Chapter 1

Linear Regression

> People change. Problem is they don't always change together.
>
> unknown

We would like to have some way to mathematically *describe* the relationship between two quantitative (numerical) variables. Recall that quantitative are those variables which consist of numbers in contrast to the categorical ones which contribute categories (ex. yellow/blue/red). Associations between categorical variables we've analyzed in the previous chapter using chi-squared tests only on the basis exist/does not exist any association.

For quantitative variables we want to go further and find the mathematic formula that will exactly describe the relationship, answering not only on the existence of relationship but how exactly it will be.

So, we believe that some variable Y depends on some other variable X and thus can be described with it. We will then call X **independent** (explanatory) variable, and Y **dependent** (response) variable.

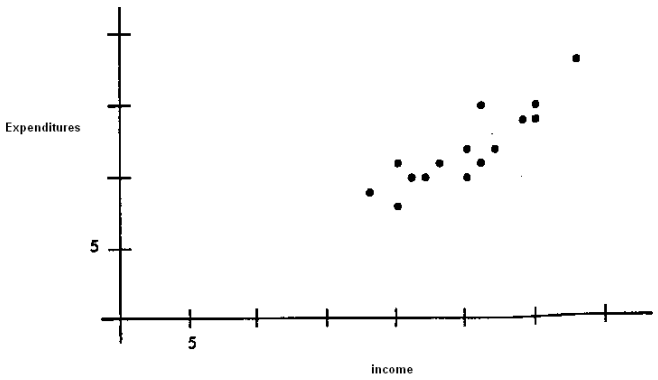
In our course we will consider only a case with two variables, one dependent and one independent, and will call it **simple regression**. It is also possible to construct a regression with many variables in which case it will be called **multiple regression**. You will study it in future next years.

Scatterplot

Imagine you are given a sample of observations.

Income (X)	65	35	37	110	45	72	30	60	55	50	35	70	20	45	100
Expenditures (Y)	58	15	21	70	25	30	29	25	45	20	20	40	18	20	50

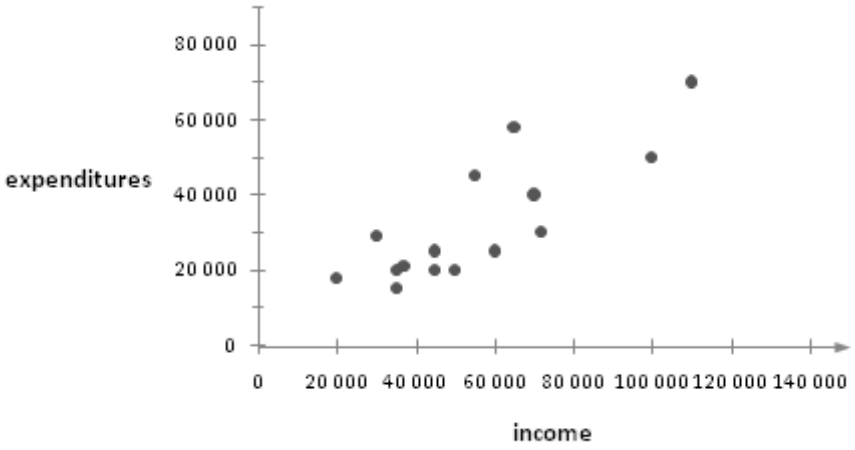
Each observation, plotted as a dot, has *two characteristics*, X and Y. It could be price and number of resulting units sold, or income and resulting expenditures, etc. Plotting these dots on a graph with $X - Y$ coordinates gives us a graph called **scatterplot** where every dot has two coordinates (x_i, y_i).



Stop shopping. Masha is a typical girl who loves... shopping! However, she studies Economics and doesn't want to spend her money irrationally. How much should she spend? She noticed that in general expenditures tend to rise with higher incomes. By how much exactly do they rise? Let us examine this relationship.

She obtains a sample of 15 randomly chosen people with the information on their incomes and corresponding expenditures available, per month in RUB. The research question is whether spendings agree with incomes or whether consumerism prevails over rationality and people spend disregard of earnings.

In thousands, $65 = 65,000$. This information is summarized in the scatterplot bellow.



Scatterplot provides us with a quick *visual impression* of a relationship, or its absence. It also allows to define whether it is linear or not.

Next step is how this relationship could be described in functional terms. What could be the expected expenditures for a given level of income? To answer that question we need to find the equation to describe how expenditures Y depend on income X.

Note that the assumptions necessary for this kind of analysis are:

1. There is a dependence of Y on X (and not vice versa)

2. This dependence is linear

We do that visually by looking at a scatterplot or by common sense, and all our analysis is made keeping in mind these assumptions.

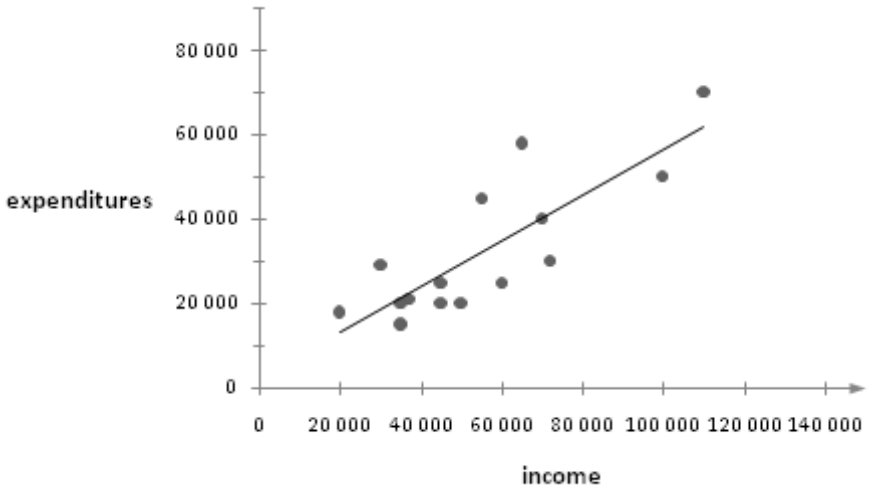
Regression Line (Ordinary Least Squares)

Once we have drawn a scatterplot and checked that Y seems to be linearly related to X, we can start the analysis. We need to find the equation of the form $\widehat{Y} = a + bX$ (a and b are some constants) which would provide predicted values of expenditures $\widehat{Y}$ as precise as possible to the observed values of expenditures (Y) for each level of income (X).

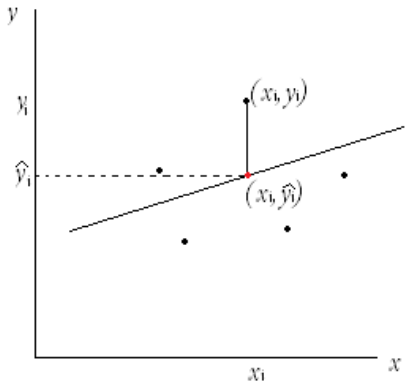
$$\widehat{Y} = a + bX$$

Y is called dependent or response variable. X is independent or explanatory variable, a and b are **regression coefficients**, $\widehat{Y}$ represents the estimated or **predicted value** of expenditures for given x.

Look at the scatterplot. To find the regression means to find a and b such that the resulting line $\widehat{Y} = a + bX$ is the best fit to the observations. It means that the line goes as closely to the scatterplot dots as possible. For that purpose we should minimize the *distances* from observed values Y to values of $\widehat{Y}$ predicted by the line.



Values of Y are simply found as y-coordinates of the dots and predicted values $\widehat{Y}$ are found as the height of the line corresponding to given level of X (remember, we are constructing a model to explain how Y depends on X). Thus, the coefficients a and b should be such that to minimize overall vertical distances from the observed dots to the resulting line.



This vertical distance for i-th observation is called the **residual** e_i and equals the difference between the actual value of expenditures y_i and the predicted value for i-th observation $\widehat{y}_i$: $e_i = y_i - \widehat{y}_i$. How to minimize the overall distances? Minimizing the sum of them $\sum_{i=1}^{n}(y_i - \widehat{y}_i)$ is useless, since it will always provide zero. It will happen because for observations above the line value of $(y_i - \widehat{y}_i)$ is positive, and for those below the line – it is negative, the sum will tend to reduce to zero. So, the conventional way to minimize the overall deviations (distances) is to minimize the sum of squared deviations: $\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$. First, each summand will be non-negative, thus, the sum will not reduce to zero. Second, minimizing squares of deviations allows to avoid especially high prediction errors (deviations) and to pay little attention to minor differences between y_i and $\widehat{y}_i$.

