Module 6: Regression 1 – The Basic Model

Learning Goal for Module 6

This Module (and the next) presents regression, a core tool in data analytics and economic analysis. It explains relationships among economic variables and serves as a foundation for prediction and forecasting.

By the end of this Module, you will:

- Understand measures of association and correlation
- Understand the difference between dependent (effect/outcome) variables and independent (cause/input) variables
- Obtain insight into data and algorithmic approaches to analytics
- Understand the form of the linear regression
- Specify and run a linear regression using the Data Analysis ToolPak
- Interpret regression coefficients

Modules 6 and 7 present the regression tools in Excel. Module 12 shows how to use matrix capabilities in Excel, especially dynamic arrays that have just become available, to build regression models "from scratch."

1. Introduction

Dependent variables are effects and independent variables are causes.

Regression, the workhorse in economic and business analytics, has two primary roles. First, regression models support the testing of economic theories and help make a case for causal relationships. They attempt to explain the variation in an outcome (dependent) variable, using the changes in a range of input (independent) variables

- **Example:** What is the relationship between changes in housing sales and interest rates?
- **Example:** What happens to university enrolments with the elimination of tuition fees?

Second, forecasting models often start with regression equations. These models attempt to predict the changes in the outcomes based on changes in the independent variables.

- **Example:** If central banks increase interest by 1% (100 basis points), what will happen to home sales?
- **Example**: If the university increases fees for international students by 25%, by how much will enrolments from China change?

One way to start building a regression is to explore typical associations that can appear to reflect a plausible explanation of reality. Often, upon further reflection and analysis, more complex relationships among social and economic variables emerge.

- Example: How does the sex of the worker affect variations in income?
- Example: What role do occupation and seniority play in determining incomes?
- Example: Does having children impose a penalty on lifetime incomes?

A block diagram can clarify meaning. Here a change in price and income will change the outcome; that is, the quantity demanded. If price is the "own price," namely the price of the product/service, then an increase in price typically induces consumers to purchase less and quantity demanded falls.

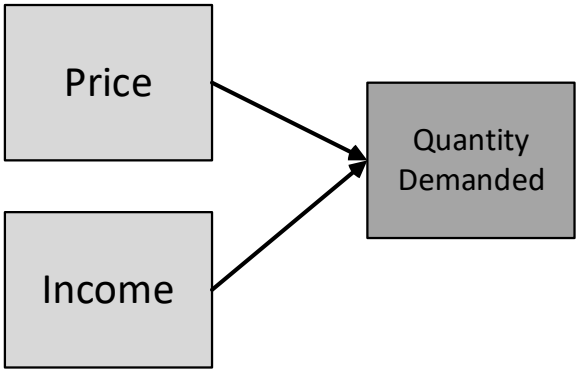


Figure 1: Basic regression model

When price increases, the quantity demanded typically falls. Exceptions exist, such as luxury goods with snob appeal. These general causal relationships among the factors of demand have wide acceptance in economics. Therefore, in addition to associations, theory plays a key role in developing models.

However, reality is nothing if not complex. In the background, other factors can affect what we measure and may produce counterintuitive associations. Termed "lurking variables," the economic analyst must remain alert t o all social, economic, and environmental influences on data patterns. In the example of price and income affecting quantity demand, the price of a competitive product may fall, which could reduce the effect of a decline in its own price.

A block diagram of the regression model, such as Figure 1, imposes a theory of causal relationship. Creating a regression model implicitly creates an understanding of the relationship among the variables. Price and income cause changes in quantity demanded, not the other way around. Yet it is easy to imagine instances when restrictions on the availability of a good or service affects its price. Stores often will have "sales" of excess inventory, where the surplus product has triggered a price reduction. This is what makes economic analytics endlessly fascinating (or frustrating)!

1.1. Two approaches to statistical modelling

- Data modelling: Assumes that the outcomes are the result of a specific equation and tests whether the data conforms to that model. In refined models, theory will produce a

specific equation, and depending on statistical testing, the analyst will (or will not) accept the theory. The linear model (shown below) remains the starting point for many studies where outcomes = f (independent variables and errors in a defined algebraic model).

$$Y= a_0 + a_1X_1 + a_2X_2 .... + e,$$

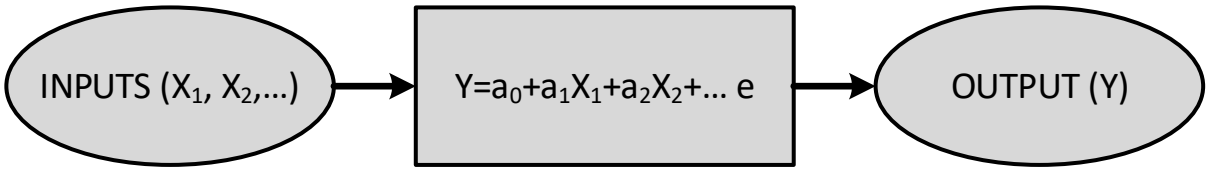


Figure 2: Data model

Using the example price (P), income (Y), and quantity demanded (Q_d), the model appears as

$$Q_d = f(P,Y) = a_0 + a_1P + a_2Y + e$$

- Algorithmic modelling: Tries to find the set of rules that does the best job of explaining the variation in outcomes. Also known as "machine learning," this approach process is atheoretical in the sense that the primary goal is to find the simplest set of rules that link inputs to outputs. The rule generation process may be a "black box" and change with the data.

Outcomes = f(rules that predict Y)

Figure 3: Algorithmic model

The result of algorithmic modelling will usually not be the linear model. Some see this as an important advantage of algorithmic modelling. Others see the lack of theoretical foundations as a fatal flaw since it does not lead to clear policy interventions.

1.2. Confounding factors: Simpson's Paradox

The National Science Foundation in the US conducted a study of persons who received a degree in social science, life/physical science, or engineering in 1977 or 1978. The study found that, at the bachelor's degree level, women with a full-time job had salaries that were 77% of men's

salaries. However, comparing salaries within each field, the average salary for women was at least 92% of the average male salary.

But why???

The explanation here is a **lurking or confounding variable**:

- Women concentrated in the social and life sciences professions, which have lower salaries than engineering, a profession that had a much higher number of men than women.
- Considering all fields, the large numbers of women in social sciences combined to create an aggregate measure that showed much lower salaries for women than men when model does not include field (occupation). Occupation was a lurking variable in the simpler model with just SEX as the explanatory variable.

See Simpson's Paradox.xlsx

Video: Simpson's Paradox

1.3. Correlation

The terms *correlation* and *association* are closely related concepts.

