Module 7: Regression 2 – Extending the regression model

Learning Goal for Module 7

Module 7 introduces more sophisticated specification of dummy variables, assessing the quality of the regression of the regression coefficients, and using regression as a forecasting model.

By the end of this Module, you will:

- Be able to create dummy variables for any situation
- Interpret the coefficients of a multiple regression
- Assess regression quality by testing the statistical significance of individual coefficients and the overall explanatory power of the model
- Use regression to predict and forecast.

1. Introduction

The regression model introduced in Module 6 used a dummy variable (also know as indicator or qualitative variable) to define sex. In that model the indicator variable (X_1) became 0 for a man and 1 for a woman. Exercise 4 in that Module explored the idea that it does not matter how one codes gender (it could be 1 for a man and 0 for the woman), but such reversals change the sign of the coefficient a_1 and therefore the interpretation of the regression.

As a rule, use “1” for the state you *expect* will show the increase in the intercept or slope. Moving from the state of “0” (female) to a state of “1” (male) results in a positive change for the intercept, making the calculation a little more intuitive. With the variable Sex coded 0 for a female, intercept stands for the salaries for women, when *all other* independent variables take the value 0. The coefficient on the dummy variable (Sex) then reads the salary boost for males over females.

Of course, the data may not confirm your expectations. The sign could be negative and the coefficient not statistically significant. It is always best to work through the implications of the dummy variables, by setting all variables to “0” and then vary the value of the dummies between 0 and 1 and calculate the implication.

But before exploring regression quality, this Module corrects a mistake in the salary model developed as part of Module 6.

1.1. Dummy variables

The issue is the coding of the variable field which used the following scheme

- Social Sciences = 1

- Life/Physical Sciences = 2
- Engineering = 3

The mistake with this coding is that it embeds the idea that Life/Physical Science are "worth" twice as much as Sciences and Engineering is "worth" three times as much as Social Sciences. A statistical model should never presume or embed values within the data construction. To see how to correct this, recall a dummy variable turns on and off and takes the value of 0 or 1 to signal a two-state variable such as sex male/female, or buyer/non-buyer. This may imply that this technique applies only to variables that assume two states, such as gender (male or female, buyer, or non-buyer…) which is not true. Variables such with fields coded as three values can use a set of ***two*** dummy variables; variables with k discrete states can use k-1 dummy variables. How does this work in the case of the salary regression problem?

In general, K-1 binary (0-1)dummy variables can describe a variable with "K" discrete states

Define a new variable **FIELD 1** as equal to 1, for Physical Sciences and 0 otherwise; and define **FIELD 2** as equal 1 for Engineering and 0 otherwise. Using this logic, when Field 1 = 0 and Field 2 = 0 implicitly defines the Field as Social/Life Sciences.

Transforming Field (coded 1,2 3) into Field 1 and Field 2 uses the =IF statement. Recall from Module 2 that the =IF() function is a workhorse of Excel; here it creates dummy variables. In the Salary data, we need to create two dummy variables from one variable (Field) that originally coded as 1, 2, and 3. Note the construction of the =IF statement. It reads "If C2 = 3, then set the target cell (F2) = 1, otherwise set it to 0."

See Salary Data with Field Fixed.xlsx

Use F1 in Excel to explore these operations.
=AVERAGEIF
=SUMIF
=COUNTIF
See the Annex to this module.

Video: Using =IF to Code Dummy Variables

Exercise 1 presents the original salary data and asks you to correct the coding, re-run the regressions and compare the interpretations from the two models.

Video: Creating Monthly Dummies

1.2. Non-linear regression

Here the term "non-linear" refers to *variables*, not non-linear in *parameters.* Compare these two models; the first is non-linear in variables and linear regression techniques can estimate a_0, a_1, and a_3, but the second is a non-linear expression that requires more sophisticated methods.

$$Y_i = a_0 + a_1 X_{1i}^{\;2} + a_2 X_{2i}^{\;1/2}$$

$$Y_i = a_0 \frac{a_1 X_{1i}}{a_2 + X_{1i}}$$

The linear regression can adapt to include certain non-linear specifications. Basic algebra can change variables such as power (e.g., $X_k^{\;2}, X_k^{\;3} \ldots$) and logarithms(LN, LOG10). For certain equations, applying operators to an entire expression renders it to the linear form such as celebrated Cobb-Douglas model.

$$Y = a_0 X_1^{a_1} X_2^{a_2} ,$$

where Y designates output and X_1 and X_2 factors of production such as capital and labour. This is a multiplicative equation, but using logarithms (natural or to any base), creates a standard linear regression

$$\ln Y = a_o + a_1 \ln X_1 + a_2 \ln X_2$$

.

The Cobb-Douglas *functional form* has the interesting property that by taking logs of both sides, the estimated coefficients a_1 and a_2 measure elasticities as discussed in Section 1.4.

Logarithms also rescale data such as in **Figure *1*** and Figure *2*. **Figure *1*** shows the per capita incomes growing at increasing rates, especially for China and the US. We can create a more insightful chart by changing the vertical scale to logs. This produces Figure *2* shows that China and the US are now close in terms of rate. Whenever series grow at exponential rates, using logarithms to “linearize” the trends reveal the growth rates more accurately.

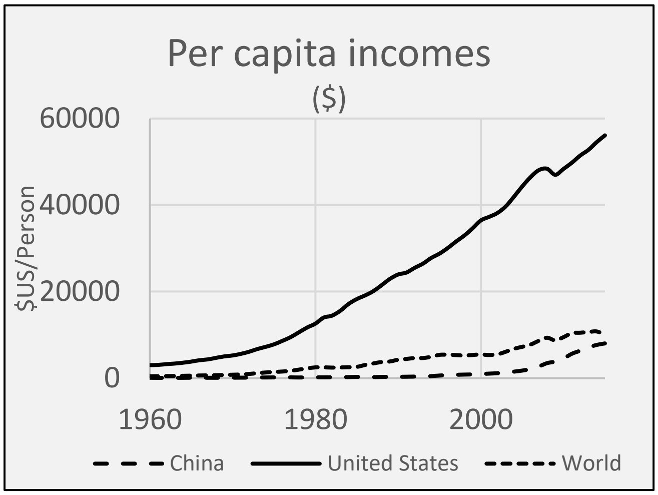


Figure 1: Per-capita incomes

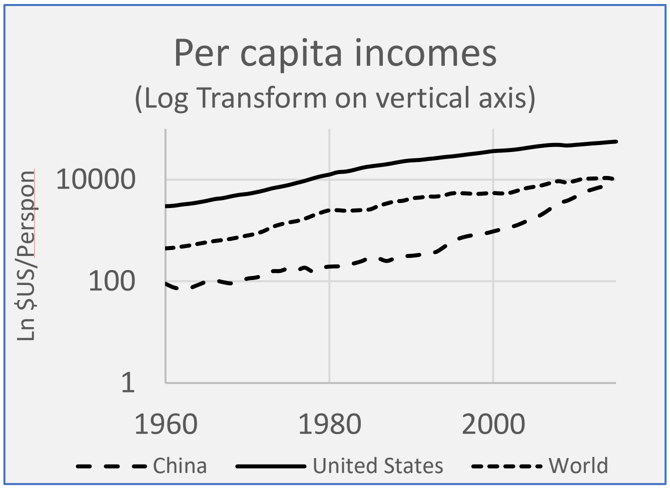


Figure 2: Log transform on per-capita incomes

1.3. Specification

The term "specification" refers to three elements of an econometric model:

- Functional form – is the model linear, additive and with untransformed variables (no power or logarithmic transformations, i.e., the standard model), or is it multiplicative with variables transformed such as the Cobb-Douglas model. Econometricians often use mathematics to transform theoretical models into linear additive forms since these are often easier to estimate and interpret.
- Variables – inclusions/exclusions refer to the choice of variables. For example, the salary model has Sex, Seniority and Field as explanatory(independent variables). What about years of education as another independent variable? Often, datasets do not include all the relevant variables because of information deficits, or the analyst throws every possible variable into the equation, which is usually poor practice.
- Complexity – often relationships exist with the set of independent variables, or the dependent variable can affect an independent variable (reverse causality), or a model may include multiple equations.

This course does not explore issues of specification, but it is important to be aware that the linear additive model used in Modules 6 and 7 are just the start of econometrics.

1.4. Elasticities

We often use a linear regression model of demand to measure price and income elasticities. To review, the regular regression coefficient has a simple interpretation. As an example, the linear regression models

$$Y_i = a_o + a_1X_1 + a_2X_2$$

a_0 is the value of Y when X_1 and X_2 both = 0 (this is also the simple average of Y)

a_1 is the change in Y for a unit change in X_1 (holding X_2 constant)

a_2 is the change in Y for a unit change in X_2 (holding X_1 constant).

An important measure in microeconomics is *elasticity*, which measures the percentage change in Y for a one percent change in X. Literally, we define ε_1 as the elasticity of Y with respect to X_1,

$\varepsilon_1 = \left[\frac{\Delta Y}{Y}\right] / \left[\frac{\Delta X_1}{X_1}\right]$, which after a little arrangement becomes $\left[\frac{\Delta Y}{\Delta X_1}\right] * \left[\frac{X_1}{Y}\right]$.