Now, how do we find coefficients a and b? We have a sample, so we are given n pairs of values $(x_i,\ y_i)$. Since $\widehat{y}_i = a + bx_i$ we minimize the function $f(a, b) =$

$\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$. Here x_i and y_i are known, a and b should be found so that to minimize $f(a, b)$.

To solve this minimization problem, according to first order condition we take partial derivatives with respect to a and b and equate them to zero:

$$\begin{cases} \frac{\partial f(a,\ b)}{\partial a} = 0, \\ \frac{\partial f(a,\ b)}{\partial\ b} = 0. \end{cases}$$

This way we get the system of two equations with two unknowns. This system results in the following formulas for regression coefficients:

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

You do not need to remember these formulas since they are provided in the standard list of formulas that you'll be given on the exam. If you are interested in where did these formulas come from, the derivation is provided in the "Awesome guys stuff" section at the end of this chapter, under a stamp of secrecy, of course.

Thus, using the formulas you can find the line $\widehat{Y} = a + bX$ for any given set of observations $(x_i,\ y_i)$. This line provides the best linear prediction of how Y depends on X. Since the formulas are derived from minimization of squared deviations, this strategy is called **Ordinary Least Squares** (OLS) and the resulting equation is called the **OLS regression**.

Values a and b can be automatically calculated by your graphical calculator.

There are two important properties of OLS coefficients:

1. Sum of residuals is zero: $\sum_{i=1}^{n} e_i = 0$

> One of the first order conditions in the minimization problem discussed is $\frac{\partial f(a,\ b)}{\partial a} = 0$, where $f(a),\ b = \sum_{i=1}^{n} (y_i - a - bx_i)^2$. Taking derivative we get: $-2 \cdot \sum_{i=1}^{n} (y_i - a - bx_i) = 0$. Thus, $\sum_{i=1}^{n} (y_i - a - bx_i) = \sum_{i=1}^{n} (y_i - \widehat{y}_i) = \sum_{i=1}^{n} e_i = 0$.

2. The regression line always goes through the point of mean values $(\bar{x}, \bar{y})$.

> This simply follows from the formula for b_0. Since $a = \bar{y} - b\bar{x}$, $\bar{y} = a + b\bar{x}$ and thus the point $(\bar{x}, \bar{y})$ belongs to the regression line.

Now, using the formulas, we can find the regression for the "STOP SHOPPING" example.

Masha has calculated that: $b = 0.54$, $a = 2534$

Thus, she gets the following regression equation: $\widehat{y} = 2534 + 0.54x$,

or in the context of the problem: $\widehat{\text{Expenditures}} = 2534 + 0.54\text{Income}$.

Note that this equation was calculated based on the assumption that Y depends on X. If the reverse is true and X is a dependent variable, the resulting equation to predict X on Y will be different!

Interpretation of regression coefficients

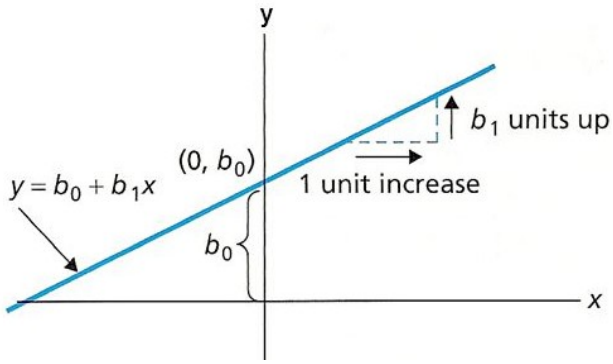
Regression equation can be used for prediction of value of Y given some value of X. $\widehat{Y}$ denotes the estimated or predicted or mean value of Y for given X.

For example, we can find the level of expenditures we should expect for a person with monthly income of 50,000 roubles as:

$\widehat{y}(50,000) = 2,534 + 0.54 \cdot 50,000 = 29,534$ (Roubles)

Thus, a person with income of 50,000, on average tends to spend about 29,534 rub per month.

The two coefficients of the regression equation have their names. a is called an **intercept** because the regression line intersects y-axis at a. b is a slope of the line, therefore it is called a **slope coefficient**.



Value of the slope coefficient is the most important result of the regression analysis. It shows the expected change in the dependent variable Y as a response to a unit change in explanatory variable X.

When X increases by 1, Y is expected to increase by b.

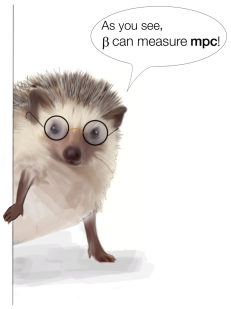
Proof

$\widehat{y}(x) = a + bx$

$\widehat{y}(x+1) = a + b(x+1)$

Change in y equals $\widehat{y}(x+1) - \widehat{y}(x) = a + b(x+1) - a - bx = a + bx + b - a - bx = b$

In our example slope $b_1 = 0.54$. It means that, every additional rouble of income increases monthly spendings by 0.54 roubles *on average*.



Recall from the Economic theory: b is the marginal propensity to consume (mpc), the proportion of disposable income which individuals spend on consumption. If a household earns an extra dollar of disposable income marginal propensity to consume of 0.54 means that the household will spend 54 cents and save 46 cents out of this additional dollar.

Intercept coefficient a

Intercept also called a **constant** shows expected value of Y given X equals zero.

a is the expected value of Y when X equals 0

Proof

The regression equation is $\widehat{Y} = a + bX$.

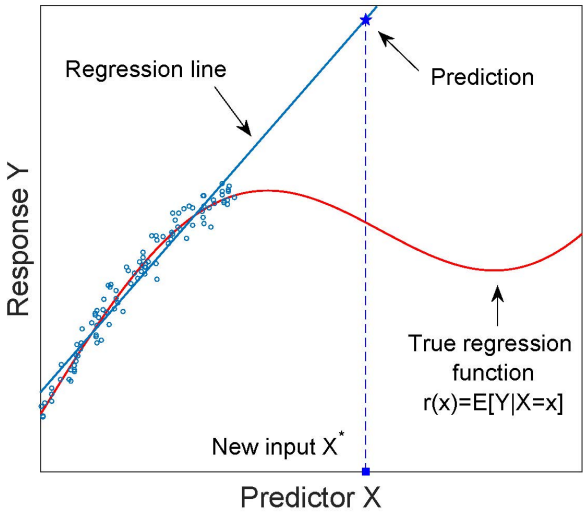
Predicted value of Y when X is zero is: $\widehat{y}(0) = a + b \cdot 0 = b$

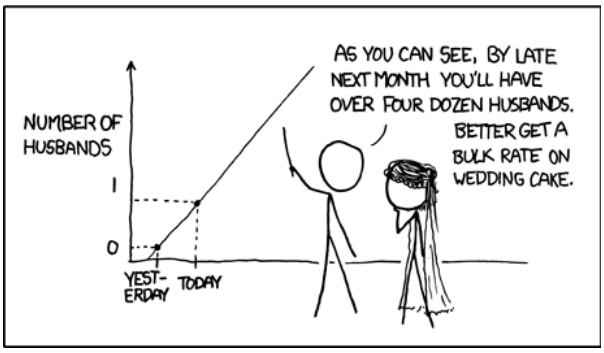
In our example constant equals 2534. It means that people which do not earn anything still spend 2534 roubles per month on average.

Note however, that sometimes intercept does not have meaningful interpretation. For example, in regression of weight of a person on her height intercept is the weight of a human with zero height, which is nonsense. This happens because of extrapolation.

Extrapolation

Extrapolation happens when you extend the relationship revealed in regression on values of variables that are out of the range of sample on which the regression was estimated. It is possible that variables do not exist there or that the relationship between Y and X has different form there:





Regression results can be reasonably used to predict values of Y only based on values of X which are within the range of the dataset on which the regression was estimated. All predictions based on the values below minimum or above maximum of X represent extrapolation and can be used only under assumption that X and Y exist in the corresponding region and that the relationship between them follows the same revealed pattern.