- Correlation *suggests* the *potential* for a causal relation.
- Correlation coefficients vary between -1 (perfect negative) and +1 (perfect positive). A value of "0" implies no correlation, and values of -1 and +1 suggest a technical relationship, such as the correlation between temperatures in Fahrenheit or Celsius.
- **Correlation never implies causation;** a low correlation may disguise a complex causal relationship.

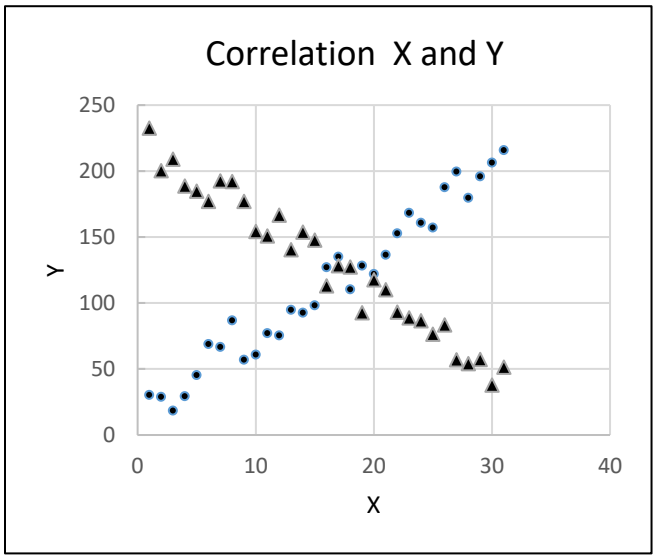


Figure 4: Correlation of X vs Y and X vs Z

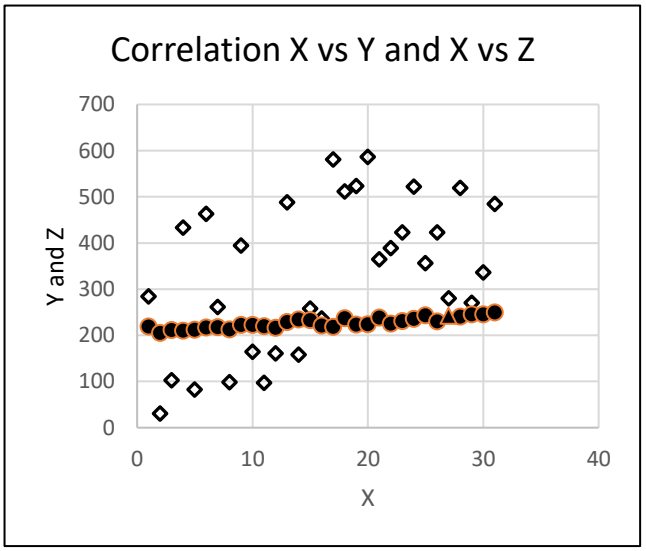


Figure 5: Correlation of X vs Y

Two common measures of correlation are:

1. The **Pearson product-moment correlation coefficient** assumes a *linear* relationship between x and y. The measure ρ_{xy} is termed the "correlation coefficient" and assumes X and Y have a linear or straight-line relationship. It is the most common measure of correlation in economic analytics. See **Pearson.xlsx**

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

The Excel Formula is =CORR(Array1,Array2)

Video: Pearson Correlation

Pearson's correlation is most common in economics and business.

2. **Spearman's rank correlation coefficient** assesses the extent to which a pair of numbers increase/decrease when placed in rank order. Numbers that move in opposite directions will have negative value; numbers that move together have positive correlations. This measure does not assume whether X and Y are linearly related. This measure applies when the two variables are order, interval, or ratio, not cardinal.

$$r = 1 - \frac{6\sum_{i=1}^{n} di^2}{n(n^2 - 1)}$$, where di is the difference in the values of the ranked pairs.

No formula for the Spearman rank correlation exists in Excel. It requires manual development for each dataset.

See **Spearman.xlsx**

Video: Spearman Correlation

Interpreting a correlation coefficient (ρ)

The *sign* (±) describes the direction, and the *absolute value* of a correlation coefficient describes the magnitude of the relationship between two variables. The greater the absolute value of a Pearson product-moment correlation coefficient, the stronger the *linear* relationship. Recall that ρ varies between -1 and +1.

- A correlation coefficient equal to 0 means no relationship between x and y. (But be careful, this does not "prove" the absence of causation; confounding variables may interact in complex ways.)

Spurious correlation occurs when two variables have a high correlation but are not causally related. Example: While most cancers tend to emerge later in life, and most cancer patients have children, having children is not a cause of cancer. The correlation between the incidence of cancer and incidence of being a parent is spurious.

- A positive correlation means that, as one variable becomes larger, the other variable also becomes larger. Conversely, a negative correlation means that, as one variable gets larger, the other variable becomes smaller.

1.4. Cross-tabulation uses "count data" that reveal a relationship between two variables

The typical bivariate display may use "discrete" or "continuous" data.

- Discrete data refer to information categories that take on a limited set of values. Examples include:
  - Sex (male/female)
  - Income level (low, medium, high)
  - Number of votes received for political parties from each sex

- Cross-tabulation (crosstabs) refers to the creation of tables to show the bivariate relationship between the two variables.

Table 1: Cross-tabulation (crosstab)

Votes→ Sex↓	Green	Liberal	Conservative	*Total*
Male	32	58	91	***181***
Female	56	62	33	***151***
Total	***88***	***120***	***124***	***332***

Imagine we have voting and sex data on 332 people. Sex has two categories, while three political parties exist. A crosstab might appear as in Table 1, where it appears that men vote Conservative, while women vote for the Green party. Support for Liberals is even between the two sexes. Note how the rows and columns sum and how the final column (row totals) and final row (column totals) sum to the total votes recorded.

1.5. Bivariate versus multivariate data: The first step to causal analysis

Bivariate analysis (primary level) means analysing pairs of variables, which is often the first step in thinking about causal relations. However, this form of analysis may obscure the lurking variables that show other factors influencing the data.

Steps to secondary level analysis include:

- conducting correlations of outcome variables with key attributes, behavioural, and attitudinal variables; and

Social sciences classify the world into data that further divides into causal and effect (outcome) factors.
The terms "factor," "influence," and "determinant" are commonly used. Labelling of one factor as a "cause" (education) and another as an "outcome" (income in later life) rests on a combination of theory and observation. Most outcomes have several factors/determinants/causes.

- using multivariate analysis, such as regression, to show the simultaneous influence of several variables on an outcome.

The object of this second-level analysis is to explore the deeper structure in the data (which sets the stage for thinking about cause and effect). In the second stage of analysis, we become alert to the possibility that more than one variable may influence the outcomes. Tertiary level analysis involves more complex models with feedback effects and multiple outcomes, where the outcomes in one period become the inputs to the analysis in later periods.

