Comments on videos from the coursera course Econometrics Methods and Applications prepared by Erasmus School of Economics

Alois Geyer
alois.geyer@wu.ac.at
http://www.wu.ac.at/~geyer

June 8, 2020

Contents

1 Simple Regression **2**

2 Multiple Regression **7**

3 Model Specification **11**

4 Endogeneity **15**

5 Binary Choice **18**

6 Time Series **21**

1 Simple Regression

Lecture 1.1 on Simple Regression: Motivation

https://www.coursera.org/learn/erasmus-econometrics/lecture/oluRX/lecture-1-1-on-simple-regression-motivation

2:32 `... and the predicted sales are on the line`:
'predicted sales' seems to be related to the 'future', but that's not necessarily the case. Fitting a linear function based on known (observed, past) values of y and x (sales and price), results in so-called 'fitted values' of y (denoted by $\hat{y}$). I rather reserve the term 'prediction' for using values of x (and obtaining corresponding values of $\hat{y}$) which *have not been used* to determine the (parameters of the) fitted line (or 'fit', in general). Obtaining values for *future* y can be called prediction or 'forecast'. It requires predicting or assuming future values of x. Making assumptions about x and 'predicting' y is not a prediction but rather a 'scenario'. Using methods to first forecast x and subsequently forecasting y, is a special issue and requires special treatment (partly covered in week 6 and section 3 of my lecture notes).

In this example it is *assumed* that the price can be fixed/set by the company, in particular as a result of current/observed sales levels. As a matter of fact, this may introduce a serious problem called 'endogeneity'. If the company acts systematically in this way, running a simple regression as shown here does not produce meaningful results (in lecture 4.1 at 7:07 this is called 'strategic behavior'). The topic of 'endogeneity' is thoroughly treated in week 4 and section 1.9 of my lecture notes.

2:38 `... as there are other factors that also cause variation in sales`:
(a) If other factors are (actually) at work this requires running a multiple regressions, otherwise the results will suffer from an omitted variable bias (see lecture 3.2; section 1.6.7 in my lecture notes).

(b) It is hardly ever appropriate to use the word 'cause'. *Causality* is a difficult issue, and requires strong assumptions. I would rather use words like 'affect' or 'has an impact on'. The lectures in week 4 on 'endogeneity' address this topic in more detail.

2:59 `predicted sales`:
For the reasons explained above, I'd prefer the term 'fitted' sales.

3:49 `This helps to set a new price if sales are felt as too low`:
This refers again to the endogeneity problem mentioned above.

4:03 `estimate the optimal price to maximize turnover`:
It is understood that the purpose of this lecture is mainly to motivate the use of regressions. However, I have to point out that calculations like those aiming at determining an 'optimal' price from a regression equation can be very problematic. They rest on the assumption that now other effects are relevant (or, can be held constant), and that the two parameters a and b are given (perfectly known), whereas they are *estimates* associated with (standard) errors (see lecture 1.4, 6:23 below and section 1.2.2 in my lecture notes). In lecture 1.5 starting from 2:32 it shown how a confidence interval for an optimal sales level can be computed (accounting for various sources of uncertainty and estimation error). My caveat regarding the role/absence of other determining factors remains valid, however.

5:56 `observations of sales are considered to be independent draws`:
More precisely, observations are independent draws from a *population* with unknown mean μ and variance σ^2.

The assumption of 'random draws' is usually made, but depends on the way the sample is actually derived from the population. For time series data this assumptions is usually not justified, because consecutive observations are often correlated (for example, sales yesterday and sales today). A major exception are financial time series (more specifically, financial assets' returns or changes in exchange rates), for which consecutive observations are usually uncorrelated. However, the volatility of financial returns (as a time series) observed on consecutive days is highly correlated.

Note that the normality assumption made here is *not* a prerequisite for regression analysis. A more important assumption, made in lecture 1.4 at 3:22 is the normality of residuals. But even that has only limited impact on the results.

6:38 `the best prediction for the next observation on sales`:
The word 'best' is problematic. If, in fact, sales depend only(!) on price, then the best prediction must be based on price. This is known as the conditional mean of y. If more than one regressor is necessary, we require a multiple regression (see lecture 3.2) to determine the conditional mean (this is implicit in the statement at 7:03). Only if there exist no regressors (or, we don't find any) which are (linearly) related to sales, the 'best' predictor is the unconditional mean μ, estimated by the sample mean $\bar{y}$.

Lecture 1.2 on Simple Regression: Representation

https://www.coursera.org/learn/erasmus-econometrics/lecture/b560B/lecture-1-2-on-simple-regression-representation

2:08 `we move from an unconditional mean to a conditional mean given a value of x`:
This is a very important aspect, already mentioned at 6:38 in lecture 1.1.