Residuals

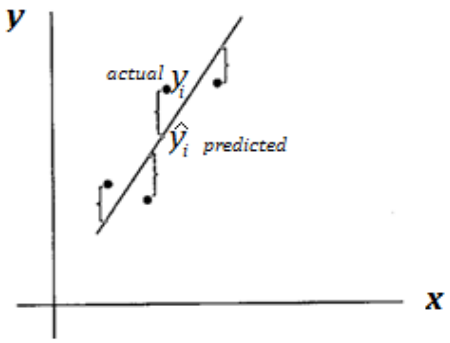
In the real world no relationship will hold *exactly*, so that to be described in one linear equation. We do not expect all points to lay exactly on a straight line or any other simple curve we can draw. Thus, the *discrepancy* from the regression line (deviation from the predicted value $\widehat{y}$) is always present. In other words, the true value of dependent variable y_i consists of the predicted (mean) value $\widehat{y}_i$ and the residual e_i:

$$y_i = \text{our expected prediction} + \text{discrepancy}$$

$$\widehat{y}_i e_i = a + bX$$

Residual e_i is the difference between the actual value of the dependent variable and the value predicted by the regression line for the corresponding value of independent variable.

$$e_i = y_i - \widehat{y}_i$$



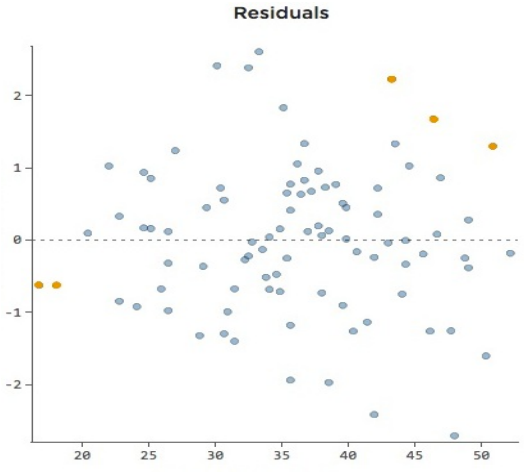
It is a vertical distance from the observed value and the regression line, as we've shown earlier. Residual shows how far the observed value lies from the fitted line.

Consider an observation (55,000, 45,000).

The predicted level of expenditures for the income of 55,000 is: $\widehat{y}(55,000) = 2534 + 0.54 \cdot 55,000 = 32,234$. The actual expenditures were 45,000.

Thus, the residual is $e_i = y_i - \widehat{y}_i = 45,000 - 32,234 = 12,766$. It shows that the actual expenditures for some household with the income of 55,000 roubles are by 12,766 roubles higher than it is predicted by the regression line.

Residual plot



Residual plot is a graph which shows values of residuals e_i (on y-axis) against independent variable X. It allows to visually evaluate the quality of the estimated regression.

If the regression fits observations well the residual plot should reveal that residuals are:

1. small (this indicates small prediction error)
2. distributed around 0 (regression contains no systematic error: $\widehat{Y}\,(X) = E(Y|X)$
3. random (reveal no trend, thus, e is not related to X).

We will talk more about these properties at the end of this chapter.

a and b are only estimates of the true regression coefficients α and β!

Let's assume that the three properties of a good regression are satisfied. Does that mean that our regression equation perfectly describes the way Y depends on X? Well, probably, if your calculations are based on the dataset including the whole population. However, in real life we do not have the complete information and can only use sample data. The true regression line is $\widehat{Y} = \alpha + \beta X$. The true equation also assumes difference between the actual y_i and estimated $\widehat{y_i}$ values of dependent variable. Thus, true dependent variable is $Y = \alpha + \beta X + \varepsilon$ where ε is the **error term** (true error of the regression).

However, without the data on the whole population we cannot *calculate* coefficients α and β. Based on sample data we can only *estimate* these parameters to provide the equation $\widehat{Y} = a + bX$, and then, the true dependent variable can be written as $Y = a + bX + e$, where e is the residual, the difference between actual y and its estimate based on sample data.

Thus, a and b are **estimated coefficients** (also called OLS estimators) for the true parameters α and β (constants): $\widehat{\alpha} = a$, $\widehat{\beta} = b$. Since coefficients a and b are based on random sample, which contains a random part of the population, a and b are

random variables. It can be shown that OLS coefficients are unbiased and efficient estimators for true coefficients α and β.

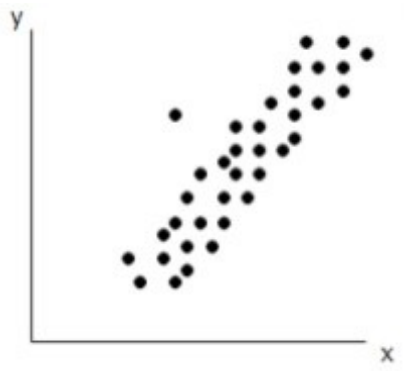
It's Ok, it only depends on whether your dataset is population or a sample: $\widehat{y}_{\text{perfect}} = \alpha + \beta X$ if you have population data and $\widehat{y}_{\text{available}} = a + bX$ based on sample data.

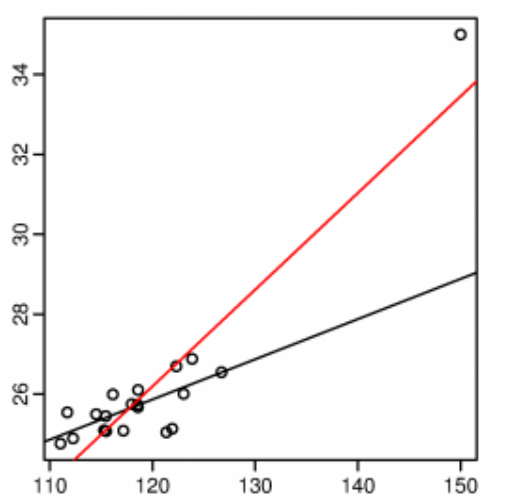
So the true value of y can be represented in both of the forms: $y_i = \widehat{y}_i^{\text{perfect}} + \varepsilon_i = \widehat{y}_i^{\text{available}} + e_i$

If you have the true coefficients α and β you can find error terms ε for all observations. If you don't – you can only find residuals e which are estimates for the true error terms.

Outliers and Influential Points

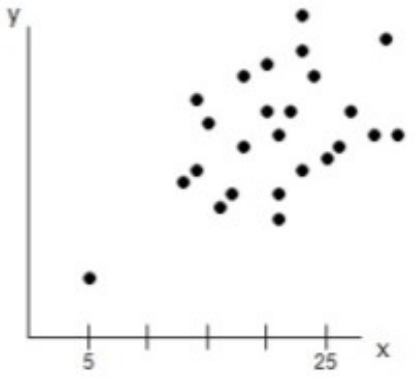
Outlier is an observation with large residual. Recall, that the residual is a vertical distance from the point to the regression line:





Influential point is an observation, which, being removed from the dataset would significantly change the regression line. Influential points are usually located at extremum values of X. On the picture below a point in the upper right corner is an influential point and an outlier at the same time.

However, it is possible that an influential point is not an outlier:



Also, outlier is not necessarily an influential point (the outstanding dot on the first picture is not an influential point).

Confidence intervals and tests for the slope coefficient

So, a and b, called OLS estimators, are random variables, taking random values determined by the sample dataset. Since they are unbiased for true regression coefficients: $E(a) = \alpha$ and $E(b) = \beta$. Standard deviations for a and b are generally denoted as SE_a and SE_b.

The true value of β is of critical interest in the regression analysis, since this parameter reflects the dependence of Y on X. When $\beta = 0$, Y does not depend on X. This is why the following two-sided test is most popular in regression analysis.

Test allows us to check whether the relationship exists in general. It is also called a test of significance, since it allows to check whether X has significant influence on Y.

H_0: $\beta = 0$ there is no dependence
H_A: $\beta \neq 0$ there exists some dependence

It can be shown that with large enough sample b is normally distributed. However, since true standard deviation of b is unknown, we use Student distribution: $\frac{b-\beta}{\text{SE}_b} \sim t(df)$, $df = n-2$ where n is the sample size. Here df equal $(n-2)$ because now we estimate two parameters instead of one: a and b. Since under null hypothesis $\beta = 0$ using the standard formula Test-statistic $= \frac{\text{Estimate}-\text{Parameter}_0}{\text{SE}_{\text{Estimator}}}$ we get the following testing statistic:

$$t_{\text{st}}\,(n-2) = \frac{b-0}{\text{SE}_b}$$

SE_b can be found by the formula $\frac{\sqrt{\frac{\sum_{i=1}^{n}(Y_i-\overline{Y_i})^2}{n-2}}}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}}$, which you should not remember in this course, it is provided in the standard set of formulas at the exam.