This formula defines the arc elasticity at a specific value of X_1 and Y; typically, computer programs offer the elasticity at the mean of X1 and the associated mean of Y. From calculus, we can derive dY / dX_1 as in the change at a point and is simply *a_1* , which implies for a specific regression equation $e_1 = a_1 * (\bar{X}_1 / \bar{Y})$, the point elasticity at the mean of *X1* and the mean of Y

Example: Imagine the regression result is $Y = 134 + 2.3X_1 - 3.08X_2$ and $\bar{Y}$ = 23.6, $\bar{X}_1$ =14.1, and $\bar{X}_2$ = 7.1, then ε_1=2.3*(14.1/23.6) and ε_2=-3.08 or (7.1/23.1).

Introductory economics and even intermediate microeconomics usually refer to elasticities in the context of how total revenue responds to changes in prices and incomes. It is a generic measure that applies to any multivariate model, measuring the response of the dependent variable to a change in any independent variable at any value of the independent variable.

Interesting Fact: The regression line pivots on the mean of X_k and the mean of Y. In other words, the means of all the independent variables ($\bar{X}_k$) and the mean of the dependent variable $\bar{Y}_i$ lie on the regression line.

For most expressions the elasticity varies with the values of the {Y, Xi} pair. Computer programs that offer measures of elasticity, usually do this for the mean value of the dependent and independent variables. Special forms such as the Cobb-Douglas have a constant elasticity for all values of {Y, Xi}. A final note ... an that elasticity exists between the dependent variable and *one* independent variable, implying that a demand equation with three independent variables (own price, the price of a substitute, and income) will offer three elasticity measures.

2. Assessing regression quality

How do we know whether a regression is "good?" What do we mean by good? How can we do to make a regression better?

Before answering these questions, here is a caution. At the start of their learning, aspiring economic analysts will develop a checklist of regression requirements. They will adjust regression models to meet one or more of these tests, with thinking about whether the changes make sense, align with other work, or align with theory. If you find something startling or unusual in your regression results, it is much more likely that errors exist in the data, or you have specified a strange model.

By good, is usually meant whether the independent variables explain the variation in Y (dependent variable). If we see all sorts of different values of salary in an organization, but when we correlate Sex, Seniority, and Field with Salary , we see no relationship, a regression model that tries to explain the variation in Salary (the different salaries) is likely to offer little insight. To understand the basis for quality assessment ion regression models, it is worth stepping back to the early days of astronomy.

2.1. Assumptions of the linear regression model

The telescopes of the 17th centuries were major scientific innovation, by far from the precision instruments of today. Errors in observation occurred because of the crude instruments and because atmospheric conditions and dust obscured the stars. Astronomers were the first to develop and use a method for averaging the observations on planets and stars.

The important insight was that adding observations and averaging, reduces error. More data leads to better understanding. Legendre, the famous French mathematician developed the *method of least squares,* which has evolved to the general linear model with the following assumptions

1. The basic form is linear and additive $\hat{Y}_i = \hat{a}_0 + \hat{a}_1 X_{1i} + \hat{a}_2 X_{2i} + \ldots + \hat{e}_i$. Notice that *Y* and the coefficients a_1 and a_2 appear as estimates (^) and that a new term e_i now appears
2. The dependent variable Y_i has measurement error, but we assume the independent variables $X_1, X_2 \ldots$ have no measurement error.

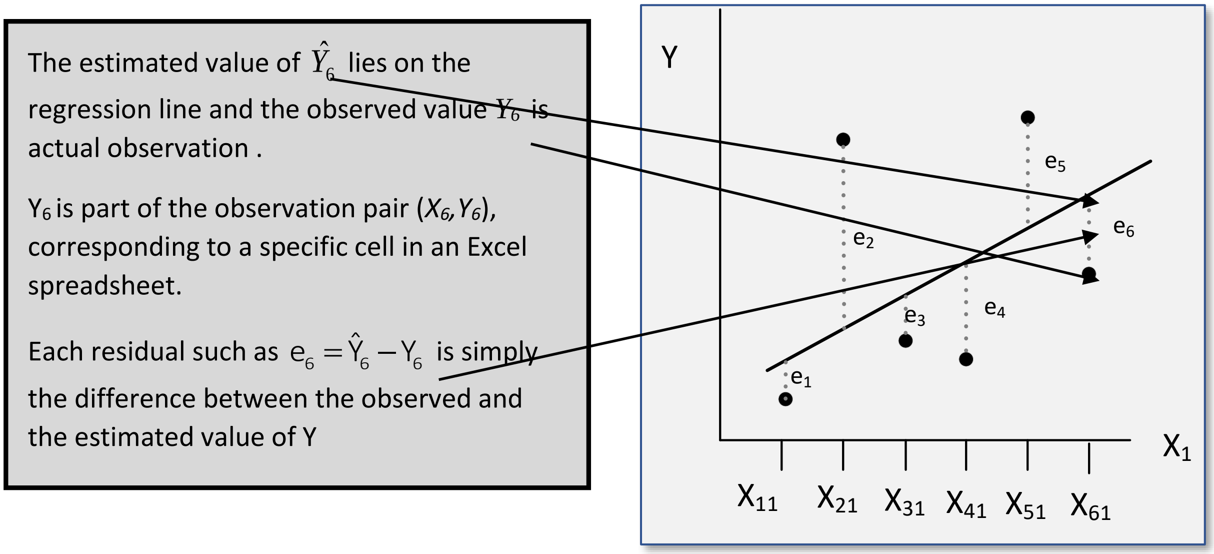
3. The estimated regression model designated as $\hat{Y}_i = \hat{a}_0 + \hat{a}_1X_{1i} + \hat{a}_2X_{2i} + \ldots + \hat{e}_i$ supports the measurement of residuals $e_i = (\hat{Y}_i - Y_i)$. The general linear model *assumes* that the residuals have a normal distribution with a constant variance.
4. The method of least squares is to find those $\hat{a}_0, \hat{a}_1, \ldots$ that makes $\sum_{i=1}^{n} e_i^2$ as small as possible.

2.2. Methods of least squares

One might try to "eyeball" or guess the best straight line that summarizes a linear relationship between X and Y. However, the most common rule for creating the regression line is the method of least squares.

> An important assumption of the linear regression model is that we specify a dependent and one or more independent variables before any statistical analysis. The literature and theory are the common starting point for developing a model. We assume that Y has no measurement error, while the independent variables X_k are fixed and known – no measurement error.

The goal of regression is to find the line that minimizes the error (vertical distance) between the observed value of Y and the estimated value of Y. From If we just summed the e_i, the positive and negative values cancel, so just like measuring the variance, each deviation must be squared and then added. (Review the calculation of the variance in Module 2.) The key notion is that by adjusting a_0, a_1, a_2..., the variance of the deviations also varies and for some unique combination, becomes minimized.

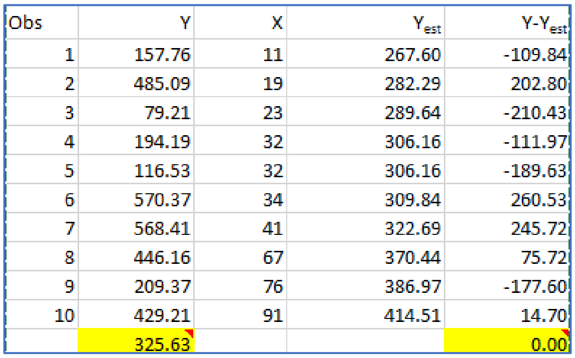


The estimated value of $\hat{Y}_6$ lies on the regression line and the observed value Y_6 is actual observation .

Y_6 is part of the observation pair (X_6,Y_6), corresponding to a specific cell in an Excel spreadsheet.

Each residual such as $e_6 = \hat{Y}_6 - Y_6$ is simply the difference between the observed and the estimated value of Y

The spreadsheet **[Regression Variation and RSQ 1.xlsx]** illustrates the sum of square residuals. Y_{est} or $\hat{Y}$ results from inserting the values of X into the estimated equation from the regression option in the Data Analysis ToolPak Option. Notice that the residuals ($Y-Y_{est}$) sum to 0 and we need to square these for the same reason we needed to square the deviations in a variance. The sum of squared residuals $(Y-Y_{est})^2$, divided by n-2 (here = 8) is the variance of the regression. Study the calculations in this spreadsheet carefully to understand the basis of regression models



Obs	Y	X	Y_{est}	$Y-Y_{est}$
1	157.76	11	267.60	-109.84
2	485.09	19	282.29	202.80
3	79.21	23	289.64	-210.43
4	194.19	32	306.16	-111.97
5	116.53	32	306.16	-189.63
6	570.37	34	309.84	260.53
7	568.41	41	322.69	245.72
8	446.16	67	370.44	75.72
9	209.37	76	386.97	-177.60
10	429.21	91	414.51	14.70
	325.63			0.00

Before considering measures of regression quality, let's consider exactly how to estimate a_0 and a_1 from the simple linear model $Y_i = a_o + a_1X_{1i} + e_i$. Any econometrics text will detail the derivation of the equations for estimating the intercept and slope, which are

$$\hat{a}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}, \text{ and}$$

$$\hat{a}_0 = \overline{Y} - \hat{a}_1\overline{X}$$

See **[Normal Equations.xlsx]** to see how these formulas produce the same results as the Data Analysis ToolPak.

Video: Regression – Normal Equations

3. Assessing regression quality

Quality in regression analysis has two general dimensions

- *Explanation:* Regression models that explain more variation in the dependent variable are more useful and therefore have higher quality
- *Prediction:* Regression models that predict well are more useful and therefore have higher quality.