2:31 `follows from demeaning y` [...] `such that a normally distribute error term with mean` μ `emerges`:
(a) The first statement refers to writing the equation as

$$\epsilon_i = y_i - \alpha - \beta x_i.$$

Note that the term 'de-meaning' usually refers to a transformation like $y_i - \bar{y}$ (for example, in the context of a panel regressions). This notion is used at 3:03 in lecture 1.3.

(b) First, it should be 'with mean 0' (rather than μ). Second, note that a normal distribution of ϵ does not necessarily follow/emerge from this transformation! Even if y is normal, substracting a non-normal x (e.g. bi-modal x) does not produce a normal ϵ. Conversely, even if y is non-normal, substracting a suitable non-normal x (e.g. if both y and x are skewed) may produce a normal ϵ.

3:18 `we did not include the prices of competing stores or the number of visitors`:
As mentioned above, if other factors are (actually) at work this requires to run a multiple regressions, otherwise the results will suffer from an omitted variable bias (see lecture 3.2; section 1.6.7 in my lecture notes).

5:43 `we often use the concept of elasticity`:
Further details on the term 'elasticity' can be found in section 1.6.1 in my lecture notes. This may also be useful for training exercise 1.2.

Lecture 1.3 on Simple Regression: Estimation

https://www.coursera.org/learn/erasmus-econometrics/lecture/HL2l0/lecture-1-3-on-simple-regression-estimation

0:00 `parameters alpha and beta that only exist in theory`:
'Theory' refers to parameters which are assumed to hold in the population. Econometrics builds on the notation that a regression equation $y = \alpha + \beta x + \epsilon$ holds in the population. α, β and ϵ are unknown, but can be estimated using a sample of y and x.

4:37 `b is equal to the sample covariance of Y and X divided by the sample variance of x`:
This can also be found at the bottom of p.2 in my lecture notes. There I also express b as a function of the correlation between y and x, and their standard deviations. This make clear that the units of measurement of b are units of y per unit of x.

5:38 `because their sample covariance is equal to 0`:
The sample covariance between x and e is 0 because of the partial derivative (first order condition) mentioned at 3:39 (`the sum of the observations on x times the residuals e is equal to 0`).

5:52 `R-squared is defined as the fraction of the variation in y`:
The 'variation in y' is not clearly defined. Does 'variation' refer to variance or standard deviation? In fact, R-squared is defined in terms of the *variance* of y and e. The equation SST=SSR+SSE only holds for sums of squares (or variances; hence also called variance decomposition). It *does not* hold for the standard deviations (or square roots of SST, SSR or SSE).

However, the term 'variation' refers to deviations from the mean which are measured in units of y! Scatter plots are shown in units of y (and x), and graphical inspection of the dispersion of actual values y around the fitted line $\hat{y}$ or $\bar{y}$ are shown in units of y. However, R-squared reflects the explanatory power in terms of squared units (for almost all cases, squared units of an entity are inconceivable).

Note that expressing the explanatory power (or goodness of fit) in terms of variance *always* paints a better picture than an equivalent measure based on original units. Unfortunately, such a measure does not exist. It is possible, however, to compare the standard deviation of y to the standard deviation of e (both being measured in the same units). In practice, such a comparison provides more useful information than R-squared, since we compare in terms of the units of the variable of interest y (its orders of magnitude are usually familiar to decision makers).

The `summary()` of an R object `lm()` reports the '`Residual standard error`'. In my lecture notes I refer to the *same* value/measure as the 'standard error of regression' (p.12). In the present example, this values is 1.189, whereas the standard deviation of the errors/residuals e is 1.183382 (using `sd(residuals())`). The (usually negligible) difference comes from using different scaling factors when computing the variance; $1/(n-1)$ for the (usual) variance and $1/(n-K)$ in a regression context (K=2 in a simple regression).

In the present case, the R-squared of the regression of sales on price is 0.72455. This suggests that a large fraction of the variance of sales is explained by the model (as stated at 5:57; I prefer to say 'explained by the variance in x'). In other words, the unexplained part (the remaining variance of e is rather small (about a quarter of y's variance). However, it is hard to relate to the variance of sales (5.084) since it is measured in squared dollars. The standard deviation of sales is 2.254777 (measured in dollars), the standard deviation

of the errors is 1.183382, and the standard error of the regression is 1.189 (both are also measured in dollars). Note that the unexplained part, if measured in terms of standard deviations, is rather large compared to the standard deviation of y. This comparison provides a more realistic assessment for the quality of a regression than R-squared.