Theory guides the search for a deeper structure (exploration of cause and effect) in the data. In social sciences, "causal" (or input) variables (also termed ***independent variables***) act on "effect" or "outcome" variables (also termed ***dependent variables***).

Common outcome (dependent) variables in economics and business include:

- employment
- increases in income
- sales (revenue or quantity)
- profits
- …

Caution: We never prove causality; rather we establish the "strength of plausibility" for thinking a causal relation exists between a set of independent and dependent variable(s).

Common causal (independent) variables include:

- gender
- education
- income
- costs
- prices
- ...

Certain independent variables support purposeful manipulation. Such *policy variables* are particularly important. Tax and interest rates are common examples of policy variables.

Hypotheses are statements of cause and effect that rest on theory but remain uncertain. A common method of establishing the *plausibility* of such hypotheses in economics and business is multivariate analysis. Multivariate analysis assumes that multiple influences or independent variables (causes) exist for changes in the dependent variable (effect). The multiple regression model introduced in Module 7 is a usual form of multivariate analysis.

1.6. Sex and gender: Navigating shifting definitions

A decade or two ago, the words "sex" and "gender" when used to classify humans were used invariably as a binary variable – male or female. Now, the terms "sex assigned at birth" and 2SLGBTQQIA+ (Two-spirit, lesbian, gay, bisexual, trans, queer, questioning, intersex, and asexual) are common. The plus sign acknowledges the many sexual and gender minority people

who don't see themselves in the umbrella acronym and prefer other identity terms, such as pansexual, gender-free, or intersex.)[1]

This is not the place to engage in cultural debate. Practically, how should an economic analyst handle sex and gender variables? Quite simply, return to the original definitions used to collect the data. For example, if one is using data from the 1970s, then unambiguously sex/gender were binary. This is apparent from the data collection forms, such as the census or employment records. In fact, it is reasonable to use this binary view for data on sex and gender collected before 2015.

However, public data collection and private records are evolving to include more classifications of sex and gender. Always check the definitions used to collect information from the original survey, data entry forms, and web-based forms. This rule applies to all variables, not just the sex/gender variables.

The reason economists include a variable representing binary classifications of sex is to create a shorthand reflecting the different life experiences of someone who is male or female. Historically, male, and female life experience reflected factors such as occupational selection and childcare obligations. These differences affected a host of behaviours that affected life outcomes, such as income and other measures of well-being.

Challenges will exist as sex/gender move from a binary variable to a multidimensional concept. Further, with increased child-care subsidies, attention to pay equity, and paternal leaves that encourage men to support childcare, the income/occupational penalty for having children that used to fall entirely on women is lessening.

An important lurking variable exists in sex/gender data collected in 2022. Younger people are more likely to report non-binary sex/gender identities than older persons. This may reflect that . younger people are more likely to accept current trends and report non-binary sex/gender identities.

All this underscores the complexity of what was once a simple binary measure – sex – with just male and female as options.

2. Understanding simple linear regression (one independent variable)

The regression model is one of the main tools used in economics to explain and predict. This section explores the simple (one independent variable) regression model as the first step in the development of your understanding of this form of analysis.

2.1. Regression: The basic idea

Here is a thought experiment. Imagine that you are a new clerk in a convenience store. It is a boring job, especially at night, and to pass the time you note and write down the height of

[1] https://employees.viu.ca/human-resources/equity-diversity-inclusion/learning/2slgbtqqia-terms

people coming through the front door. (Many convenience stores have a measuring tape on the door jamb that allows clerks to estimate the height of each person as they pass through or run out after a robbery!) After 10 people had come through the door, what would be the best prediction for the height of the 11th person?

Answer: The average height of the 10 people that had already come through the door. In the absence of any other information, use the best evidence, which for the new clerk is the 10 people who have come through the door.

When the 11th customer arrives, their height adds to the pool and the average adjusts to include this added information. In other words, we update our understanding of the world in the light of the latest information.

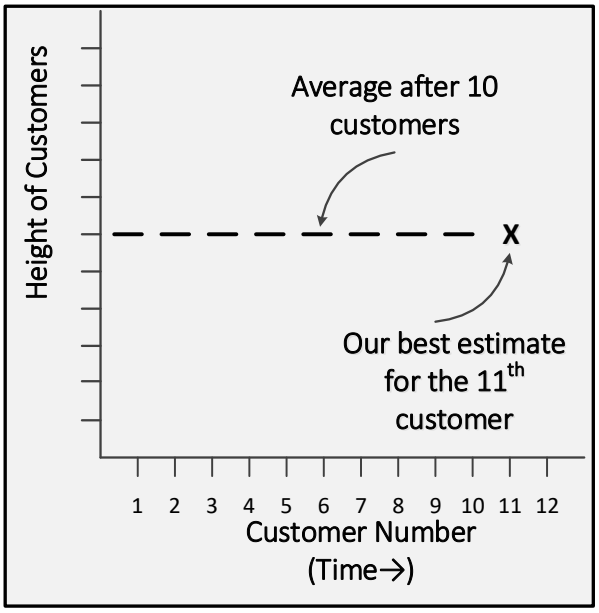


Figure 6: The average as the best predictor

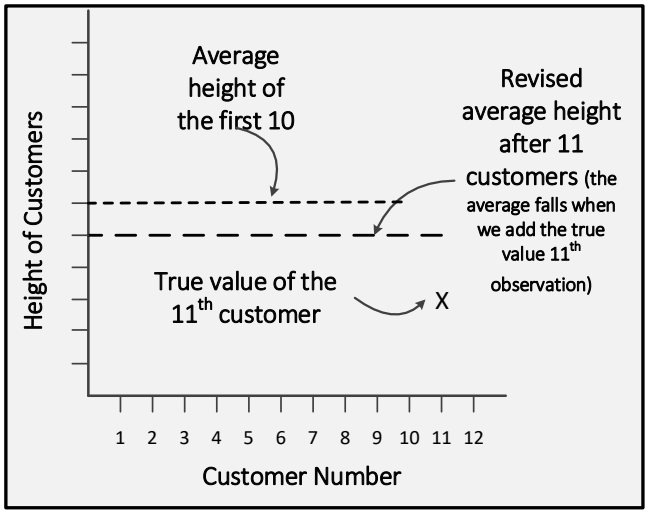


Figure 7: Adding an observation changes the average

Another way of saying this is that our knowledge is "conditional" on the available information. Any economic or business analyst (in fact anyone) who does not update their views considering current data will never make good decisions/forecasts. This is the essence of Bayesian statistics or decision-making, as mentioned in Module 5.