Of course, more detail exists, and these are just general statements.

Many statistics assess the quality of a regression – we will only cover a few of the more important ones here (specifically the R^2 and “t” scores). To start, measures of regression quality depend on the idea of ***error,*** which means that the “residuals” – the data variation not explained by our model – becomes the central element of assessing regression quality.

3.1. Goodness of fit – concept

The goodness of fit (R^2) is the most is and abused measure of regression of regression quality

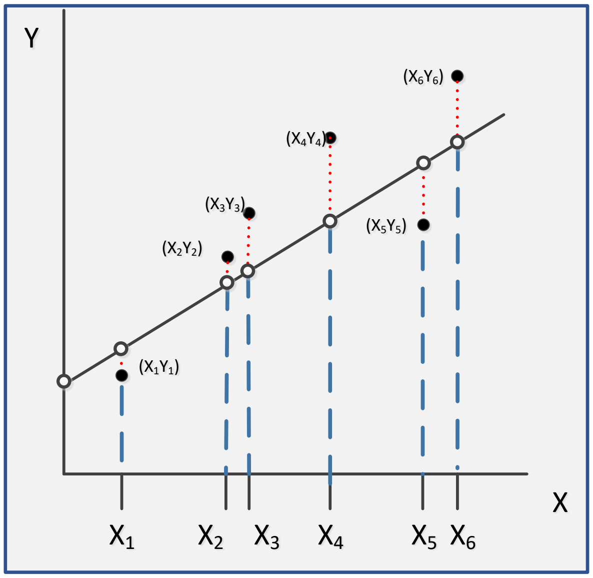


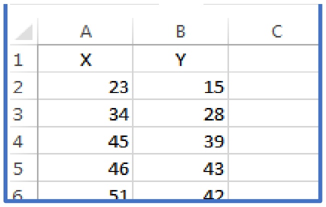
Figure 3:Goodness of fit

The figure at the left is the fundamental picture of a regression equation data consist of six pairs of numbers (x1,y1);(x2,y2)…(x6,y6). A typical excel spreadsheet with these data would appear as follows

e_6

e_5

Here (x1,y1) correspond to the pair number 23,15 and are in cells A2 and B2. In Figure 3, they are the solid dots at (X_1,Y_1)



	A	B	C
1	X	Y	
2	23	15	
3	34	28	
4	45	39	
5	46	43	
6	51	42	

The estimated equation $\hat{Y}_i = \hat{a}_0 + \hat{a}_1 X_i$ appears as the straight line. The “clear circles” are the estimated values of $\hat{Y}_i$ that form the regression line .

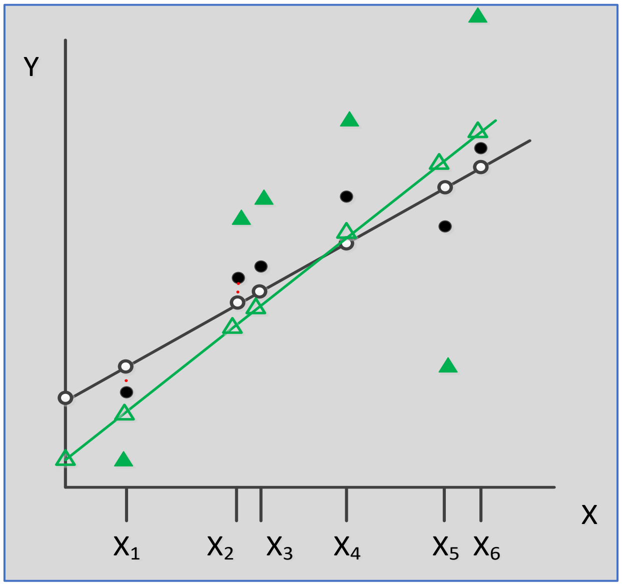
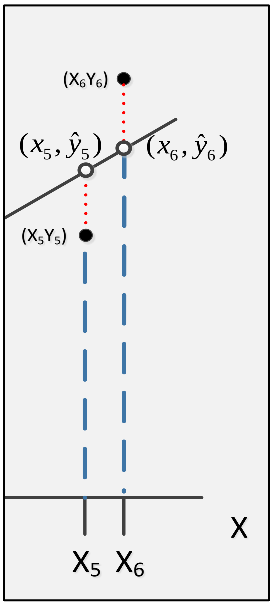


Figure 4: Two regression lines

The figure shows two possible scenarios, where the same set of X values are associated with different Y values appearing as black dots and green triangles. Each set of data supports different regression lines

In general, the “black” points tend to cluster and support the black regression more than the “green triangles” support the green regression. We say that the black regression has a higher goodness- of- fit.

The measure for goodness-of-fit for a regression is the R^2, which varies between 0 and 1, where 0 means no fit and 1 suggests a perfect fit.



The figure at the left shows the detail around points (x_5,y_5) and (x_6,y_6).

The solid dots are the data we observe, while the white circles are the estimated values shown as the data pairs $x_5, \hat{y}_5$ and $x_6, \hat{y}_6$

We say that $\hat{y}_5$ is the estimated value of y given x_5 . Note, ***and this is important***, once we decide that x is the independent variable, its value remain fixed. If we change our minds and come to see X as the dependent variable and Y as the cause, we will estimate $\hat{X}_i = \hat{b}_0 + \hat{b}_1 Y_i$ Our theory guides us in the designation of independent and dependent variables:

Example: In most cases advertising spending causes increases in sales not the reverse, although other factors boost sales (income, employment, prices…).

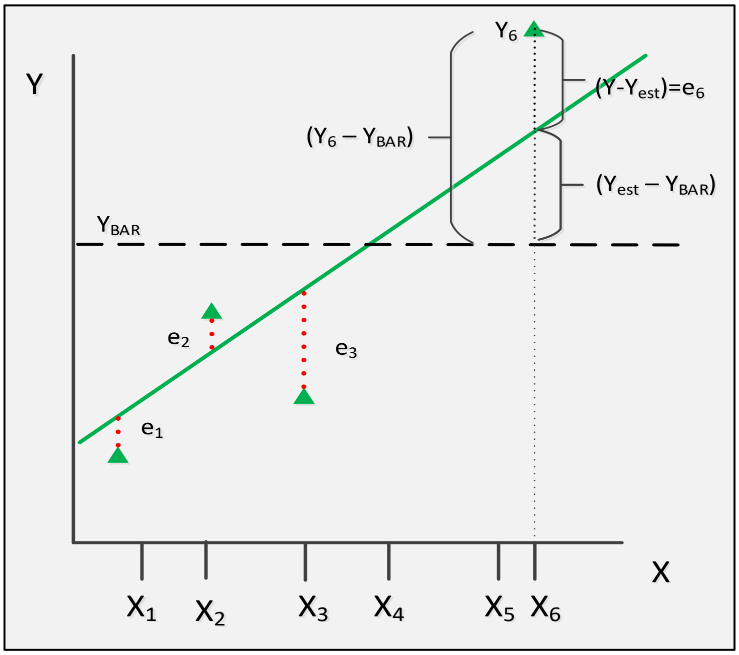
Example: Most macro theories view consumption as a function of (depends on) income.

Example: Increases in prices will usually reduce quantity demanded. (Recall the possible exceptions from microeconomics for luxury goods and Giffen goods.)

The difference between the estimated value and the actual values of y $\hat{e}_1 = (y_1 - \hat{y}_1); \hat{e}_2 = (y_1 - \hat{y}_2); ... \hat{e}_n = (y_n - \hat{y}_n)$ are the “residuals of the regression. These sum to 0, but measures of regression quality depend on the sum of the square of the residuals.

3.2. R^2 as a measure of goodness of fit

Explaining the total variation in Y forms one of the goals for a regression (prediction is the other). Recall the total variation is really the variance in Y, and we measure that, like any variance, by taking deviations $(Y_i - \bar{Y})$, squaring them $(Y_i - \bar{Y})^2$ and summing across all observations $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$. No regression model can ever explain all the variation in Y. The total variation (total sum of squares – **SST**) is comprised of the explained sum of squares (**SSE**) plus the residual sum of squares (**SSR**)



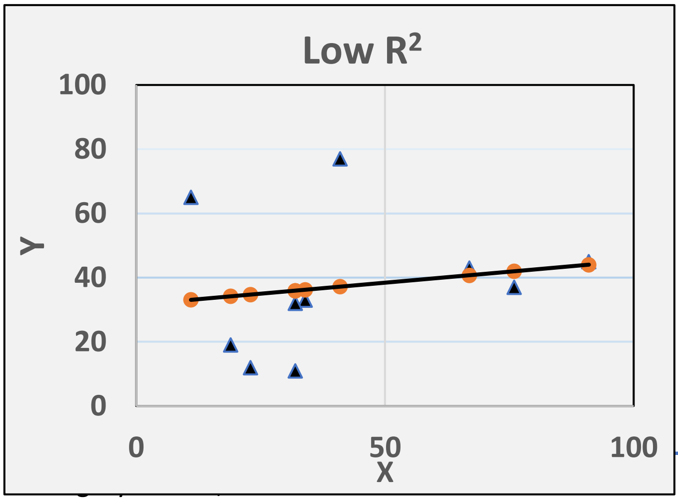
SST = SSE + SSR

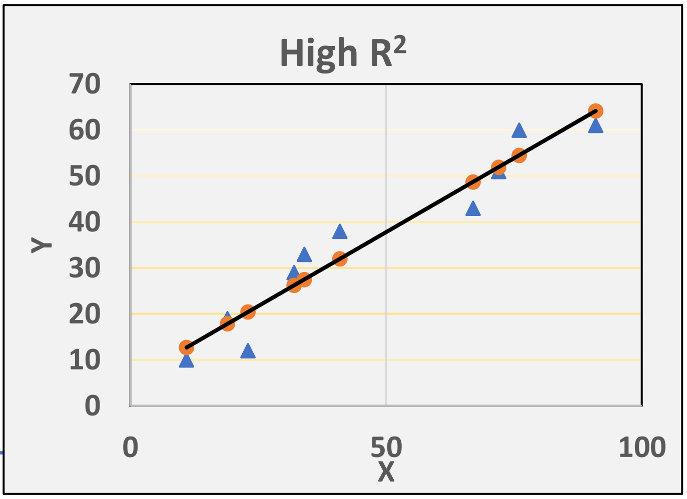
R^2 = SSE/SST = 1 – SSR/SST

Intuitively, the wider the spread of residuals, the lower the goodness of fit.