In case of a simple regression (like here) R-squared is always equal to the square of the correlation between y and x.

6:24 `Since the mean value of the residuals is 0 ('Seen before')`:
This follows from the partial derivative w.r.t. a (see 2:15, or section 1.1.2 in my lecture notes).

Lecture 1.4 on Simple Regression: Evaluation

https://www.coursera.org/learn/erasmus-econometrics/lecture/B4S56/lecture-1-4-on-simple-regression-evaluation

1:51 `seven assumptions` [...] `to link the actual value of b`:
(a) These assumptions correspond to the assumptions AL, AR, AH and AN made in my lecture notes. However, I have decided to introduce assumptions step-by-step, and only when required. The assumption AX (refering to the 'exogeneity' of regressors) is missing in this list (actually, it has been 'replaced' because of A2). The lectures in week 4 on 'endogeneity' (i.e. violations of exogeneity) address this topic in more detail.

(b) Note that 'the actual value of b' is β. The statement 'how accurate b is for beta' is treated in section 1.2.1 of my lecture notes under the assumption that x is random (other than the (frequently made) assumption A2 that x is fixed).

4:55 `b is an unbiased estimator of beta`:
Note that this important result is essentially derived as a combination of assumptions and mathematical operations.

5:25 `two steps of the derivation`:
(a) A5 corresponds to AH in my lecture notes.

(b) Regarding the variance of b I prefer to use the (equivalent) formula (6) from my lecture notes:

$$\mathrm{V}[b_1] = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2}{(n-1)\sigma_x^2}.$$

It shows that the variance of b depends on the variance of ϵ (σ^2), the sample size n, and the variance of x (σ_x^2). This means that the precision of the estimate b increases (ceteris paribus) with the sample size (the more data, the better), the variance of x (the relation between x and y can be more precisely determined the larger the variance of x is; if there is hardly any variance in x, it is difficult to figure out how x affects y, and hence, the variance of b is large), and the variance of the errors (if the fit is bad and the variance of e is large, the estimate of b is not very reliable).

5:33 `Assumption A7 states that the epsilons are distributed as normal`:
(a) A7 corresponds to AN in my lecture notes.

(b) Note that A7/AN is only necessary when we want to test the estimated parameter b. No other step so far has required this assumption. However, if we find that residuals e

are not normal this may be an indication of missing or wrong regressors (inappropriate or incomplete model specification).

6:23 `the standard deviation of b`:
The standard deviation of b (s_b) is frequently called the '*standard error*' of b.

6:29 `the null hypothesis that beta is 0 :`
Note that other values for β can be used for tests (see the more general formula at the bottom of p.12 in my lecture notes).

7:10 `derive an approximate 95% prediction interval for a new value of y`:
Further details can be found in section 1.2.6 of my lecture notes.

7:32 `should be evaluated against the measurement scale of the variable x`:
This is a key insight which results from the fact that b is measured in terms of 'units of y' per 'units of x' (for example, if sales is measured in tons and price is measured in dollars, b is measured in tons per dollar; if sales is measured in terms of kg, b must increase by a factor of 1000). I invite you to 'play around' with the data of this lecture (e.g. rescale sales and/or convert price from dollars to other currencies, and see what happens to a (the intercept) and b, their standard errors and t-statistics, R-squared, etc.)

Training Exercise 1.4:
This is also covered in section 1.9.1 (second paragraph) in my lecture notes.

2 Multiple Regression

Lecture 2.1 on Multiple Regression: Motivation

https://www.coursera.org/learn/erasmus-econometrics/lecture/LQeLK/lecture-2-1-on-multiple-regression-motivation

2:44 `which corresponds to a level effect of 22%`:
The basis for this calculation can be found on p.34 (line 2) of my lecture notes. Note that the 'level effect' is expressed as a percentage. It cannot be expressed in monetary units as long as no base level is given.

Lecture 2.2 on Multiple Regression: Representation

https://www.coursera.org/learn/erasmus-econometrics/lecture/Zknc1/lecture-2-2-on-multiple-regression-representation

3:14 `error terms epsilon represent the imperfections of the model`:
The term 'imperfections of the model' may be misleading. ε includes consequences from misspecifications (e.g. mainly missing/omitted but relevant regressors, or wrong functional form). However, it also includes unavoidable but unsystematic noise (assumptions AX and AH reflect that). In other words, it is a 'natural' part of any model, representing the unsystematic, unpredictable part of a phenomenon, which we can and need to ignore, since our main goal is to capture the systematic part $X\beta$ properly. Any modeling decision about $X\beta$ directly affects ε. The main challenge for any empirical research is to properly separate those two parts. This separation and the nature of ε will be a core element of the course, and will be repeatedly discussed.