Let's test whether expenditures depend on income based on the "STOP SHOPPING" dataset:

H_0: $\beta = 0$

H_A: $\beta \neq 0$

$t_{\text{st}}(\text{df}) = \frac{b-0}{\text{SE}_b}\frac{0.54}{0.103} = 5.24$, $\text{df} = 15-2 = 13$

p-value $= 2P(t(13) > 5.24) = 2 \cdot 0.0000798 = 0.0001596 \approx 0$

P-value is smaller than any reasonable significance level α. So, we have enough evidence to reject H_0 and conclude that expenditures Y reveal relation with income X.

We can also construct confidence intervals for the slope coefficient β. Using the standard formula Parameter $=$ Estimator $\pm$ $\text{Statistic}_{\text{crit.}} \times \text{SE}_{\text{Estimator}}$ we get the following confidence interval:

$$\beta = b \pm t_{\alpha/2}(n-2) \cdot \text{SE}_b$$

Let's construct a 95% confidence interval for the slope b of regression line for our example:

$\beta = 0.54 \pm t_{0.025}(13) \cdot \text{SE}_b = 0.54 \pm 2.16 \cdot 0.103 = 0.54 \pm 0.2225$

Thus, $\beta \in [0.3174, 0.7626]$ at 95% confidence level.

We are 95% confident that the slope is between 0.317 and 0.763 rubles

How to find SE_b in your Casio calculator

Your calculator performs test and gives t-value. Having t and b you can find SE_b from the equation.

$t = \frac{b-\beta}{\text{SE}_b} = \frac{b-0}{\text{SE}_b}$, $t = \frac{b}{\text{SE}_b}$, $\text{SE}_b = \frac{b}{t} = \frac{0.54}{5.24} = \mathbf{0.103}$ Hint!

$t = 5.24$. Please, don't write these formulas in your solutions, it does not fit official solution, we only provide it for you as a hint for faster calculations.

Using Statistical Software

You should be familiar with the common output that is given by a statistical software after fitting regression equation. It always looks like this:

Term	Coef	SE Coef	T	P
Constant	2534.5	6220	0.40	0.695
Income	0.54	0.103	5.24	0.000

R-sq = 67.8

From the output regression equation we can reconstruct that: $\widehat{y} = 2534.5 + 0.54 \cdot$ Income

The first row corresponds to regression constant a and the second – to the slope coefficient for independent variable (income in our example).

The first column shows values of a and b.

The second column shows estimates of standard deviations of a and b: SE_a and SE_b

The last two columns represent results of the two-sided test of significance for a and b (under null hypothesis that true regression coefficients equal zero). The third column provides t-statistics and the fourth column shows the corresponding p-values for these tests.

Thus, we can see that for the slope coefficient:

$t_{\mathrm{st}}(13) = 5.24$, p-value ≈ 0. We conclude that Y reveals dependence on X.

for intercept:

$t_{\mathrm{st}}(13) = 0.4$, p-value ≈ 0.695. p-value exceeds any reasonable α, so we have no evidence to reject that $\alpha = 0$.

Correlation

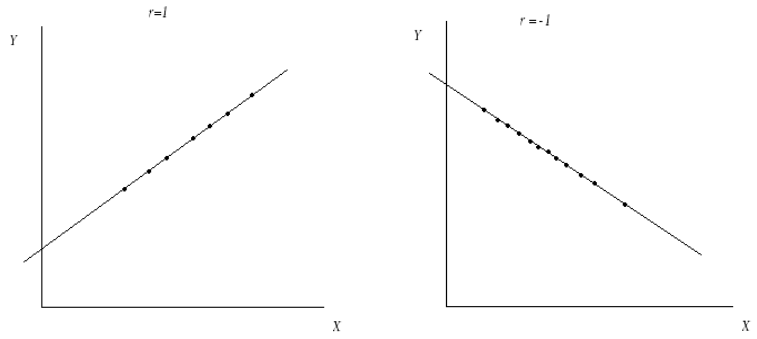
You are already familiar with the notion of correlation. Correlation coefficient shows the strength of *linear* relationship between two variables. It takes values between -1 and 1.

True (population) correlation coefficient is denoted by ρ_{XY}. However, we usually have sample data, and thus can only estimate correlation based on it. Estimated value of correlation is denoted by r.

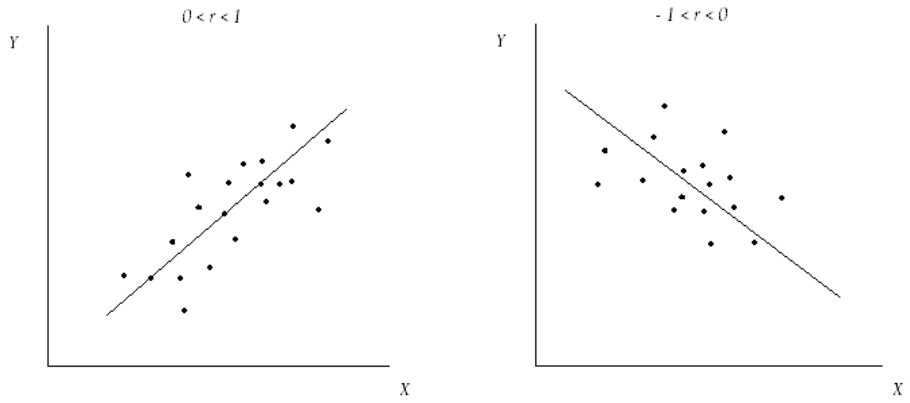
Recall that the sign of correlation indicates the direction of relation between variables. Positive correlation indicates that X and Y tend to move in the same direction, and negative correlation indicates that X and Y tend to move in the opposite directions. The closer is the absolute value of coefficient to 1, the closer is the relationship between X and Y to linear, so that there are constants a, b, such that: $Y = a + bX$. When $|r| = 1$ we have the case of perfect correlation. Zero correlation shows that two variables have no linear relations.

Perfect correlation means that $r = 1$ or $r = -1$ and the relation can be strictly described by a linear function:

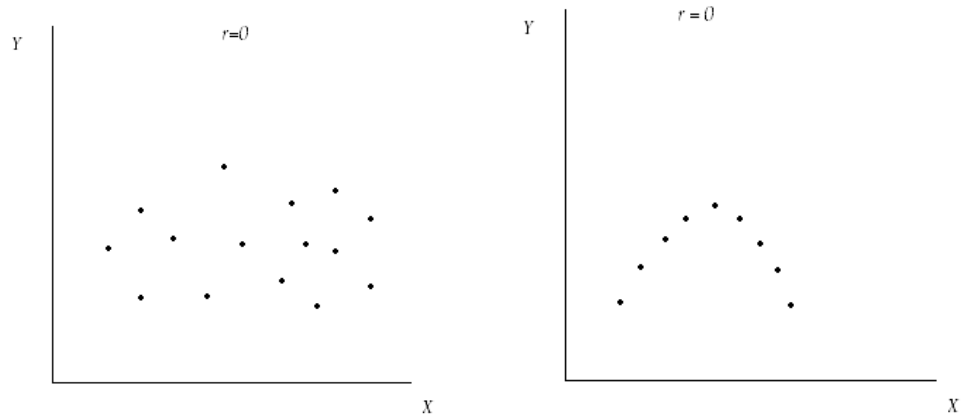
$$Y = a + bX, b > 0 \quad Y = a + bX, b < 0$$



However, we do not expect relationships studied in social sciences like Economics to be so strong, so that all observations (dots) simply lie on a straight line. In contrast, many relationships we are going to analyze will be rather weak, with correlation coefficient less than 1 in absolute value. Pictures below show scatterplots for the case of **imperfect correlation**:



Note that $r = 0$ does not necessarily mean the absence of relationship, it indicates the absence of linear relationship. So, $r = 0$ either in case when X and Y are not related *or* when the relationship between X and Y is not linear:



Previously, we have shown the following formulas for correlation:

$$\text{Corr}\,(X,Y) = \rho_{\text{XY}} = \frac{\text{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{E\left((X - E(X)) \cdot (Y - E(Y))\right)}{\sigma_X \cdot \sigma_Y}$$

.

