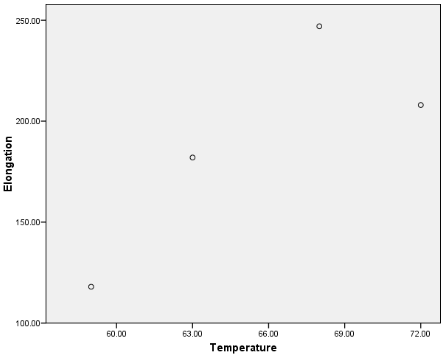
x	59	63	68	72
y	118	182	247	208

(a) Compute the summary statistics

$$\sum x_i = \qquad\qquad , \sum y_i = \qquad\qquad , \sum x_i^2 = \qquad\qquad , \sum y_i^2 =$$

$$\sum x_i y_i =$$

(b) Draw an approximate regression line into the scatterplot of the (few) observations in this example. Find the approximate slope of the line.



(c) Compute parameter estimates for the slope and intercept of your regression line.

Estimating σ^2

The residual variance parameter σ^2 describes the variability in the model. How far, on average, do the observations lie from the regression line (close to the line $\Rightarrow$ small σ^2; far from the line $\Rightarrow$ large σ^2). Recall that the residuals ϵ_i are the vertical distances from the observations to the regression line.

Definition: The error sum of squares or residual sum of squares is the sum of the squared residuals:

$$\text{SSE } = \sum (y_i - \hat{y}_i)^2 = \sum \epsilon_i^2$$

and the estimate of σ^2 is

$$\hat{\sigma}^2 = s^2 = \frac{\text{SSE}}{n-2}$$

Note: The degree of freedom associated with $\hat{\sigma}^2$ is $n-2$, because two degrees of freedom are necessary to estimate the slope and intercept.

For computation, the following formula may be more convenient. It depends on the data summaries that you have probably already computed in estimating the slope and intercept:

$$\text{SSE} = \sum y_i^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i$$

Example: (cont.)

(d) Compute the estimate of σ^2 for the previous Mozarella cheese example.

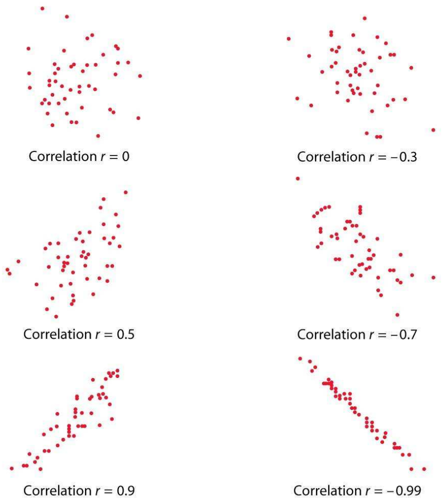
Correlation

Review: In Chapter 5 of your textbook, the concept of correlation between two random variables has been discussed. If more than one quantitative variable is observed in an experiment, it is often of interest to study the association between the variables.

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

This correlation coefficient measures the strength and direction of a linear relationship between two quantitative variables.

The sign of r describes the direction of the association. Positive correlation means that large x-values are associated (on average) with large y values. Negative correlation means that large x-values are associated (on average) with small y-values. The absolute value of r describes the strength of the association. Values close to 0 represent weak association, values close to ± 1 represent strong association.



Correlation $r = 0$

Correlation $r = -0.3$

Correlation $r = 0.5$

Correlation $r = -0.7$

Correlation $r = 0.9$

Correlation $r = -0.99$

Caution: Correlation does not imply causation!

Example: Researchers in Lower Saxony, Germany, have found a strong positive correlation between the number of storks breeding in a community and the number of babies born (per year) in the same community. Can you think of a reason?