Wide spreads in residuals are the same as having a high ratio of unexplained to explained variation.

Video: Regression - R Square





> The two figures show two situations with different regression. R^2 is insufficient to assess regression quality on its own. The reliability of the parameters a_0 and especially a_1 are important in validating theoretical models. For this we need tests on the statistical significance of regression coefficients

3.3. Testing the regression parameters - t tests

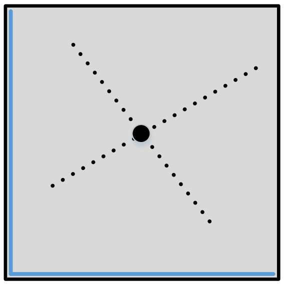
More important than the R^2, the regression parameters a_k are central to testing the quality of a model. To recall, a null hypothesis is any statement/belief that is "nullifiable" (falsifiable) because evidence shows a low likelihood of making a mistake in rejecting that statement.

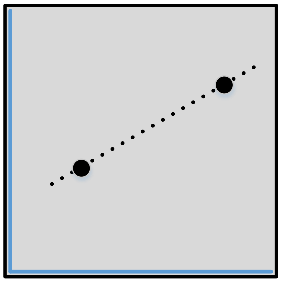
Recall the Z score from the standard normal distribution in Module 5, where a typical null hypothesis was H_0: Male wages and female wages are equal, or H_0: $\mu_m-\mu_f = 0$ (for a two tailed test) and H_0: $\mu_m>\mu_f$ when testing whether male wages exceed female wages.

The t distribution is one of the most useful probability functions for testing statistical reliability of regression parameters. The Z test works only when we have full information on the population (mean and variance/standard deviation).The t distribution supports hypothesis testing in small samples, where we can estimate means and variances only from the sample information. William S. Gosset developed this statistic to support quality control in testing batches of Guinness beer in Dublin.

The t distribution supports statistical decision making in small samples and is of statistical tests that require us to acknowledge *degrees of freedom*, the F and Chi-square tests being the other tests common to econometrics. A rigorous treatment of the degrees of freedom is beyond the scope of this text. Informally, the concept of degrees of freedom measures the extent to which "surplus" information exists, beyond the minimum needed to estimate the model. The more information the more reliable the test and the more valid the estimate of the regression parameters a_0 .... a_k

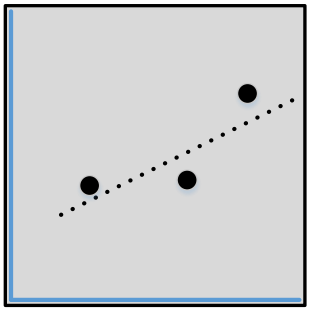
With one point, it is impossible to estimate a line – an infinite number exist





The two variable linear regression has two parameters (a_0 and a_1), and two points is the minimum information needed.

With three points, we have a "surplus" of 1 point and need a regression (method of least squares) to estimate the "best" line.
The more "surplus" information we have, the higher the degrees of freedom.



The t distribution has the following formula (you have permission to roll your eyes) $f(x) = \frac{\Gamma(\frac{v+1}{2})}{\sqrt{v\pi}\,\Gamma(v/2)}(1+\frac{t^2}{v})^{-\frac{v+1}{2}}$, where x is observation and v the degrees of freedom (n=k).

See t-Dist Generator.xlsx and Regression – t stat.xlsx

The null hypotheses associated with the regression model are for the intercept and slope(s). Think of the following model

$$Y_i = a_0 + a_1 X_{1i} + a_2 X_{2i}$$

The null hypotheses associated with the basic tests on the regression coefficients $a_0, a_1,$ and a_2 are

$H_0:\ a_0 = 0$	$H_0:\ a_1 = 0$	$H_0:\ a_2 = 0$
$H_1:\ a_0 \neq 0$	$H_1:\ a_1 \neq 0$	$H_1:\ a_2 \neq 0$

More complex null hypotheses are possible such as

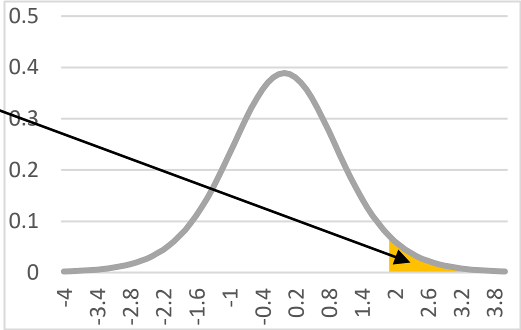
$H_0: a_0 \geq 5$	$H_0: a_1 = 3$	$H_0:\ a_2 + a_1 = 1$
$H_1: a_0 < 5$	H1: $a_1 \neq 3$	$H_1:\ a_2 + a_1 \neq 1$

Complex coefficient tests are common in advanced econometrics will not be part of this course

The ratio $a_1 / se(a_1)$ forms the **t statistic** (analogous to the Z score) and is really a **risk ratio …** the probability (risk) of being wrong if we reject H_0, when in fact it is true. The standard error of the intercept ($se(a_0)$) and slope ($se(a_1)$).

This risk ratio reflects the risk of a Type 1 error when we reject the null hypothesis. Critical values of the t statistics appear in the table. See [Sampling Distribution of Slope and Intercept.xlsx]

Critical Values of t				
Probability→ $df_{(n-k)}$↓	*0.1*	*0.05*	*0.025*	*0.01*
	0.9	*0.95*	*0.975*	*0.99*
5	1.476	2.015	2.571	3.365
10	1.372	1.812	2.228	2.764
25	1.316	1.708	2.060	2.485
50	1.299	1.676	2.009	2.403
100	1.290	1.660	1.984	2.364



Video: Sampling Distribution – Slope and Intercept

3.4. The contribution of an independent variable – t tests

Most commonly, we test the null that a coefficient is statistically significant from 0, and often with a one tailed test, since theory creates an expectation that economic relationships are positive (e.g., marginal propensities to consume) or negative (e.g., own price elasticity).

Table 1:Excel Regression Output					
Multiple R	0.974				
R Square	0.949				
Observations	68.000				
ANOVA					
	df	*SS*			
Regression	4	7532582800			
Residual	63	403136906.1			
Total	67	7935719706			
	Coefficients	*Standard Error*	*t Stat*	*P value*	
Intercept	$24,290.86	998.990	24.315	0.000	
Sex	-$1,926.49	709.185	-2.716	0.009	
Seniority (Years)	$104.24	50.077	2.082	0.041	
Field 1	$12,439.68	947.009	13.136	0.000	
Field 2	$22,560.53	1226.532	18.394	0.000	

=(Multiple R)^2

R^2 = Regression SS/Total SS

a_0
a_1
a_2
a_3
a_4

The R2, shows that this linear model with four independent variables and an intercept explains 94.9% of the variation in the dependent variable – Salary. The Multiple R is the correlation between the predicted and actual values. (Multiple R)^2 = R^2.

The ANOVA reports the SSE, SSR, and SST that form the R^2.

Since Salary is measured in $, it helps to translate the coefficients into dollars to support interpretation. The t statistics is Coefficient/Standard Error. The Standard Error is the standard deviation of the sampling distribution of the coefficient. The p value is the probability of a Type 1 error.

The most important aspect of regression, aside from setting up the model, developing the data, and defining variables (e.g., recoding Field into two variables Field 1 and Field 2) is interpreting the results. In this case the Intercept, Sex, Field 1, and Field 2 are statistically (P values smaller than .009), and Seniority is significant at .041. We take a small risk in rejecting the null the Field 1 (Life Sciences) and Field 2 (Engineering) do not affect salary. Recall Field 1 = 1 for Life Sciences and 0 otherwise and Field 2 = 1 for Engineering and 0 otherwise. When Field 1 = 0 and Field 2 = 0, the regression defines Social Sciences as the occupation which is the value of the intercept. Recall also that Sex = 0 for a male and 1 for a female.

See Regression – Salary on Sex, Seniority and Field.xlsx

Video: Regression on Salary on Sex, Seniority, and Field

Read this next section carefully and make sure you understand – something like this always appears on tests and the final exam,

Here is the interpretation of the regression. Let Field 1 = Field 2 =0. Then for someone in the Social Sciences, their starting salary (Seniority = 0 years) is $24,290.96 if they are male, and $22,364,47 if they are female ($24,290,86 – $1,926.46). Each addition year of work adds $104.24.