As you can see, it requires to know true means and standard deviations of X and Y. Thus, it can only be calculated if you know the true joint probability distribution of random variables. When you only have sample data, some other calculation is required. How to estimate $\text{Cov}(X,Y) = E\left[(X - \mu_X) \cdot (Y - \mu_Y)\right]$ based on sample data? Expectation is the sum of values multiplied by their probabilities. If population includes N observations, we can say that each observation (X_i, Y_i) happens with probability of $\frac{1}{N}$. Thus, Covariance can be calculated as: $\frac{\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y)}{N}$. When true population means μ_X and μ_y are unknown, they can be replaced by sample means $\overline{X}$ and $\overline{Y}$, in which case n is replaced[1] by $(n-1)$:

$$\widehat{\text{Cov}}\,(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

Then, how to calculate sample correlation? Initially, $\text{Corr}\,(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y}$. Since true population standard deviations are unknown, we should replace it by sample standard deviations s_X and s_Y.

Thus, $\widehat{\rho}\,(X,Y) = r = \frac{\widehat{\text{Cov}}(X,Y)}{s_X \cdot s_Y} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1} \cdot \frac{1}{s_X \cdot s_Y}$. We can rewrite this formula as:

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

As you can see, correlation is the sum of the products of z-scores of X and Y. Recall, that z-score is a standardized version of a variable and the term "standardized" means, that they have 0 expectation and variance equal to 1. Therefore correlation can be viewed as the covariance between "standardized" versions of X and Y. This explains why changing units of measurement of X and Y does not change the value of correlation coefficient.

Relation with b

Note that formulas for correlation and for the OLS slope coefficient are very similar:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1} \cdot \frac{1}{s_X \cdot s_Y}$$

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} \cdot \frac{1}{{s_X}^2}$$

So, they can be expressed one from another:

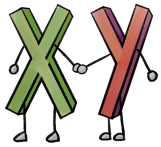
$$b = r \cdot \frac{s_Y}{s_X}$$

This close relationship between the two coefficients explains why Regression analysis is sometimes called a correlation analysis.

[1] Proof is the same as for sample standard deviation. Dividing by $(n-1)$ instead of n allows the estimate to be unbiased.

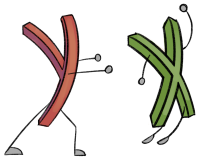
Correlation versus Causation

Recall that when you see that variable X and Y have significant correlation, all you can say is that they tend to move together, or simultaneously:

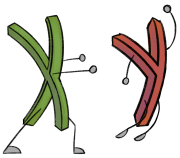


We are correlated! We go together!

The reasons behind this tendency might be very different. It possible that:

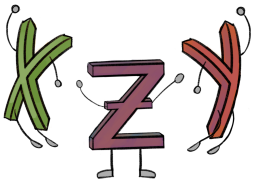


X depends on Y



Y depends on X

Or that both X and Y depend on some other variable Z, and this is what makes them related:



Do you remember "SMART PEOPLE WEAR BIGGER SHOES" example? Age was a lurking variable, which influenced both shoe size and IQ level, which therefore had significantly high correlation.

So, correlation does not imply causation. The only way to prove causation is by a controlled experiment.

It is you to decide which variable to take as dependent!

Analogously, significant slope coefficient in regression does not mean that Y depends on X. Contrary, it is before starting the analysis, that you state the assumption about the direction of possible linear dependence and then run regression. Usually this assumption is based on theoretical knowledge about the nature of variables to be analysed.

For funny examples of accidentally correlated variables go here: `http://tylervigen.com/spurious-correlations`

Coefficient of determination R^2

We construct regression in order to explain the relationship between Y and X and to create an instrument for prediction of Y based on value of X. How can we evaluate a quality of the resulting prediction tool?

R^2, or coefficient of determination, is the first quick tool to see how good our regression model is. It shows what part of behaviour of dependent variable Y we managed to explain by the regression equation. Particularly:

R^2 is proportion of variation in dependent variable Y, explained by variation in independent variable X:

$$R^2 = \frac{\sum_{i=1}^{n} \left(\widehat{y_i} - \bar{y}\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2}.$$

Denominator $\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2$ represents total sum of squares which is the measure of total variation observed in Y. Numerator $\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2$ represents the part of "behavior" of dependent variable which we can explain be the regression. R^2 takes values between 0 and 1. The closer is it to 1, the higher is the prediction power of the regression.

In our "STOP SHOPPING" example $R^2 = 0.72$, which means that 72% of the variation in expenditures is explained by the change in income.

It can be shown that R^2 is the square of coefficient of determination r:

$$R^2 = r^2$$

Model building

Note, that although we construct a linear model, OLS regression may be used to describe non-linear relations. For example, we may estimate regressions of the following forms:

$$\widehat{Y} = a + b \ln X \text{ or}$$

$$\ln \widehat{Y} = a + b \ln X \text{ or}$$

$$\widehat{Y} = a + bX^2, \text{ etc.}$$

It still the same OLS regression, but with transformed variables. For example X can be replaced by $\ln X$ or X^2 and then the same procedure is applied.

How to choose the right specification?

You should start your analysis with visual investigation of the scatter plot. If it reveals non-linearity, you should adopt your model correspondingly. Additional tool is analysis of residual plot after the regression was run.

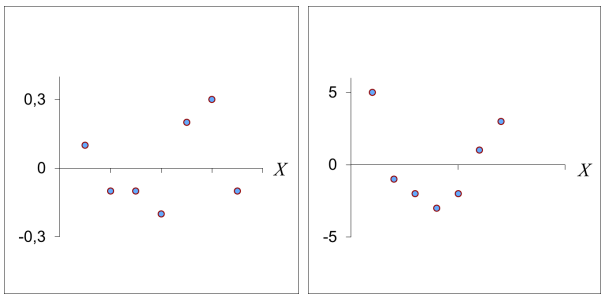
Remember the requirements for residuals of a good regression we stated earlier? The third one stated that the residual plot should show that residuals are random, that is, they are not related to X.

Intuition behind the third property is the following. Residual is a part of the value of dependent variable: $y(x) = \widehat{y}(x) + e_i$. It is a prediction error – a part of the behaviour of dependent variable, that we failed to explain by the regression on X: $e = y(x) - \widehat{y}(x)$. If this error e_i still reveals some dependence on X the regression function does not describe the dependence of y on x exhaustively: there still exists some influence of y on x, seen in the dependence of residual e on X (remember, e is a part of y). This means that the regression should be modified so that to incorporate this revealed aspect of dependence of Y on X, so that that the residual would becomes random and independent from X. This would mean that we've explained all possible effects on y caused by changes in x.

So if we find that residual plot follow certain pattern we need to change the formula of regression.

Sometimes the residual plot follows some definite nonlinear pattern. It is an indication that a linear model (straight line) is *not* a good fit. Probably some nonlinear model will be a better fit.

Consider an example. Drying times y (in minutes) are tabulated for various concentrations of pigment x (in ounces per gallon) for a certain brand of latex paint. A linear regression model gives with the following residual plot:



A non-linear model gives with the following residual plot:

1. What does each model predict for the drying time when using 5 ounces of pigment per gallon?

2. Which model is a better fit? Explain.

Solution.

a) Model predictions: $y = -5.1(5) + 59.7 = 34.2$ minutes, and $y = -16.8 \ln 5 + 59.8 = 32.8$ minutes.

b) Model 2 is the better fit. First, the residuals are much smaller for model 2, indicating that model 2 gives values much closer to the observed values. Second, a curved residual pattern as from model 1 indicates that a non-linear model would be better. A more uniform residual scatter as in model 2 indicates a better fit.

The goal is to achieve a model that is sufficiently simple for convenient interpretation but not so oversimplified that the important influences are ignored.

You must be able to reproduce even being half-awake

- $\widehat{y} = a + bx$ equation of regression line
- $\widehat{y}$ - mean predicted value of dependent or response variable, x – explanatory or independent variable
- interpretation of slope, constant
- residual and its interpretation
- residual plot
- R-sq
- extrapolation

Calculator BOX

To calculate linear regression plug observations in `List1(x)`, `List2(y)` and press: `CALC` → `REG` → Linear → EXE

To draw scatter plot: `GRPH` → Set Graph type: `Scat`, `XList: List1`, `YList: List2`, GRPH → S-Gph1 → EXE

From here you may calculate equation too! `CALC` → `Linear` → `EXE`

Top secret information

Derivation of OLS estimators

$$f(a, b = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \rightarrow min$$

Here x_i and y_i are known, a and b should be found so that to minimize $f(a, b)$.