Let us continue with the thought experiment and imagine the clerk also notes the sex[2] of people coming through the door. It is generally true that men are, on average, 10 to 15 centimeters taller than women. (More on this just below.) Adding this second variable (sex) supports an improvement in our prediction since we can create an average height for women and an average height for men. So, when the clerk sees the 11th person in the parking lot and can also determine their gender, they can make a better estimate of their height. We say our prediction of height is "conditional" on the sex of the customer.

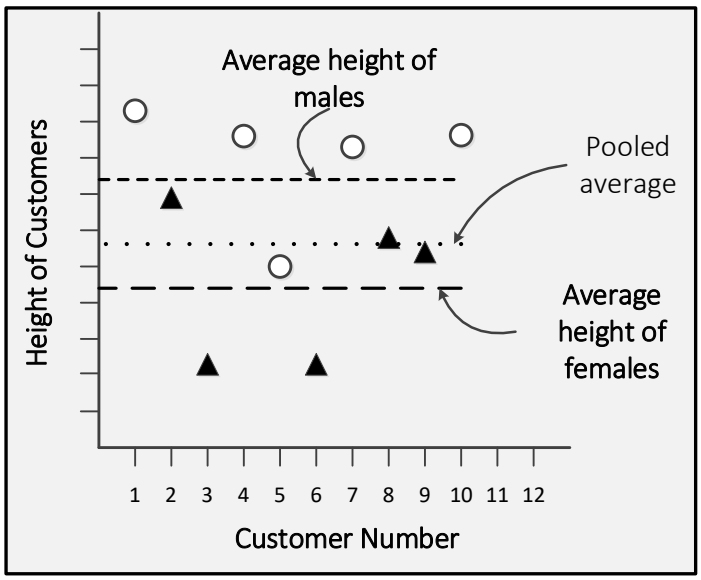


Figure 8: Adding sex as a predictor

We often start with a simple trend line experiment where time is a dependent variable. You can also perform this within a linear chart. See

Simple Trend Line Experiment.xlsx

Simple Regression.xlsx

2.2. Creating the regression equation

Regression (and all multivariate models) have two basic purposes in economics:

- explanation
- prediction

The regression model is an efficient way to analyse the structure of the data and to explain the variation in the dependent variable as a function of the independent variable. This model also allows one to predict the value of the dependent variable for potential values of the independent variables. The regression equation is really a **sentence** that summarizes the influence of the independent variables (causes or drivers) on the dependent variables (effects or outcomes).

$$\text{Height} = 165 + 5*D$$

(D = 1 for a man and 0 for a woman)

The predicted height for people coming through the door is 165 cm plus 5 cm if that person is a man, and 165 cm if that person is a woman. In other words, women have an expected height of 165

Independent variables that take on the value 0 or 1 are termed "dummy" or "indicator" variables.

They indicate a two-possibility state: in this case, either male or female. As shall become apparent, it makes no difference whether to designate (D) and standing for the male or the female as =1.

[2] Sex and gender are used interchangeably.

cm, and men have an expected height of 170 cm. The simplest regression states that a dependent variable "Y" is explained/predicted by a single independent variable "X", or

$$Y_i = a_0 + a_1 X_{1i} + e_i$$

The subscript "i" refers to different units of observation and here stands for individuals; other models could reference households, countries, schools, etc. With time series data, the subscript "t" is common and can reference years, quarters, months, etc.

Once we estimate the regression, the equation appears as

$$\hat{Y}_i = \hat{a}_0 + \hat{a}_i X_{1i} + \hat{e}_i ,$$

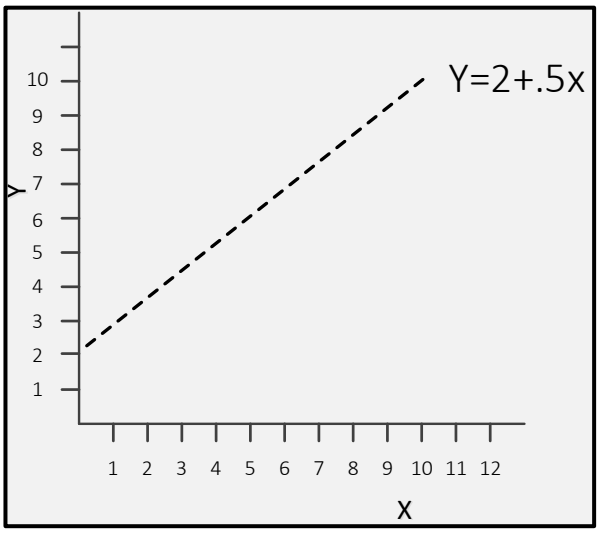
where $\hat{Y}_i$ is the estimated value of the dependent variable; $\hat{a}_0$ is the predicted value of the intercept (or the predicted value of Y when x is 0); $\hat{a}_1$ is the predicted value of the slope that measures the change in $\hat{Y}_i$ for a unit change in X; and $\hat{e}_i$ is the estimated errors, more commonly termed the residuals. Notice that X has no tilde (^) since it is not an estimated value.

The data for simple regression is a *flat file*, with the units of analysis – individuals, firms, countries, etc. for cross-sectional data, and weeks, months, years, etc. for time series data – as rows, and variables as columns. Here we have a dependent variable (Y) as salary and an independent variable seniority (years of service). Each row represents a different case (person). In this case, no column enumerating the person is needed, but often the cases have unique IDs, such as the family number in Module 2's Mincome data (see Module 2's

Mincome Baseline Extract.xlsx

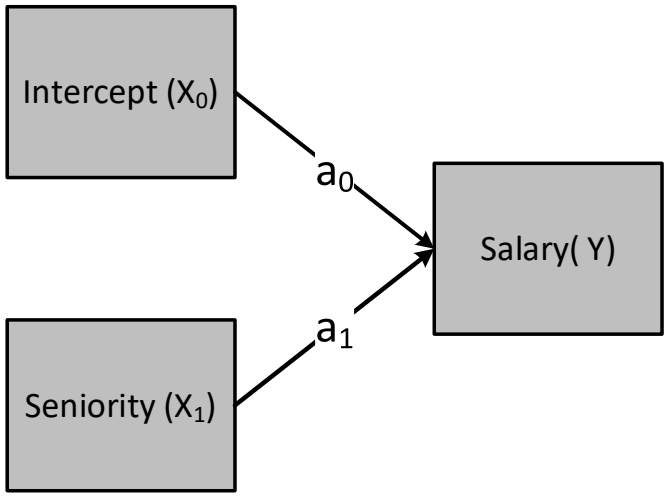
Salary	Seniority (Years)
24000	7.4
28000	15.9
22000	14.3
23000	5.1
29000	14.9
25500	7.2
23000	8.5
25000	9.0
29000	26.5
21000	21.2
20000	6.7
19000	12.7

An example appears to the right, with the associated graph, which reads as "Y equals 2 plus .5X." Another way to read it is "When X = 0, Y=2" or "When X = 10, Y=7." The intercept (a_0) = 2 and the slope (a_1) =.5



2.3. Three ways to talk about and interpret regression

Text, a block, diagram (causal path), or an equation are three ways to define a regression. Using the data from [Simpson's Paradox.xlsx], we can hypothesize that salary is a function of seniority (years of employment), *or* salary is a linear function of seniority. These are two ways to describe a regression relationship, with the second being more precise about the fact that it is a linear regression.