Now, for a male in the Life Sciences, their starting salary is $36,730.54 ($24,290.86 + $12,449.68) and for a male Engineer, the starting salary is $46,851.39. Females would earn $1,926.49 less for each occupation. Note that seniority has a small impact on salary, and it is occupation that is by far the most important influence. The role of Sex, while statistically significant, is much less important. Students focus on statistical significance, and while it is important, the numerical size of the coefficient is also relevant to interpreting a regression. A coefficient with a value of .000004 that is statistically significant at the .001 level is relevant for most forecasting and policy studies.

Balancing statistical significance and coefficient value can be tricky. Imagine that the coefficient on Field 2 (Engineering) was $22,560 as estimated here, but the p value was .15 and not .000. The null hypothesis is that an Engineering degree adds no extra income. You take a 15% chance of being wrong if you reject this null. Your gut tells you that Engineers get a salary boost, but the data do not support this. Most statisticians would advise to treat the value $$22,560 as if it were 0.

Notice that Seniority has a p value of .041 and, meaning that you take a 4% chance of being wrong if you reject the null (additional years of work – seniority – have no impact on salary). But the value of $104 for each addition year is trifling. Someone with a Social Science degree would need 216 years to catch up to an Engineer ($22,560/$104)! It is probably easier just to suck it up and get the Engineering degree.

3.5. Seasonal adjustment of data and forecasting

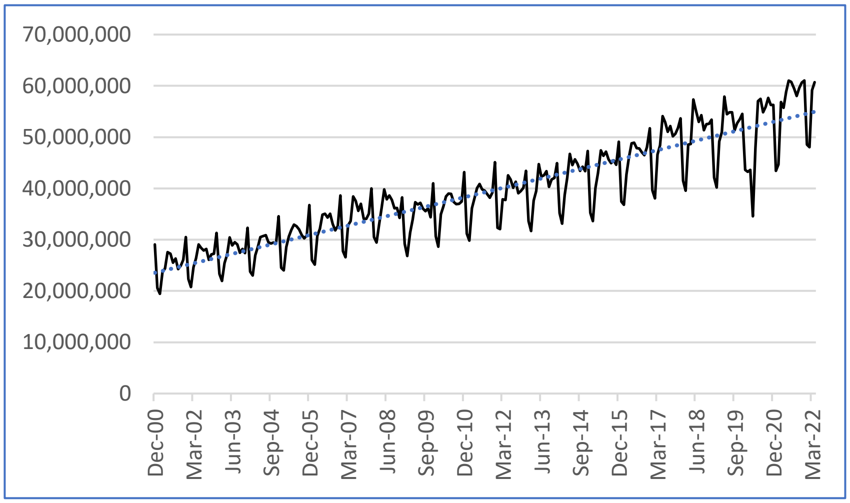
Time series data represent are a format for economic information. Certainly in 2020, marking both the progress of COVID and the collapse of the economy requires time series information. As noted in Module 2, time series data comprise one or more variables that have a time stamp. That time stamp may be minutes, hours, days weeks, months, quarters, or years.

As described in Module 2, time series data typically show one or more of four components – trends, cycles, seasonality, structural breaks, and random events. This section examines seasonality in time series and the creation of basic forecasting models.

3.5.1. Trend and cycle component

In

an overall *trend* exists. This is the basic relationship, and here the independent variables per capita incomes and population influence retail sales.



See

RETAIL.xlsx

Figure 5 Retail Sales Canada ($'000) Table: 20-10-0008-01:

The *cycle* is harder to see. One can create trends for subsets of retail sales. Figure 6 shows a series of trends for subsets of retail sales over the Dec 2000 – Apr 2022 period. The pre and post Great Recession (Dec 2008 – Jun 2009) are close, suggesting that retail sales did not experience much change through the business cycle, most likely because households must maintain housing, food, gasoline, and other routine spending

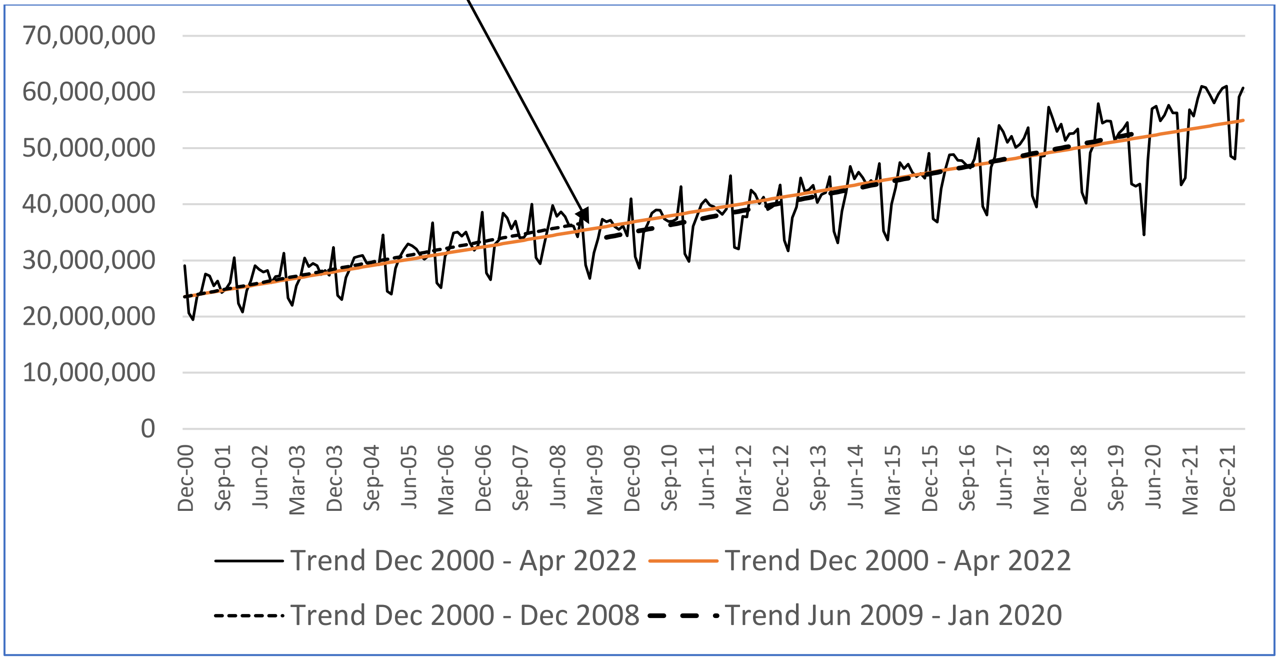


Figure 6: Retail trends

Study RETAIL Trends.xlsx to understand how to set up and graph these trends.

The various trend regressions are simple time trends

$$Y_t = a_o + a_1 t \, ,$$

where t indicates a date. This is peculiar regression; time is not a "variable" and does not "cause" changes in Y_t . Time is simply a tape measure against which to record changes in Yt, here retail sales. By looking at the twists and turns of Y_t, we can align events, such as the Great Recession, on the time tape, but in no sense did December 2007 *cause* the financial crises. Measuring rates of change assessing the severity of structural change is a main purpose of a time trend regression. By adjusting the sample, one can estimate trends for data subsets.

Structural breaks occur in 2008/2009 (Great Recession) and the precipitous drop in March/April 2020 due to COVID. But even after the sharp dip in the summer of 2020, retails rebounded to previous levels very quickly.

Video: Retail Sales Canada

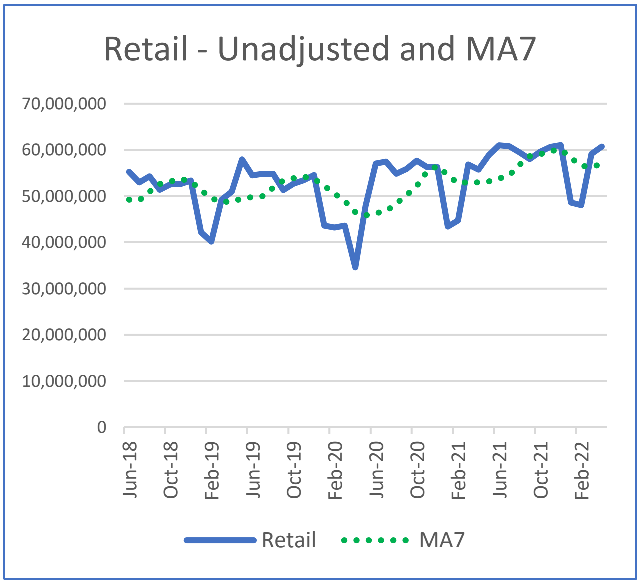
3.5.2. Seasonal adjustment

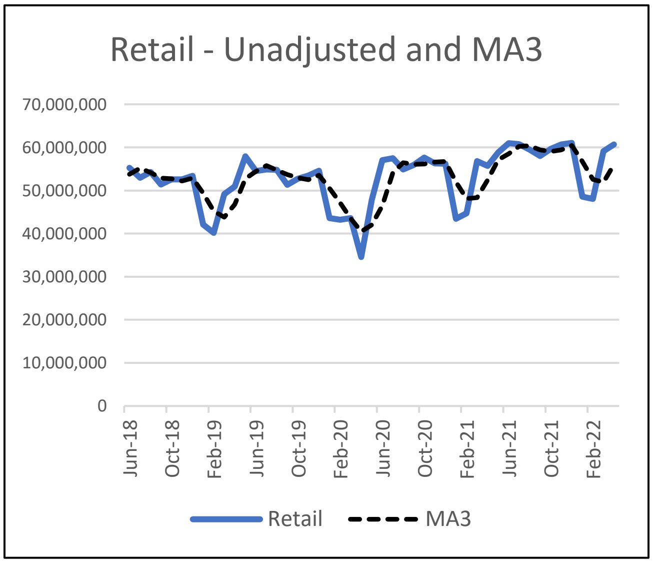
Retail trade reflects the equilibrium between demand and supply, typically showing all four time series components. The seasonal component in the figure is apparent. In Canada, Peak retail sales typically occur in December, with a minimum in February. Often, analysts compare peak-to-peak or trough- to-trough. The data in the figure are unadjusted for seasonality; that option exists for most time series maintained by Statistics Canada and other official statistics. Most often downloading the seasonally adjusted data makes sense, since central statistical agencies all use similar adjustment mechanisms that are quite sophisticated. However, since corporate data and information from other governments may not be seasonality adjusted (de-seasonalized), it is helpful to acquire some basic procedures for modifying the data.