To solve this minimization problem we take partial derivatives with respect to a and b and equate them to zero:

$$\begin{cases} \frac{\partial f(a,\ b)}{\partial a} = 0 \\ \frac{\partial f(a,\ b)}{\partial b} = 0 \end{cases}$$

1. $\frac{\partial f(a,\ b)}{\partial a} = -2 \cdot \sum_{i=1}^{n} (y_i - a - bx_i) = -2(\sum y_i - na - b \sum x_i) = 0$
   Divide equation by n: $\frac{\Sigma y_i}{n} - a - b\frac{\Sigma x_i}{n} = 0$
   $\overline{Y} - a - b\overline{X} = 0$
   $a = \overline{Y} - b\overline{X}$

2. Plug the result for b into $f(a,\ b)$:
   $f(a,\ b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2 = \sum \left(y_i - \overline{Y} + b\overline{X} - bx_i\right)^2 = \sum \left((y_i - \overline{Y}) - b(x_i - \overline{X})\right)^2$
   Now take derivative with respect to b_1:
   $\frac{\partial f(a,\ b)}{\partial b} = -2 \sum \left(x_i - \overline{X}\right) \left(\left(y_i - \overline{Y}\right) - b\left(x_i - \overline{X}\right)\right) =$
   $= -2 \left(\sum \left(x_i - \overline{X}\right) \left(y_i - \overline{Y}\right) - b \sum \left(x_i - \overline{X}\right)^2\right) = 0$
   $\left(\sum \left(x_i - \overline{X}\right) \left(y_i - \overline{Y}\right) - b \sum \left(x_i - \overline{X}\right)^2\right) = 0$

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

Sample AP practice problems

Problem 1. AP 2015 №5

A student measured the heights and the arm spans, rounded to the nearest inch, of each person in a random sample of 12 seniors at a high school. A scatterplot of arm span versus height for the 12 seniors is shown.

a) Based on the scatterplot, describe the relationship between arm span and height for sample of 12 seniors.

Let x represent height, in inches, and let y represent arm span, in inches. Two scatterplots of the same data are shown below. Graph 1 shows the data with the least squares regression line $\hat{y} = 11.74 + 0.8247x$, and graph 2 shows the data with the line $y = x$.

b) The criteria described in the table below can be used to classify people into one of three body shape categories: square, tall rectangle, or short rectangle.

Square	Tall Rectangle	Short Rectangle
Arm span is equal to height.	Arm span is less than height.	Arm span is greater than height.

i) For which graph, 1 or 2, is the line helpful in classifying a student's body shape as square, tall rectangle, or short rectangle? Explain.

ii) Complete the table of classifications for the 12 seniors.

Classification	Square	Tall Rectangle	Short Rectangle
Frequency			

c) Using the best model for prediction, calculate the predicted arm span for a senior with height 61 inches.

Solution:

a) There is a moderately strong, positive linear relationship (association) between height and arm span of a senior. Thus taller students tend to have longer arm spans.

b) i) The line depicted on Graph 2 is more helpful for the exact purpose described in the task. For every point (student) the graph illustrate whether the arm span is equal to the height ($y = x$, points on the line), arm span is greater than the height ($y > x$, points from above the line) or is less than the height ($y < x$, points from below the line).

ii) by counting the points the way exactly described in (i) we fill the table

Classification	Square	Tall Rectangle	Short Rectangle
Frequency	3	4	5

c) The predicted arm span for a senior with height 61 inches is: $\widehat{y} = 11.74 + 0.8247x = 11.74 + 0.8247 \cdot (61) =$ **62.05 inches** from the equation of regression line depicted on Graph 1. $\widehat{y}$ is the mean predicted height in inches.

Problem 2. AP 2014 №6

Jamal is researching the characteristics of a car that might be useful in predicting the fuel consumption rate (FCR); that is, the number of gallons of gasoline that the car requires to travel 100 miles under conditions of typical city driving. The length of a car is one explanatory variable that can be used to predict FCR. Graph I is a scatterplot showing the lengths of 66 cars plotted with the corresponding FCR. One point on the graph is labeled A.

Jamal examined the scatterplot and determined that a linear model would be a reasonable way to express the relationship between FCR and length. A computer output from a linear regression is shown below.

Linear Fit FCR $= -1.595789 + 0.0372614 \cdot$ Length

Summary of Fit RSquare	0.250401
Root Mean Square Error	0.902382
Observations	66

a) The point on the graph labeled A represents one car of length 175 inches and an FCR of 5.88. Calculate and interpret the residual for the car relative to the least squares regression line.

Jamal knows that it is possible to predict a response variable using more than one explanatory variable. He wants to see if he can improve the original model of predicting FCR from length by including a second explanatory variable in addition to length. He is considering including engine size, in liters, or wheel base (the length between axles), in inches. Graph II is a scatterplot showing the engine size of the 66 cars plotted with the corresponding residuals from the regression of FCR on length. Graph III is a scatterplot showing the wheel base of the 66 cars plotted with the corresponding residuals from the regression of FCR on length.

b) In graph II, the point labeled A corresponds to the same car whose point was labeled A in graph I. The measurements for the car represented by point A are given below.

FCR	Length (inches)	Engine Size (liters)	Wheel Base (inches)
5.88	175	3.6	93

i) Circle the point on graph III that corresponds to the car represented by point A on graphs I and II.

ii) There is a point on graph III labeled B. It is very close to the horizontal line at 0. What does that indicate about the FCR of the car represented by point B?

c) Write a few sentences to compare the association between the variables in graph II with the association between the variables in graph III.

d) Jamal wants to predict FCR using length and one of the other variables, engine size or wheel base. Based on your response to part (c), which variable, engine size or wheel base, should Jamal use in addition to length if he wants to improve the prediction? Explain why you chose that variable.

Solution:

a) For a car with length 175 inches, the predicted value for the car's FCR, based on the least squares regression line, is:

$$\text{predicted FCR} = -1.595789 + 0.0372614(175) \approx 4.92 \text{gallons per 100 miles.}$$

The actual FCR for the car was 5.88, so the residual is:

$$\text{Resid} = \text{Actual} - \text{Predicted} = 5.88 - 4.92 = 0.96$$

The residual value means that the car's FCR is 0.96 gallons per 100 miles *greater* than it would be predicted for a car of its length.

b) i) The point with a wheel base of 93 inches and a residual of 0.96 gallons per 100 miles is circled in graph III below.

ii) Point B corresponds to a car with an actual FCR that is very close to the predicted one, for a car with its length, by the regression model.

c) Graph II reveals a *moderate association* that is positive and linear. In contrast, there is a *weak association* that is positive and linear in graph III. The association between engine size and residual (from predicting FCR based on length) is stronger than the association between wheel base and residual (from predicting FCR based on length).

d) Engine size is a better choice than wheel base for including with length in a regression model for predicting FCR. The stronger association between engine size and residual (from predicting FCR based on length) indicates that engine size is more useful than wheel base for **reducing the variability in FCR values that remains unexplained** (as indicated by residuals) after predicting FCR based on length.

Problem 3. AP 2012 №1

The scatterplot below displays the price in dollars and quality rating for 14 different sewing machines.

a) Describe the nature of the association between price and quality rating for the sewing machines.

b) One of the 14 sewing machines substantially affects the appropriateness of using a linear regression model to predict quality rating based on price. Report the approximate price and quality rating of that machine and explain your choice.

c) Chris is interested in buying one of the 14 sewing machines. He will consider buying only those machines for which there is no other machine that has both higher quality and lower price. On the scatterplot reproduced above, circle all data points corresponding to machines that Chris will consider buying.