The block diagram shows in picture format what the general equation shows.

$$Y_i = a_0 + a_i X_{1i}$$

The only change to note is that the intercept is actually "variable" X_0 that never varies and always has the value 1. Most often we just write a_0, but the truth is that it is there.

The "variable" X_0 is odd, and you will never see it again in this course (at least not until Module 12). This is a "variable" with a constant value of 1, and it supports the estimate of the intercept term a_0. This becomes clearer in Module 12, which explains the matrix approach to estimating regression models.

The equation in Figure 9 has the following interpretation. The value of a_0 is equal to the value of Y (income) when X (years of seniority) is 0 is the "starting salary" and has a value of $15,000. The value of a_1 comes from basic trigonometry and states that every year adds $667, which means after 60 years on the job, someone can expect to earn $55,020 ($15,000+ $667*60).

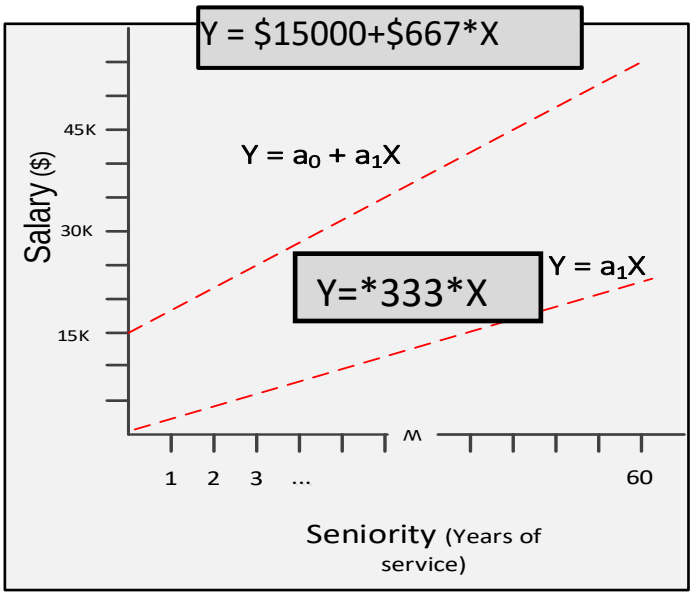


Figure 9: Regression example

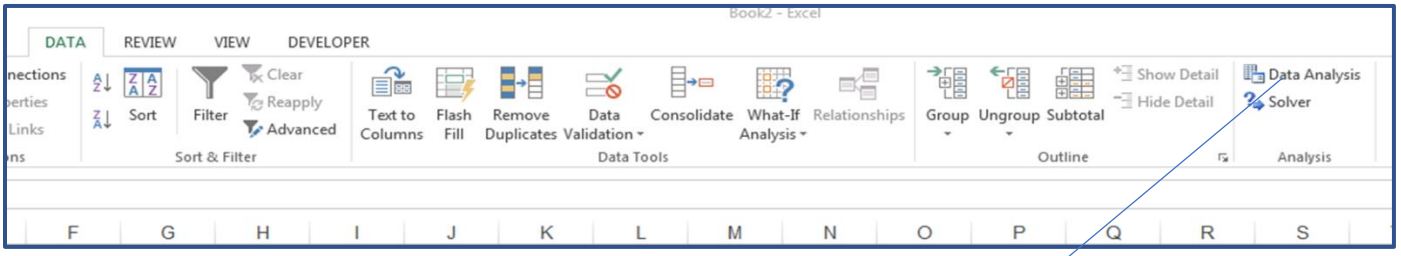
So, if the starting salary is forced to be 0, the annual increase is about $20,000/60 or $333. Deciding to suppress the intercept (setting $a_0 = 0$) is usually wrong (although in some specific circumstances certain theories do require it).

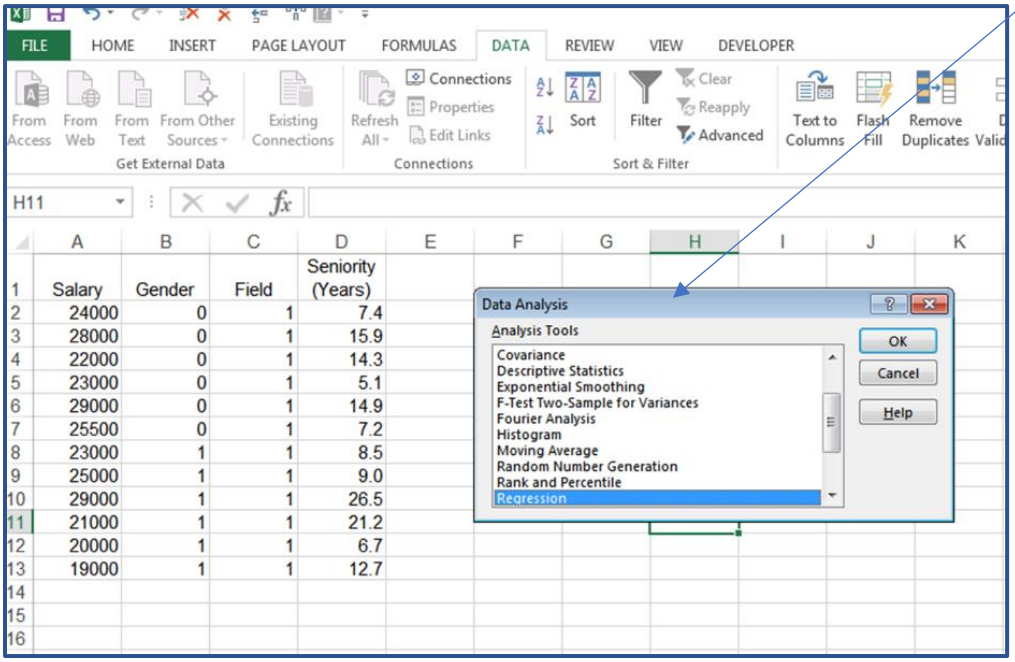
Another model, just using a dummy variable for sex, appears at; [Salary Data.xlsx

study this carefully to see all the features embedded in the model.