Modelling the seasonal component of time series serves three purposes. First, seasonal adjustment clarifies the average movement of a variable data series. A time trend model is often too “aggressive” and can obscure important variation. Second, incorporating seasonal variation into a time trend model creates a framework for “story-telling”. Why do sales peak in December, sag in February, recover in July? These are sales for Canada. Do you imagine the same pattern would occur in a Southern hemisphere country” A Muslim country? Third, adding the seasonal component to a model will usually increase forecasting accuracy.

A common approach to seasonal adjustment is the moving average (MA). This requires taking successive averages of data spans, typically 3, 5 or 7. The conventional MA process places the average in the last of the set of 3, 5, 7…). This has the effect of shifting the resulting observations, so no gaps exist in the most recent series. The methods used in Col 3 and 4, has gaps in the first and last positions, which can be a disadvantage.

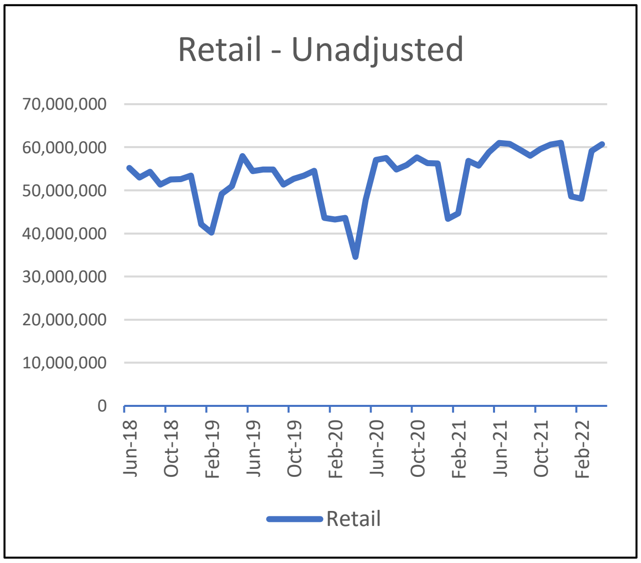
Choosing the period for an MA requires judgement. The trade-off is between smoothing the variability and seeing the underlying movement in a series, versus tracking all “twists and turns” in the data.





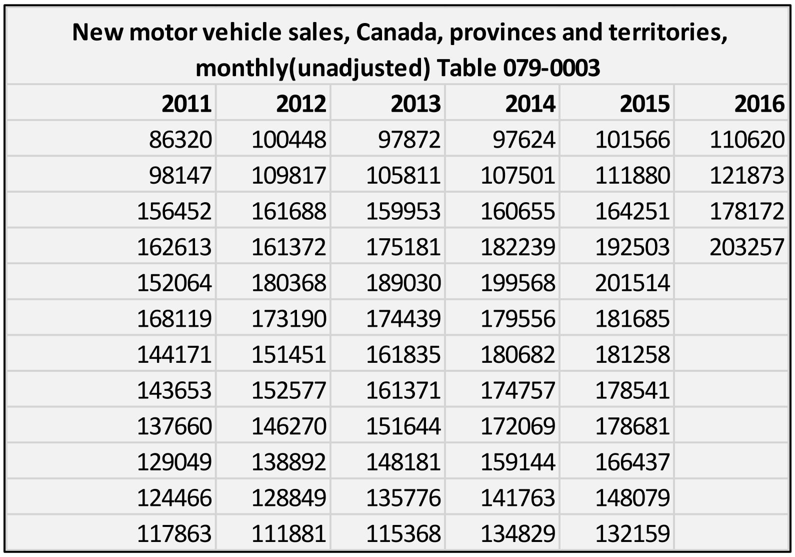
These three figures show the original Retail data (Jun 2018 – Apr 2022), and the effect of a moving average of 3 and 7 periods respectively. See [RETAIL MA.xlsx] making sure you work through the formulas. Note that these charts are at the bottom of the spreadsheet.

In general, use odd numbers for moving averages (think about why that must be). Note that the MA process deletes observations at the start.



3.6. Creating a seasonal index

A seasonal index tracks changes across weekly/monthly/quarterly cycles. Several techniques exist for creating seasonal indexes; some rely on using a moving average while others use regression-based techniques. The easiest approach requires isolating time slices, where the seasonality repeats. The figure below shows the data for] Motor Vehicle Sales.xlsx from the examples folder.



New motor vehicle sales, Canada, provinces and territories, monthly(unadjusted) Table 079-0003

2011	2012	2013	2014	2015	2016
86320	100448	97872	97624	101566	110620
98147	109817	105811	107501	111880	121873
156452	161688	159953	160655	164251	178172
162613	161372	175181	182239	192503	203257
152064	180368	189030	199568	201514	
168119	173190	174439	179556	181685	
144171	151451	161835	180682	181258	
143653	152577	161371	174757	178541	
137660	146270	151644	172069	178681	
129049	138892	148181	159144	166437	
124466	128849	135776	141763	148079	
117863	111881	115368	134829	132159	

The **Average Percent Method** has three steps

- Drop Incomplete years (2016)
- Sum each year and for each month create a new variable showing the percentage of the total year's sales
- Summarize across each month using mean or median and use the resulting column as the seasonal index. See Motor Vehicle Sales-Creating Median&Average Index.xlsx

4. Forecasting

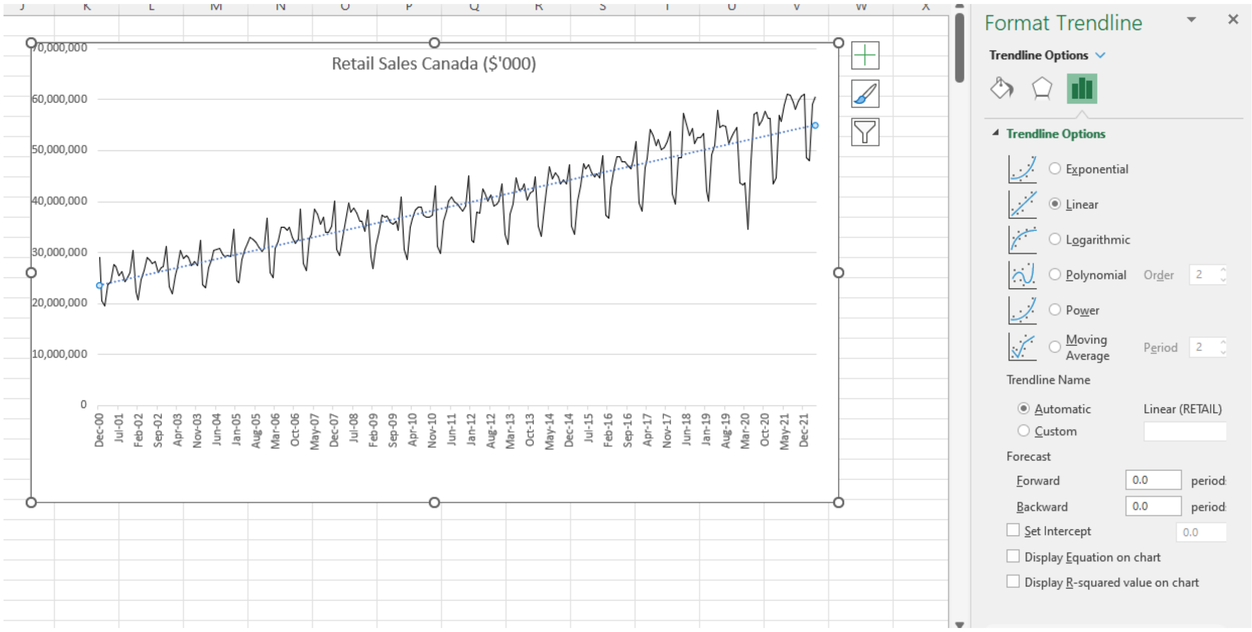
Forecasting comes in two varieties – projections based solely on the time series itself (call this *technical* forecasting) and projections based on a causal model that links the time series to causal variables (call this *"fundamentals"* forecasting).

4.1. Technical forecasting

Technical forecasting builds its projections using trend, cycles, and seasonality to extend a series beyond the present.