Solution:

a) The data show a weak but positive association between price and quality rating for these sewing machines. The form of the association does not appear to be linear. Among machines that cost less than \$500, there appears to be very little association between price and quality rating. But the machines that cost more than \$500 do generally have better quality ratings than those that cost less than \$500, which causes the overall association to be positive.

b) The sewing machine that most affects the appropriateness of using a linear regression model is the one that costs about \$2,200 and has a quality rating of about 65. Although the other four sewing machines costing more than \$500 generally have higher quality ratings than those costing under \$500, their prices and quality ratings follow a trend that suggests that quality ratings may not continue to increase with higher prices, but instead may approach a maximum possible quality rating. The \$2,200 sewing machine is the most expensive of all but has a relatively low quality rating, which is consistent with a nonlinear model that approaches a maximum possible quality rating and then perhaps decreases. If a linear model were fit to all of the data, this one machine would substantially pull the regression line toward it, resulting in a poor overall fit of the line to the data.

c) According to Chris's criterion, there are two sewing machine models that he will consider buying:

   1. The model that costs a bit more than \$100 and has a quality rating of 65.
   2. The model that costs a bit below \$500 and has a quality rating of 81 or 82.

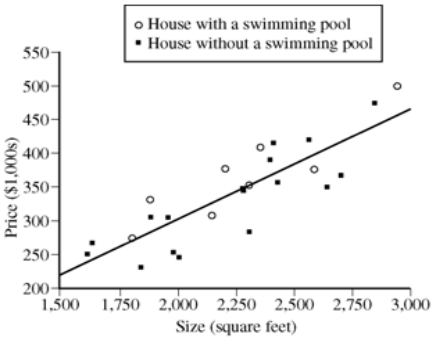
   Thus Chris would choose either the cheapest, or the one with the highest quality. The data points corresponding to these two machines have been circled on the scatterplot below.

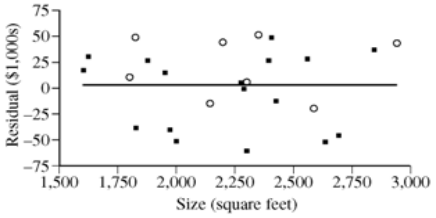
Problem 4. AP 2010 Form B №6

A real estate agent is interested in developing a model to estimate the prices of houses in a particular part of a large city. She takes a random sample of 25 recent sales and,

for each house, records the price (in thousands of dollars), the size of the house (in square feet), and whether or not the house has a swimming pool. This information, along with regression output for a linear model using size to predict price, is shown below.

Price ($1,000s)	Size (square feet)	Pool	Residual ($1,000s)
274	1,799	yes	6
330	1,875	yes	49
307	2,145	yes	–18
376	2,200	yes	42
352	2,300	yes	1
409	2,350	yes	50
375	2,589	yes	–23
498	2,943	yes	42
248	1,600	no	13
265	1,623	no	26
228	1,829	no	–45
303	1,875	no	22
303	1,950	no	10
251	1,975	no	–46
244	2,000	no	–57
347	2,274	no	1
345	2,279	no	–2
282	2,300	no	–69
389	2,392	no	23
413	2,410	no	44
353	2,428	no	–19
419	2,560	no	26
348	2,639	no	–58
365	2,701	no	–52
474	2,849	no	33





Linear Fit
Price = –28.144 + 0.165 Size

Summary of Fit
RSquare 0.722

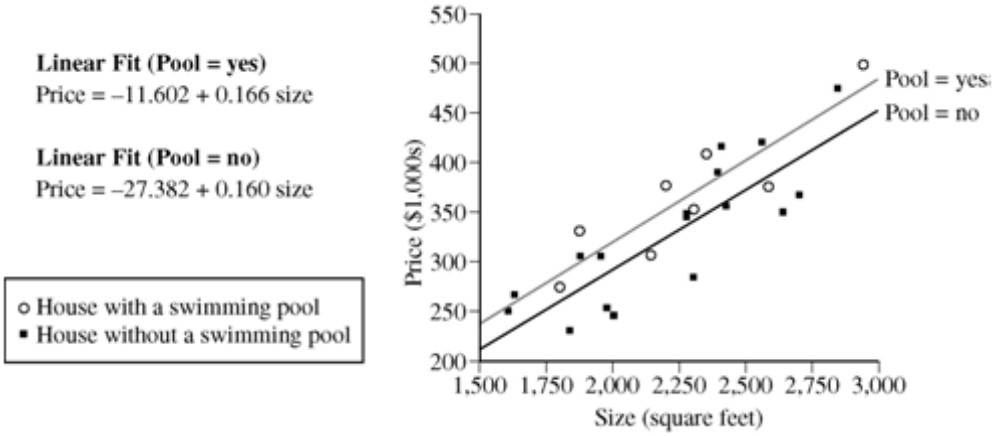
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	–28.144	48.259	–0.58	0.5654
Size	0.165	0.0213	7.72	<.0001

a) Interpret the slope of the least squares regression line in the context of the study.

b) The second house in the table has a residual of 49. Interpret this residual value in the context of the study. The real estate agent is interested in investigating the effect of having a swimming pool on the price of a house.

c) Use the residuals from all 25 houses to estimate how much greater the price for a house with a swimming pool would be, on average, than the price for a house of the same size without a swimming pool. To further investigate the effect of having a swimming pool on the price of a house, the real estate agent creates two regression models, one for houses with a swimming pool and one for houses without a swimming pool. Regression output for these two models is shown below.



d) The conditions for inference have been checked and verified, and a 95 percent confidence interval for the true difference in the two slopes is $(-0.099, 0.110)$. Based on this interval, is there a significant difference in the two slopes? Explain your answer.

e) Use the regression model for houses with a swimming pool and the regression model for houses without a swimming pool to estimate how much greater the price for a house with a swimming pool would be than the price for a house of the same size without a swimming pool. How does this estimate compare with your result from part (c)?

Solution:

a) The slope coefficient is 0.165. This means that for each additional square foot of size, the predicted price of the house increases by 0.165 thousand dollars, which is \$165. In other words, this model predicts that the average price of a house increases by \$165 for each additional square foot of a house's size.

b) The residual value of 49 for this house indicates that its actual price is 49 thousand dollars higher than the model would predict for a house of its size.

c) The average residual value for the eight houses with a swimming pool is: $\frac{6+49+(-18)+42+1+50+9+(-23)+42}{8} = \frac{149}{8} = 18.6$ thousand dollars. The average residual value for the 17 houses with no swimming pool is:

$$\frac{13 + 26 + (-45) + \ldots + (-58) + (-52) + 33}{17} = -\frac{150}{17} = -8.8 \text{thousand dollars.}$$

The residual averages suggest that the regression line tends to underestimate the price of homes with a swimming pool by about 18.6 thousand dollars and to overestimate the price of homes with no pool by about 8.8 thousand dollars. The difference between these two residual averages is $18.6 - (-8.8) = 27.4$ thousand dollars. This suggests that, for two houses of the same size, the

house with a swimming pool would be estimated to cost \$27,400 more than the house with no swimming pool.

d) No, this confidence interval does *not* indicate a significant difference (at the 95 percent confidence level, equivalent to the 5 percent significance level) between the two slope coefficients because the interval includes the value zero.

e) If the two population regression lines do in fact have the same slope, the impact of a swimming pool is the (constant) vertical distance between the two lines. However, because the two fitted lines do not have the same slope, the distance between the two fitted lines depends on the size of the house. Using the available information, there are two acceptable approaches to estimating the impact of having a swimming pool.

Approach 1: Because the slopes of the two sample regression lines were judged not to be significantly different, another acceptable approach would be to use the difference in the intercepts of the two fitted lines as an estimate of the vertical distance between the two population regression lines. The difference in the intercepts of the two fitted lines is $-11.602-(-27.382) = 15.780$ thousand dollars, which is an estimate of the impact of a swimming pool on the predicted price of a house, assuming this difference does not change with the size of the house. This is quite different from the estimate based on residuals in part (c).

Approach 2: Use the two fitted lines to predict the price of a house with and without a pool for a particular house size. For example, using the value of size = 2,250 square feet (which is near the middle of the distribution of house sizes), we find:

Predicted price for a 2,250 square-foot house with a swimming pool $= -11.602+ 0.166 \cdot 2,250 = 361.898$ thousand dollars.

Predicted price for a 2,250 square-foot house with no swimming pool $= -27.382+ 0.160 \cdot 2,250 = 332.618$ thousand dollars.

The difference in these predicted prices is $361.898-332.618 = 29.280$ thousand dollars, which is an estimate of the impact of a swimming pool on the predicted price of a 2,250 square-foot house. This is quite similar to the estimate based on residuals in part (c).

You can use either depending on what you like more or understand better.