2.4. Simple (one independent variable) regression in Excel

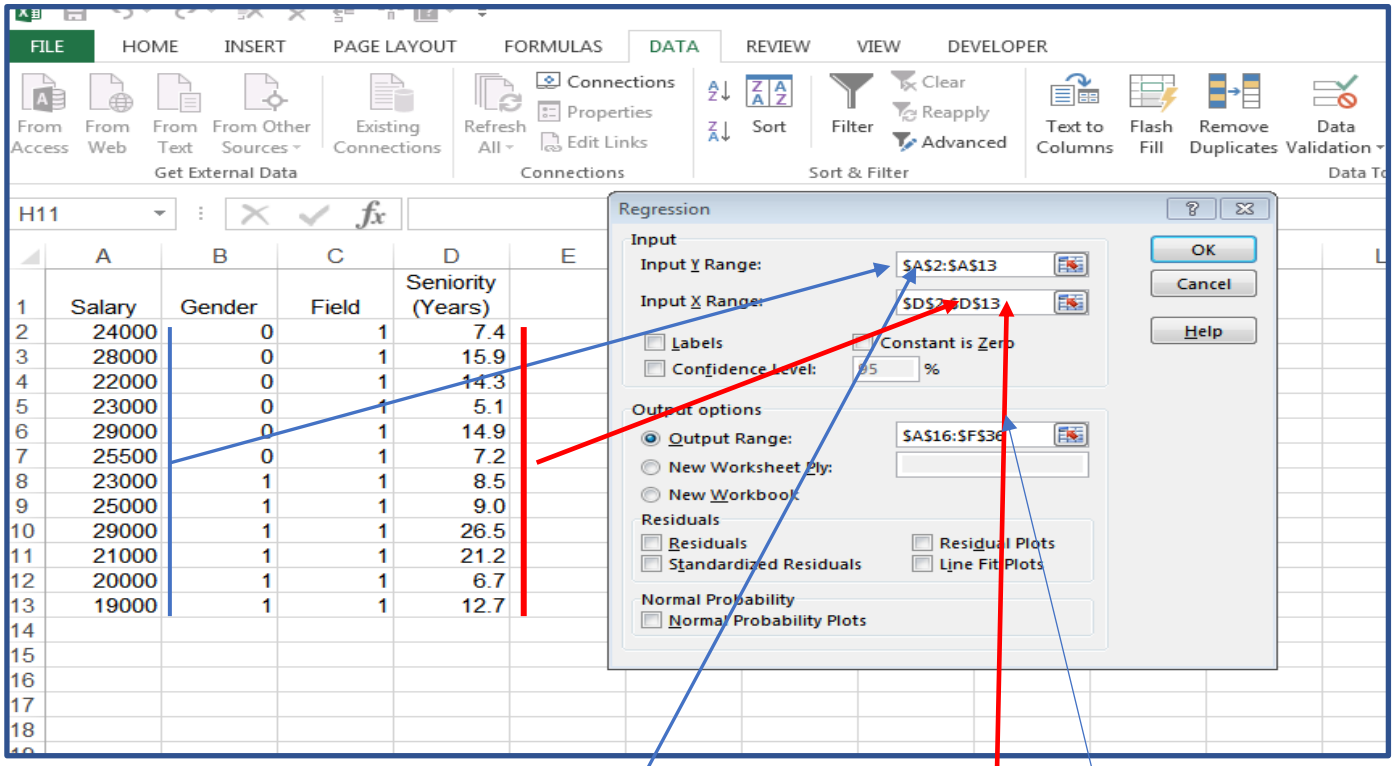
The regression model in Excel appears under the Data Analysis ToolPak.





Video: Estimating and interpreting simple regression

Module 6: Regression 1 – The Basic Model



	A	B	C	D
1	Salary	Gender	Field	Seniority (Years)
2	24000	0	1	7.4
3	28000	0	1	15.9
4	22000	0	1	14.3
5	23000	0	1	5.1
6	29000	0	1	14.9
7	25500	0	1	7.2
8	23000	1	1	8.5
9	25000	1	1	9.0
10	29000	1	1	26.5
11	21000	1	1	21.2
12	20000	1	1	6.7
13	19000	1	1	12.7

We use the menus to select the dependent (Y) variable and the independent (X) variable. We retain the intercept (by **not** selecting "Constant is Zero" and defining an output range on the current sheet where the results will appear). Selecting "New Worksheet" or New Workbook" are other options. See

Simple Regression.xlsx

Salary	Seniority (Years)	Sex	Field
24000	7.4	0	1
28000	15.9	0	1
22000	14.3	0	1
23000	5.1	0	1
29000	14.9	0	1
25500	7.2	0	1
23000	8.5	1	1
25000	9.0	1	1
29000	26.5	1	1
21000	21.2	1	1
20000	6.7	1	1
19000	12.7	1	1

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.365954369
R Square	0.1339226
Adjusted R	0.04731486
Standard E	3293.298515
Observatic	12

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regressio	1	16771015.59	16771016	1.546312	0.242042896
Residual	10	108458151.1	10845815		
Total	11	125229166.7			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	21667.42731	2132.902645	10.15866	1.38E-06	16915.02406	26419.83	16915.02406	26419.83056
Seniority (Y	190.5172	153.2094689	1.243508	0.242043	-150.8547702	531.8892	-150.8547702	531.8891702

Even the simple regression process in the Data Analysis ToolPak produces much information. In this course, we will not delve too deeply – save this for your econometrics courses!

The key elements for now are:

- The intercept
- The slope coefficient, which shows the increase in salary for an additional year of seniority

3. Multivariate regression

3.1. Regression as a weighted average

Linear regression is a form of weighted average. This is easy to see by comparing the formula for a weighted average to the formula for a regression model.

$$\overline{X} = w_1 x_1 + w_2 x_2 + \ldots w_n x_n \quad \text{weighted average}$$

$$\hat{Y}_i = \hat{a}_0 + \hat{a}_1 X_{1i} + \hat{a}_2 X_{2i} + \ldots + \hat{a}_k X_{ni} \quad \text{regression equation}$$

When we omit all independent variables, $\hat{Y}$ becomes the average ($\overline{Y}$), or $\hat{Y} = \overline{Y} = \hat{a}_0$. As noted above, the ^ denotes an estimated value. Y_i is the exact value for the "i'th" value of Y, for which a total of n values exists.

3.2. Multivariate (multiple) regression

Most economic and business data have **multiple influences**:

- Our income as adults depends on education, the cultural and social legacies of our family and friends, innate character and personality traits, and some luck.
- The profits of a firm depend on, among other things, the alignment of a product and consumer needs, access to skilled labour, management skills, general national and international economic conditions, and government policy.

Multiple regression attempts to capture multiple influences (*independent variable**s***) on outcomes (*dependent variables*), such as lifetime income or business profits. The extension of the simple linear regression to the multiple regression context is straightforward.