All forecasting is *conditional.* Technical forecasts are conditional on trends, cycles and the seasonal patterns remaining stable. Fundamental forecasts require both the projection on the independent variables and that the coefficients linking the independent variables and the dependent variable remain stable. This text focuses on technical forecasts.

The graphical options in Excel offer a simple (simplistic) approach to forecasting. It simply extends the trend using options in the menu.



Format Trendline

Trendline Options

○ Exponential

● Linear

○ Logarithmic

○ Polynomial Order 2

○ Power

○ Moving Average Period 2

Trendline Name

● Automatic Linear (RETAIL)

○ Custom

Forecast

Forward 0.0 periods

Backward 0.0 periods

□ Set Intercept 0.0

□ Display Equation on chart

□ Display R-squared value on chart

Video: Trends in Excel Charts

For prediction, the important question is how well a model based on historical (past) information can project into the future. Other Important questions include:

- Does the model capture the trend (linear or non-linear)?
- Does the model reflect seasonal and cyclical factors?

Imagine we wished to understand car and truck sales in Canada as shown in Figure 7.

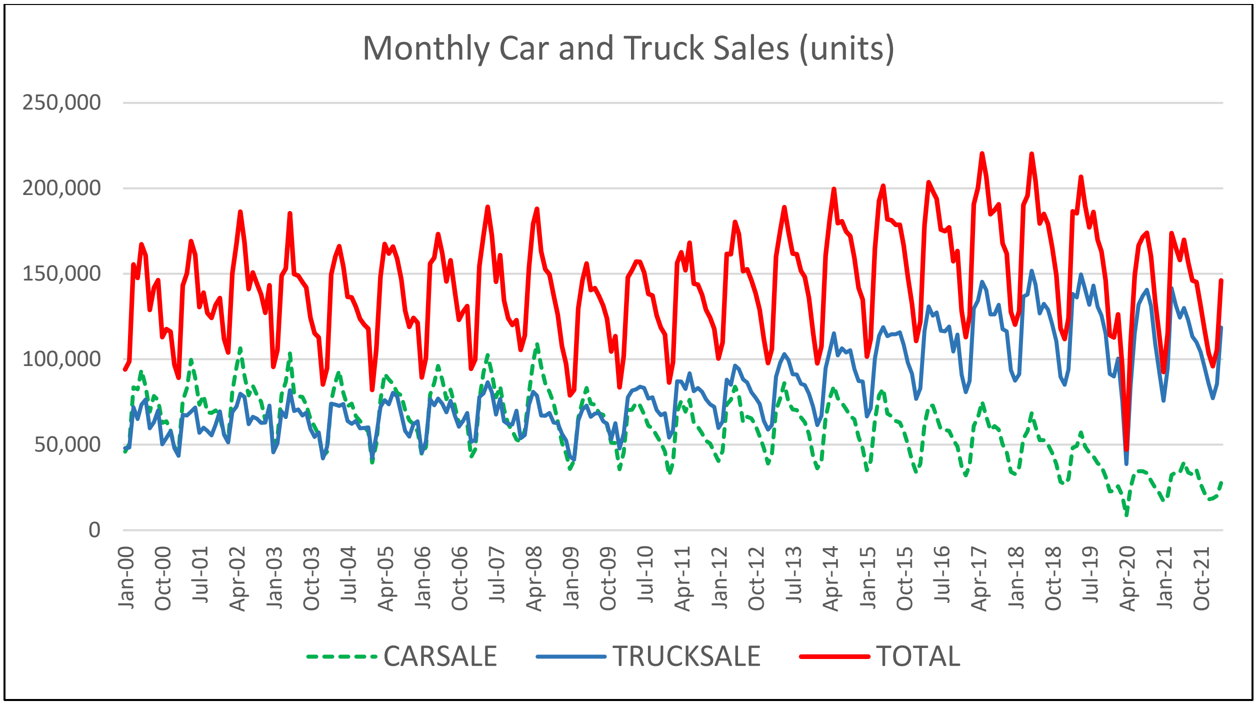


Figure 7: Vehicle sales

At first glance these time series show extreme seasonality. It also appears Canadians have been purchasing fewer cars over the period 2001 to the present, while the number of trucks (which include SUVs) has steadily increased. Also note the impact of the Great Recession (2008-2009) and COVID starting in January 2020 . (It is also always good to look at the raw data in the spreadsheet and not relay on the graph. See

Vehicle Sales.xlsx

[])

A forecasting model needs to capture the trend and the obvious seasonality in sales. Since these are monthly data, the model must specify 11 (k-1) dummy variables.

$$Y_t = a_0 + a_{1t}T_t + a_2D1_t + a_3D2_t \ldots + a_{12}D11_t \text{ ,}$$

where Yt is total vehicle unit sales in month t, T denotes the month, and $D1_t, D2_t \ldots D11_t$ are the 11 dummy variables for the 12 months. Setting up the dummy variables requires an =IF statement, but once you see it done, it becomes straightforward. See [Passenger Car Sales Forecast.xlsx] and [Passenger Car Sales with Seasonal Dummies.xlsx] for details.

Video: Passenger car sales with seasonal dummies

Time series analysis, where past values of a dependent variable and random variables form the basis for a forecast is an important form of technical forecasting. This is beyond the scope of this text; students going on in econometrics can expect to learn and use these models.

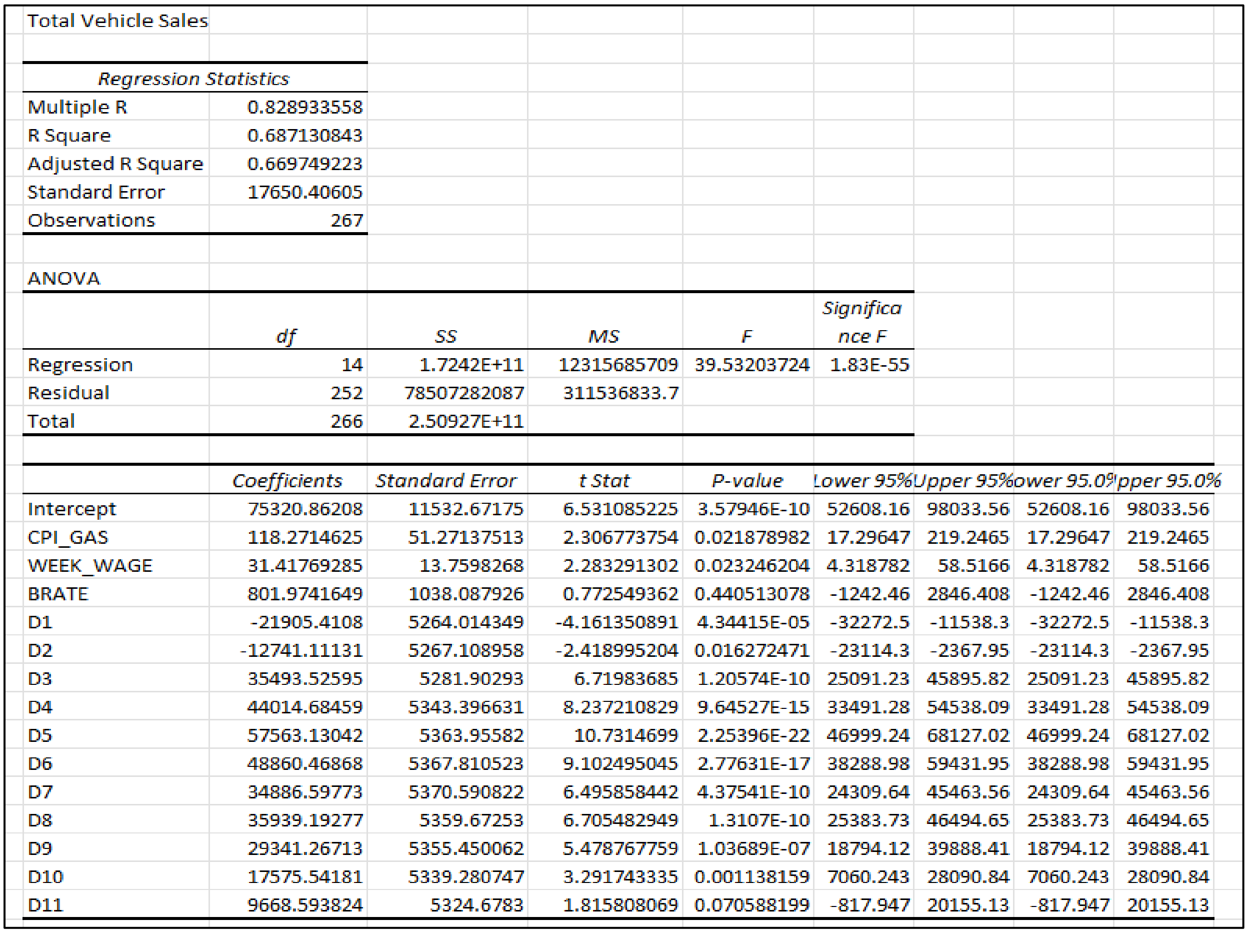
4.2. Fundamentals forecasting

Fundamental analysis uses economic theory to build a model of the time series, where a dependent variable, for example, motor vehicle sales could be a function of wages, vehicle prices, and the price of gasoline. A multiple regression model assesses the relationship between the independent variables (wages, vehicle prices and the price of gasoline) obtaining values for elasticities and other measures.

The dataset Vehicle Sales.xlsx includes several potential independent variables, such as the CPI for gas, weekly wages, and the bank rate.