Problem 5. AP 2008 №6 b,c,d

Administrators in a large school district wanted to determine whether students who attended a new magnet school for one year achieved greater improvement in science test performance than students who did not attend the magnet school. Knowing that more parents would want to enroll their children in the magnet school than there was space available for those children, the district administrators decided to conduct a lottery of all families who expressed interest in participating. In their data analysis, the administrators would then compare the change in test scores of those children who were selected to attend the magnet school with the change in test scores of those

who applied to attend the magnet school but who were not selected.

a) *Topic: Hypothesis testing-Mean Difference*

Administrators were also interested in using pretest scores on this test as a predictor of posttest scores on the test. The following computer output contains the results from separate regression analyses on the magnet school scores and on the original school scores. The accompanying graph displays the data and separate regression lines for the magnet and original schools.

Regression Analysis: Post_Magnet versus Pre_Magnet

Predictor	Coef	SE Coef	T	P
Constant	73.27	34.55	2.12	0.078
Pre_Magnet	0.1811	0.4583	0.40	0.706

S = 8.20920 R-Sq = 2.5% R-Sq(adj) = 0.0%

Regression Analysis: Post_Original versus Pre_Original

Predictor	Coef	SE Coef	T	P
Constant	9.24	11.91	0.78	0.456
Pre_Original	0.9204	0.1512	6.09	0.000

S = 4.11463 R-Sq = 78.8% R-Sq(adj) = 76.6%

b) i) State the equation of the regression line for the magnet school and interpret its slope in the context of the question.

ii) State the equation of the regression line for the original school and interpret its slope in the context of the question.

c) To determine whether there is a significant correlation between pretest score and posttest score, a test of the following hypotheses will be performed.

H_0 : There is no correlation between pretest score and posttest score (true slope = 0)

H_a : There is a correlation between pretest score and posttest score (true slope $\neq$ 0)

i) Using the regression output, state the p-value and conclusion for this test at the magnet school. Assume the conditions for inference have been met.

ii) Using the regression output, state the p-value and conclusion for this test at the original school. Assume the conditions for inference have been met.

d) What additional information do the regression analyses give you about student performance on the science test at the two schools beyond the comparison of mean differences in part (a)?

Solution:

b) Let y = posttest score and x = pretest score.

i) The predicted regression equation for the magnet school is $\hat{y} = 73.27 + 0.1811x$. For students at the magnet school, a 1-point increase in the pretest score is associated with a predicted increase of 0.1181 points on the posttest (i.e., the slope is positive but close to zero).

ii) The predicted regression equation for the original school is $\hat{y} = 9.24 + 0.9204x$. For students at the original school, a 1-point increase in the pretest score is associated with a predicted increase of 0.9204 points on the posttest (i.e., the slope is positive and close to 1).

c) i) The test statistic is $t = 0.40$ with a **p-value of 0.706**. Because the p-value is greater than any reasonable significance level, say 0.05, we fail to reject H_0. We conclude that there is insufficient evidence to state that pretest score is a significant predictor of posttest score at the magnet school. The data do not support a conclusion that a correlation exists between pretest and posttest scores at the magnet school.

ii) The test statistic is $t = 6.09$ with a **p-value of 0.000**. Because the p-value is less than any reasonable significance level, say 0.05, we reject H_0 and conclude that there is sufficient evidence to state that pretest score is a significant predictor of posttest score at the original school. The data support a conclusion that a correlation exists between pretest and posttest scores at the original school.

d) Unlike the two-sample analysis of differences in part (a), the regression analyses allow us to explore the relationship between pretest and posttest scores at each school. From the regression output and graph, we see that students with low pretest scores benefit more from attending magnet schools, as compared with students with low pretest scores at the original school. Also at the magnet school, students with low pretest scores benefit more than students with high pretest scores. In other words, students at the magnet school all score high on the posttest, regardless of how they scored on the pretest. But at the original school, only students who scored high on the pretest scored high on the posttest.

Problem 6. AP 2008 Form B №6

The nerves that supply sensation to the front portion of a person's foot run between the long bones of the foot.

Tight-fitting shoes can squeeze these nerves between the bones, causing pain when the nerves swell. This condition is called Morton's neuroma. Because most people have a dominant foot, muscular development is not the same in both feet. People who have Morton's neuroma may have the condition in only one foot or they may have it in both feet.

Investigators selected a random sample of 12 adult female patients with Morton's neuroma to study this disease further. The data below are measurements of nerve swelling as recorded by a physician. A value of 1.0 is considered "normal," and 2.0 is considered extreme swelling. The population distribution of the swelling measurements is approximately normal for adult females who have Morton's neuroma.

Dominant Foot	Swelling in Dominant Foot	Swelling in Nondominant Foot	Foot with Neur
Left	1.40	1.10	Left
Left	1.55	1.25	Left
Left	1.65	1.20	Left
Left	1.55	1.40	Both
Left	1.70	1.40	Left
Left	1.85	1.50	Both
Right	1.45	1.20	Right
Right	1.65	1.30	Right
Right	1.60	1.40	Right
Right	1.70	1.45	Both
Right	1.85	1.45	Both
Right	1.75	1.60	Both

a) A scatterplot of the ordered pairs (swelling in left foot, swelling in right foot), is shown below. The scatterplot suggests there are two distinct groups of patients. Patients within each group share a common trait. Use the scatterplot above and the table on page 10 to determine the common trait and explain how this trait differs for the two groups.

b) A scatterplot of the ordered pairs (swelling in dominant foot, swelling in nondominant foot), is shown below. What conclusion can be drawn from this scatterplot that is not apparent from the scatterplot in part (a) ?

c) See Chapter 11

Solution:

a) The trait that distinguishes the two groups in the scatterplot is dominant foot (left or right). All of the points in the upper left cluster represent measurements from individuals whose dominant foot is the right foot, whereas all of the points in the lower right cluster represent measurements from individuals whose dominant foot is the left foot.

b) There is a positive linear relationship between swelling in the dominant foot and swelling in the nondominant foot. Swelling in the dominant foot tends to be greater than swelling in the nondominant foot.

Practice AP problems

Problem 1. AP 2016 №6

Problem 2. AP 2011 №5

Problem 3. AP 2010 №1 b

After performing the experiment, the researchers recorded the data shown in the table below.

Garlic oil concentration	0%	2%	10%	25%	50%
Mean number of food granules consumed	58	48	29	24	20
Number of birds	8	8	8	8	8

i) Construct a graph of the data that could be used to investigate the appropriateness of a linear regression model for analyzing the results of the experiment.

ii) Based on your graph, do you think a linear regression model is appropriate? Explain.

Problem 4. AP 2008 №4 ab
Problem 5. AP 2007 Form B №4
Problem 6. AP 2007 Form B №6 bcd
Problem 7. AP 2006 №2
Problem 8. 2017 №1
Problem 9. 2007 №6
Problem 10 2005 B №5
Problem 11 2005 №3
Problem 12 2004 B №1
Problem 13 2003 B №1
Problem 14 2002 B №1
Problem 15 2002 №4
Problem 16 2001 №6 (bc)
Problem 17 2001 №1
Problem 18 1999 №6 (c)
Problem 19 1999 №1
Problem 20 1998 №4

Problem 21. AP 2018 №1

The manager of a grocery store selected a random sample of 11 customers to investigate the relationship between the number of customers in a checkout line and the time to finish checkout. As soon as the selected customer entered the end of a checkout line, data were collected on the number of customers in line who were in front of the selected customer and the time, in seconds, until the selected customer was finished with the checkout. The data are shown in the following scatterplot along with the corresponding least-squares regression line and computer output.

(a)Identify and interpret in context the estimate of the intercept for the least-squares regression line.

(b)Identify and interpret in context the coefficient of determination, r^2.

(c)One of the data points was determined to be an outlier. Circle the point on the scatterplot and explain why the point is considered an outlier.

Answers to practice problems

Problem 21. (a)The estimate of the intercept is 72.95. It is estimated that the average time to finish checkout if there are no other customers in line is 72.95 seconds.

(b)The coefficient of determination is $r^2 = 73.33$ %. This value indicates that 73.33 % of the variability in the times it takes customers to finish checkout, including time waiting in line, can be explained by knowing how many customers are in line in front of the selected customer.

(c)The outlier is the point with x=3 and y close to 0. This point is considered an outlier because the combination of x and y values differs from the pattern of the rest of the data. Specifically, the value of y (time to finish checkout) is much lower than would be expected when there are x=3 customers in line in front of the selected customer, given the remaining data.