As with simple regression, multiple regression may use text, block diagrams, and algebra to specify the model. As an example of a text description, consider the data for Simpson's paradox and the models one may test using multiple regression.

Table 2: Possible regression models associated with Simpson's paradox			
Simple regressions	Salary is a linear function of sex (salary varies with sex).	$Y_i = a_0 + a_1X_{1i}$	1
	Salary is a linear function of seniority (salary increases with seniority).	$Y_i = a_0 + a_1X_{2i}$	2
	Salary is a linear function of field (salary varies with field [occupation]).	$Y_i = a_o + a_1X_{3i}$	3
Regressions with two independent variables	Salary is a linear function of sex and seniority (sex and seniority jointly determine salary).	$Y_i = a_0 + a_1X_{1i} + a_2X_{3i}$	4
	Salary is a linear function of sex and field.	$Y_i = a_0 + a_1X_{1i} + a_2X_{3i}$	5
	Salary is a linear function of seniority and field.	$Y_i = a_o + a_1X_{2i} + a_2X_{3i}$	6
Regressions with three independent variables	Salary is a linear function of sex, seniority, and field.	$Y_i = a_0 + a_1X_{1i} + a_2X_{2i} + a_3X_{3i}$	7
Y = Salary	X2 = Seniority (years)		
X1 = Sex (0 = Female, 1 = Male)	X3 = Field (1 = liberal arts, 2 = social science, 3 = engineering)		
To understand the equations, align the variables X1, X2, and X3 with their definitions.			

Three important points emerge from this table. First, the subscripts of the regression parameters (a_o, a_1, a_2, a_3) refer to their position in the regression and do not reflect the variable number. Look at the equations in detail. Most often, models use more descriptive variable names than X1, X2, etc. Here the variables could be named "salary," "sex," "field," and "years." As a rule, always name your variables with short (< 8 characters) and try to be descriptive. It will make the output from statistical processing much easier to interpret. Using caps is also good practice to set these off in your descriptions. So here we might name the variables as SEX, FIELD, YEARS, and SALARY.

Second, the ^ refers to estimated values and does not appear in general formats as in the table. Third, two of the variables – sex and field – have qualitative coding with the variables: SEX as a 0/1 variable to denote females and males, and Field having three values. CAUTION - As discussed below and in Module 7 , the *coding* for SEX is correct (and it makes no difference whether males are represented by 0 and females by 1), but the coding for FIELD is incorrect, as discussed in Module 7.

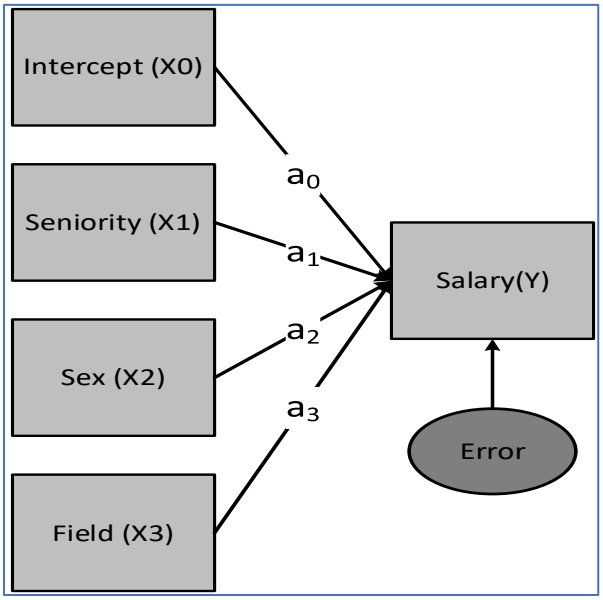


Figure 10: Multiple regression model

In this block representation of a multiple regression, a new term appears – error, usually represented by e_i. Multiple regression (in fact most models in econometrics) assume that the independent variables X_k have no error in measurement. However, the dependent variable Y has two sources of error – measurement and specification. Measurement errors stem from recording mistakes, survey errors, and other "noise" in the process of collecting and processing the data. Specification errors reflect many issues, such as 1) not including the correct independent variables, 2) using the wrong form of equation, and 3) violations of regression assumptions.

3.3. Extending the analysis of Simpson's paradox

In the analysis of Simpson's paradox above, rearranging the information in a table revealed the impact of lurking variables, which increase the insight into the factors that determine salary. Multiple regression can do the same task but more efficiently and with more insight. Note the adoption of a naming convention for the variables.

Note the labels in first row.

	A	B	C	D
1	SALARY	SEX	FIELD	YEARS
2	24000	0	1	19.29073
3	28000	0	1	23.64834
4	22000	0	1	22.21952
5	23000	0	1	8.64679
6	29000	0	1	21.49346
7	25500	0	1	14.46658
8	23000	1	1	8.452595
9	25000	1	1	24.69793
10	29000	1	1	16.1944
11	21000	1	1	8.954525
12	20000	1	1	14.78708
13	19000	1	1	6.962435
14	25000	1	1	19.07648
15	28000	1	1	22.85081
16	25500	1	1	18.1885
17	22000	1	1	15.72742
18	25000	1	1	17.8042
19	26500	1	1	8.109229
20	22000	1	1	18.76467
21	27000	1	1	13.23686
22	24000	1	1	10.59066

So far, we have explored the explanation for salary levels (SALARY) using only seniority.

Does SEX help **explain** salary differences? Or does knowing the FIELD of the worker increase our ability to **predict** the SALARY of a worker who is not currently part of the data?

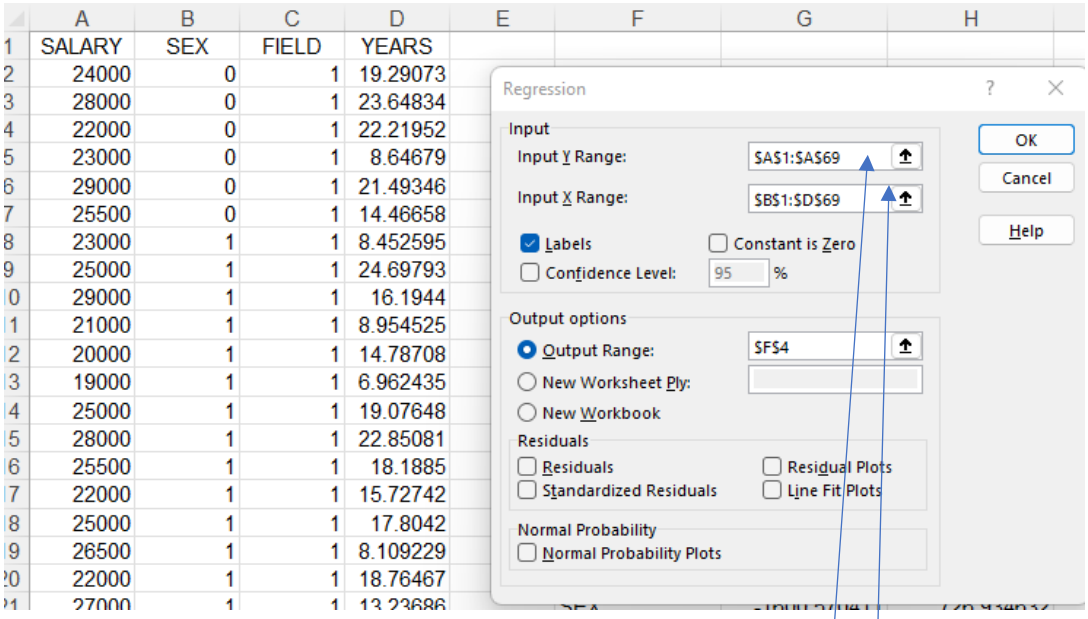
Here is a portion of the data we will use. The full data are in the spreadsheet.