The Coefficient of Determination

The total sum of squares can be used as a measure for the total variation in the data.

$$\text{SST } = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2$$

Recall, that the sum of squared residuals describes the amount of variation in the residuals:

$$\text{SSE } = \sum (y_i - \hat{y}_i)^2 = \sum \epsilon_i^2$$

Definition: The coefficient of determination r^2 is given by

$$r^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

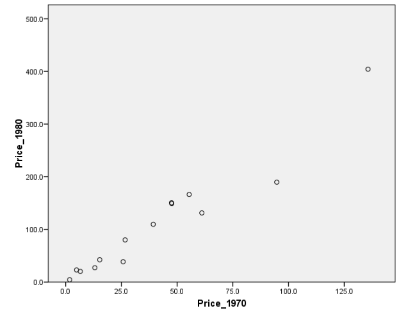
It can be interpreted as the percentage of variation observed in the response variable that can be explained by the linear relationship with x.

Note: In simple linear regression the coefficient of determination is the square of the correlation coefficient. Large values, close to 1, mean that the points all lie close to the regression line.

Example: The price for fresh fish has increased drastically over the years. The following table contains prices (in cents per pound) paid to fishermen for different types of fish and shellfish in 1970 and 1980.

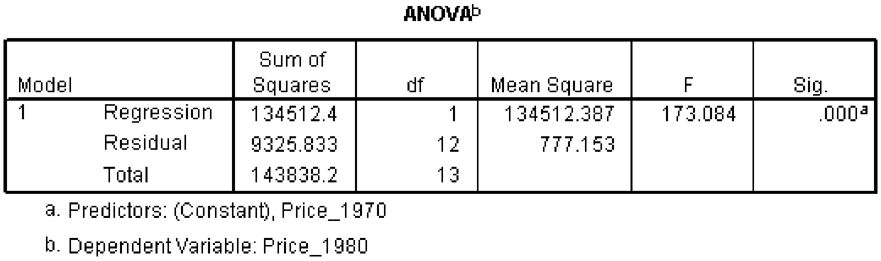
Fish	1970	1980
COD	13.1	27.3
FLOUNDER	15.3	42.4
HADDOCK	25.8	38.7
MENHADEN	1.8	4.5
OCEAN PERCH	4.9	23
SALMON, CHINOOK	55.4	166.3
SALMON, COHO	39.3	109.7
TUNA, ALBACORE	26.7	80.1
CLAMS, SOFT-SHELLED	47.5	150.7
CLAMS, BLUE HARD-SHELLED	6.6	20.3
LOBSTERS, AMERICAN	94.7	189.7
OYSTERS, EASTERN	61.1	131.3
SEA SCALLOPS	135.6	404.2
SHRIMP	47.6	149

(a) Take a look at the scatterplot of values. Do you think it is appropriate to fit a linear regression model in this case?



(b) Would you say that there is an outlier in the data set? Do outliers have a weak or strong influence on the fitted regression line?

(c) Take a look at the output of a regression analysis conducted by SPSS. The ANOVA table looks very similar to those produced in the ANOVA examples.



ANOVA[b]

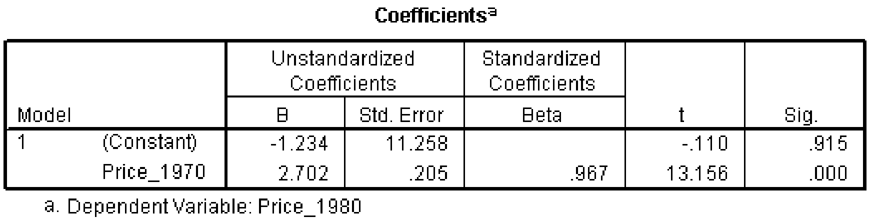
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	134512.4	1	134512.387	173.084	.000[a]
	Residual	9325.833	12	777.153		
	Total	143838.2	13			

a. Predictors: (Constant), Price_1970

b. Dependent Variable: Price_1980

(d) Find the value of $\hat{\sigma}^2$ in the ANOVA table.

(e) Look at the regression coefficients computed by SPSS:



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1.234	11.258		-.110	.915
	Price_1970	2.702	.205	.967	13.156	.000

a. Dependent Variable: Price_1980

(f) What percentage of the variation in prices in 1980 is explained through the regression on previous prices?