Imagine that we wished to build a forecasting model that utilize these variables. The equation appears as this. Note that the time variable T no longer remains in the equation.

$$Y_t = a_0 + a_1 CPI\_GAS_t + a_2 WEEK\_WAGE_t + a3BRATE_t + a4D1_t + a_5 D2_t \ldots + a_{15} D11_t$$



Total Vehicle Sales

Regression Statistics	
Multiple R	0.828933558
R Square	0.687130843
Adjusted R Square	0.669749223
Standard Error	17650.40605
Observations	267

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	14	1.7242E+11	12315685709	39.53203724	1.83E-55
Residual	252	78507282087	311536833.7		
Total	266	2.50927E+11			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	75320.86208	11532.67175	6.531085225	3.57946E-10	52608.16	98033.56	52608.16	98033.56
CPI_GAS	118.2714625	51.27137513	2.306773754	0.021878982	17.29647	219.2465	17.29647	219.2465
WEEK_WAGE	31.41769285	13.7598268	2.283291302	0.023246204	4.318782	58.5166	4.318782	58.5166
BRATE	801.9741649	1038.087926	0.772549362	0.440513078	-1242.46	2846.408	-1242.46	2846.408
D1	-21905.4108	5264.014349	-4.161350891	4.34415E-05	-32272.5	-11538.3	-32272.5	-11538.3
D2	-12741.11131	5267.108958	-2.418995204	0.016272471	-23114.3	-2367.95	-23114.3	-2367.95
D3	35493.52595	5281.90293	6.71983685	1.20574E-10	25091.23	45895.82	25091.23	45895.82
D4	44014.68459	5343.396631	8.237210829	9.64527E-15	33491.28	54538.09	33491.28	54538.09
D5	57563.13042	5363.95582	10.7314699	2.25396E-22	46999.24	68127.02	46999.24	68127.02
D6	48860.46868	5367.810523	9.102495045	2.77631E-17	38288.98	59431.95	38288.98	59431.95
D7	34886.59773	5370.590822	6.495858442	4.37541E-10	24309.64	45463.56	24309.64	45463.56
D8	35939.19277	5359.67253	6.705482949	1.3107E-10	25383.73	46494.65	25383.73	46494.65
D9	29341.26713	5355.450062	5.478767759	1.03689E-07	18794.12	39888.41	18794.12	39888.41
D10	17575.54181	5339.280747	3.291743335	0.001138159	7060.243	28090.84	7060.243	28090.84
D11	9668.593824	5324.6783	1.815808069	0.070588199	-817.947	20155.13	-817.947	20155.13

Figure 8: Excel output for vehicle model

The Excel result appears in **Error! Reference source not found.**. Simplifying the output produces Table 2 below in Section 5.

Is this a valid regression model? Well, it depends! First, think about the measurement of the variables. Using the Consumer Price Index for gasoline is probably the best measure of a national price. Similarly, weekly wages are also a reasonable measure for household capacity to purchase vehicles. Third, the bank rate likely understates the true borrowing costs, but all interest rates tend to move in unison, so this seems a reasonable proxy (approximation).

The individual coefficients are all statistically significant at the 5% level or better (P<.05) except for the bank rate, which has a P=.44, implying that one takes a 44% chance of being wrong in rejecting the null hypothesis that the value of a_3 is 0. It is best to treat a3 as 0 and that the bank rate has no effect vehicle sales, *at least with these data*. Reread the italicized phrase, it is most important.

The dummy variables are all statistically significant with January and February (D1 and D2) being negative, and the intercept (standing for December) being large (75321). Study the Excel data and make sure you understand the definition of these dummy variables.

Now a problem does exist with the sign of the CPI-GAS coefficient. Intuitively one would expect that increasing gas prices would discourage vehicle purchases, yet here CPI-GAS has positive sign. Remember this is not a macroeconomic model; it is a forecasting model. The price of gas responds to market forces and macroeconomic trends. This is a weakness in this model.

Now to produce a forecast for the next 12 months, one needs to project the independent variables forward. For example, one might anticipate that the price of gas will rise 1% per month, that weekly wages will rise .5% per month and that the bank rate will remain steady. The example [] shows how to accomplish this. Make sure to work though all aspects including the graph.

Vehicle Sales Forecast.xlsx

5. Formatting Excel output

After specifying and testing a variety of models, you will want to bring present the "best" models in a report, typically in Word. You can edit the Excel regression output to present them to a current professional standard. The organization for which you are preparing your analysis may have "corporate" standards, and academic journals certainly has exacting specifications for their results.

Video: Formatting Excel output for Word

Table 2 shows the essential output from the case study in Section 4 The first step is to eliminate the "extraneous" information, with extraneous being defined by the contexts of the research. At a minimum, regression results should include the R2, sample size, coefficient, and one of t value or p value.

It is straight forward to edit the Excel output to create a subset of regression diagnostics that you can cut and past into Word.

Table 2 Total Vehicle Sales - Estimates

	Coefficients	Standard Error	t stat	P value
Intercept	75321	11533	6.53	0.000
CPI_GAS	118	51	2.31	0.022
WEEK_WAGE	31	14	2.28	0.023
BRATE	802	1038	0.77	0.441
D1	-21905	5264	-4.16	0.000
D2	-12741	5267	-2.42	0.016
D3	35494	5282	6.72	0.000
D4	44015	5343	8.24	0.000
D5	57563	5364	10.73	0.000
D6	48860	5368	9.10	0.000
D7	34887	5371	6.50	0.000
D8	35939	5360	6.71	0.000
D9	29341	5355	5.48	0.000
D10	17576	5339	3.29	0.001
R Square	0.687			
Observations	267			

By rearranging cells associated with regression output in Excel and deleting unnecessary output, it is possible to import a table into Word that communicates clearly. Right justify your entries and use only two or three significant digits (numbers after the decimal) for the t stat and P value. Depending on the scale of the original data, you may wish to have no digits after the decimal. Use your judgement.

6. Summary

By now, you should be quite adept at specifying a linear regression with multiple independent variables, creating dummy variables to mark regular variation in states, and interpreting the results. You also can now make a linear regression "non-linear" by using logs and exponents to transform columns. Finally, this Module has demonstrated reformatting Excel output for import directly to Word.

Annex A: Excel Functions and Formulas

The Data Analysis add-in (under the Data tab on the Ribbon bar) provides a menu driven way to estimate a regression function. Learning certain logical functions opens more opportunities for problem solving in Excel. Make sure you are familiar with the online help (F1) where many examples demonstrate the use of these logical formulas.

Excel Functions and Formulas		
Function	Example	Explanation
=IF(criteria, True, False	=IF(C5<C6,F2=25,F2=30) =IF(C5<C6,"Yes","NO")	If C5 is less than C6, then make F2 equal to 25, otherwise (C5≥C6) make F2= 30. Note that the "otherwise" is that C5 is greater than or equal to C6. It is important to be careful. In the second example, if C5<C6 is true, then place Yes in the cell where the =IF appears, otherwise place "NO".
=AVERAGEIF(range, criteria, average_ range)	=AVERAGEIF(A1:A10,">0")	In the cell containing the formula, place the average of all numbers less than 0 in the range A1:A10. If the average_ range not specified, then the range is used)
=SUMIF(range, criteria, sum_range)	=SUMIF(A1:A10,">10",C1:C10)	Sum the values in C1:C10 for all the elements >10 in A1:A10. Study the examines in F1 of Excel as this can get tricky
=COUNTIF (range, criteria)	=COUNTIF(A1:A10,"Male")	Counts the number of times "Male" appears in cells A1:A10
Note: The double quotes are important in the criteria element of logical formulas AVERAGEIF, SUMIF, AND COUNTIF		

Annex B: Interpreting log/ln transformation.

Logarithmic transformations have two primary purposes. First, they make calculations simpler – multiplication becomes addition and division becomes subtraction. Second, they transform variables from a non-linear form to a form that more closely resembles a straight line.

In economics logs often appear in investment, production, and consumption models. The best-known economic model using log transformations is the Cobb-Douglass production function with the formula appearing as:

$$Y_i = a_0 * X_{1i}^{a_1} X_{2i}^{a_2}\text{, where Y is output and } X_1\text{, } X_2 \text{ are inputs.}$$
$$\ln Y_i = a_0 + a_1 \ln X_{1i} + a_2 \ln X_{2i}$$

Log transformed regressions have three forms as shown in Table A1, using the level-level regression as a reference. Here y is the dependent variable and x the independent variable.

Table A1: Interpretation		
Model	Form	Interpretation
Standard (Level-level)	$y = a_0 + a_1 x + e$	$\Delta y = a_1 \Delta x$ Change x by 1 unit and y changes by a_1 units.
Log-Level	$\ln(y) = a_0 + a_1 x + e$	$\%\Delta y = 100 \bullet a_1 \Delta x$ Change x by 1 unit and y changes by 100▪a1Δ.
Level - Log	$y = a_0 + a_1 \ln(x) + e$	$\%\Delta y = (a_1 / 100\%)\Delta x$ Change x by 1%, and y increases by a_1/100 units.
Log-Log	$\ln(y) = a_0 + a_1 \ln(x) + e$	$\%\Delta y = a_1 \%\Delta x$ Change x by 1% and y changes by a_1 %. (This form offers a simple measure of the elasticity of y with respect to x *at the mean values of y and x)*