SalaryData.xlsx

Video: Salary Regression

The results of the multiple regression confirm the importance of field in determining salaries, with sex and seniority having less influence.

Here is the set-up for running a multiple regression using the ToolPak. Note that the Input Y range defines a vector (Column 1), which is Column A, and the Input X range is an array comprising cells B1:D69. Also, each column has a variable name in Row 1 and the labels option has a check to instruct Excel to not include this information in the calculation of regression parameters.



	A	B	C	D
1	SALARY	SEX	FIELD	YEARS
2	24000	0	1	19.29073
3	28000	0	1	23.64834
4	22000	0	1	22.21952
5	23000	0	1	8.64679
6	29000	0	1	21.49346
7	25500	0	1	14.46658
8	23000	1	1	8.452595
9	25000	1	1	24.69793
10	29000	1	1	16.1944
11	21000	1	1	8.954525
12	20000	1	1	14.78708
13	19000	1	1	6.962435
14	25000	1	1	19.07648
15	28000	1	1	22.85081
16	25500	1	1	18.1885
17	22000	1	1	15.72742
18	25000	1	1	17.8042
19	26500	1	1	8.109229
20	22000	1	1	18.76467
21	27000	1	1	13.23686

The results of the regression appear in Figure 11.

Simpson's Paradox Income vs Sex vs Field vs Seniority.xlsx

Notice the dependent variable range ($A$1:$A$69) and the ***array*** that defines the three dependent variables ($B$1:$D$69) being the columns B, C and D.

SALARY REGRESSED ON SEX, FIELD, AND YEARS

Regression Statistics	
Multiple R	0.973108157
R Square	0.946939484
Adjusted R Squa	0.944452273
Standard Error	2565.009073
Observations	68

ANOVA

	df	*SS*	*MS*	*F*	*Sign*
Regression	3	7514646327	2504882109	380.7233205	9
Residual	64	421073378.9	6579271.545		
Total	67	7935719706			

	Coefficients	Standard Error	t Stat	P-value	Lowe
Intercept	12734.98656	1220.956705	10.43033427	1.95362E-15	1(
SEX	-1600.570411	726.934632	-2.201807893	0.031290093	-3(
FIELD	11275.45314	579.0511506	19.47229208	1.37221E-28	1(
YEARS	126.3553933	55.2920797	2.285234956	0.025622005	1(

SEX=0 (M) FIELD = 1 Social sciences-education
SEX=1 (F) FIELD = 2 Life sciences
FIELD = 3 Engineering

The regression reads as follows:

- When gender is 0 (male), the field is 1 (liberal arts), and years of seniority is 0, salary will be $12,735 (the value of the intercept plus $11,275 = $24,010).
- Being female reduces this by $1,600.
- Increasing field by 1 increases salary by $11,275.
- Each year of seniority adds $126 in annual income.

Figure 11: Results of the regression (**SalaryData.xlsx**)

Notice that the column labels appear in the regression output. If you had not labelled the columns in the Excel sheet, then generic names would appear, making it harder to interpret the results.

3.4. Multiple regression – Synopsis thus far

This Module has presented regression as

- increasing understanding of what factors might be plausible explanations for the variation in a dependent variable; and
- as a weighted average that supports the assessment of the relative importance of each included independent variable (regressor) for a prediction of the dependent variable.

Regression has a symbolic representation, a mathematical expression, and a plain language description. The model in this Module is a linear regression – the weighted sum of the independent variables produces the predicted value of salary based on specified values of three variables SEX, FIELD, and YEARS).

Here are some cautions:

1. The coding used for FIELD reflects a common mistake, which we fix in the next Module.
2. Changing coding will change the coefficients and may change the interpretation.

3. Extending the model into areas beyond the range of the independent variables, such as predicting salary with 100 years of seniority, creates errors. This emphasizes the risks of analysis beyond the scope of the data used to estimate the model.
4. Thus far, only a small part of the regression output figures in the analysis – the next Module explains some important *regression diagnostics.*
5. While the linear regression model is the most common, in many cases, a non-linear model is more appropriate. Module 7 discusses these approaches.

4. Summary

This Module introduced the regression model. It is an important extension to simple bivariate correlations that often (usually) mislead. The regression model introduces other confounding variables that *together* explain variation in a dependent variable. But be careful. A regression model does not offer a causal explanation. Just because we name some variables to be "independent" and we hypothesize them to be causes does not make them so. Regression is merely a weighted average that can "explain" variation in another variable we happen to designate as dependent (or the effect).

Annex: Key Excel functions and formulas

Excel functions and formulas		
Function	**Example**	**Explanation**
=RANK(Cell Ref, Array,order)	=RANK(A2,A1:A3045,0) The two rank functions support the construction of the Spearman rank order correlation.	Finds the rank of the number in cell A2 in the array A1:3045, in descending order, with the "0." Any number in the third argument finds the rank in ascending order.
=RANK.AVG(Cell Ref, Array,order)	=RANK.AVG (56,A1:A3045)	Finds the rank of the value 56 in the array A1:A3045 in ascending order (third argument is optional and assumed to be 0 if missing).
=SLOPE(known y's, known x's)	=SLOPE(A1:A14,F1:S1) If you have a single variable regression and just need the slope and intercept, use these quick tools.	Finds the slope of the regression, with Y as the dependent variable (intercept included) and X as the independent variable. Note the known x's and known y's are two vectors of the same dimension.
=INTERCEPT(known y's, known x's)	=INTERCEPT(A1:A14,F1:S1)	Finds the intercept of the regression, with Y as the dependent variable and X as the independent variable. Note the two arrays must have the same dimension.