3:48–4:51 `test question`:
This is about the assumption AR (full rank). This assumption and the presentation until 4:51 is based on mathematical requirements. Note that the main challenge for empirical research is the related issue of multicollinearity (section 1.6 in my lecture notes) (i.e. cases which *almost* violate AR).

5:46 `keeping all other factors fixed is not possible in practice`:
The reference to a 'thought experiment' is important. Note that these 'thought experiments' are at the heart of theoretical models which generate empirically testable implications. The parameters of interest in theoretical models are usually constructed/derived under a ceteris paribus condition by isolating the mechanism of interest. Other circumstances are left out by assumption (i.e. by defining the model environment). In 'real life' many other effects are present and reflected in actual data. The main task of an econometrician is to account for all those aspects/factors (which are not part of theoretical models). As stated in this video they have to be included in the empirical model, in order to isolate the partial effects (coefficients) of the variables of interest.

6:35 `total and partial effects`:
Note that partial effects are those that are the objects of interest from the perspective of theoretical (economic/financial) models. The total effect of a variable can be estimated by using that variable as the only regressor in a simple regression. However, simple regressions are hardly ever used and are usually considered as misspecified because of omitted regressors. If at all, total effects are rather derived from a complete model specification by making use of the (more precisely) estimated partial effects/coefficients.

7:41 `Statistical tests can be formulated for the significance`:

This is covered in section 1.2.3 of my lecture notes.

7:56 `test question`:
Note that this question refers to the *true* parameter β_j. In empirical work β_j is unknown and has to be estimated from the data. In that case the answer is more complicated since (a) we have to account the estimation error associated with (the estimate) b_j, and (b) the results of hypothesis tests are subject to errors of type I and II.

Lecture 2.3 on Multiple Regression: Estimation

https://www.coursera.org/learn/erasmus-econometrics/lecture/jMr3i/lecture-2-3-on-multiple-regression-estimation

1:04 `test question`:
As in the test question at 3:48 in lecture 2.2 this test question refers to a mathematical requirement. As noted above the actual/main challenge of empirical analyses to find a good balance between the number of parameters (i.e. regressors) relative to the number of observations. Ideally, we like to find a complete specification (i.e. without any missing/omitted regressors) which does not include any irrelevant regressors (which decrease the efficiency of a model). Details are treated in section 1.6.7 of my lecture notes.

2:13-3:23 `matrix methods ... test question`:
I want to point out that I put little emphasis on these mathematical aspects. I don not consider solving equations (using matrix algebra) to obtain LS estimates (see section 1.1.1 of my lecture notes) as an important issue in empirical econometric analyses.

3:51 `this (famous) formula can be computed from the observed data X and y`:
Note that this (famous) formula consists of two parts: $\boldsymbol{X}'\boldsymbol{y}$ and the inverse of $\boldsymbol{X}'\boldsymbol{X}$. As explained on p.2 of my lecture notes, the first term reflects the covariance of $\boldsymbol{y}$ and $\boldsymbol{X}$, and $\boldsymbol{X}'\boldsymbol{X}$ the covariance among $\boldsymbol{X}$.

3:57-5:55 `to have also a geometric picture`:
I want to point out that I do not consider this geometric perspective as very important, with the following exception:

6:34 `the fact that e is orthogonal to the X plane`:
The normal equation (2) (see p.3 in my lecture notes) implies that each column of $\boldsymbol{X}$ is uncorrelated with (orthogonal to) $\boldsymbol{e}$. This can be viewed as follows: the estimates $\boldsymbol{b}$ have been chosen in a way which exploits all available information in $\boldsymbol{X}$. Hence, the residuals $\boldsymbol{e}$ are orthogonal to $\boldsymbol{X}$. If that was not the case, it would be possible to find a better estimate $\boldsymbol{b}$ which extracted more (all relevant) information from $\boldsymbol{X}$.

7:20 `R-squared`:
Whatever I have said about R-squared in my comments on lecture 1.3 at 5:52 holds here as well.

Lecture 2.4.1 on Multiple Regression: Evaluation - Statistical Properties

https://www.coursera.org/learn/erasmus-econometrics/lecture/tYw5A/lecture-2-4-1-on-multiple-regression-evaluation-statistical-properties

0:13 `assumptions`:
Whatever I have said about the assumptions in my comments on lecture 1.4 holds here as

well.

2:22 `test question`:
This is covered in section 1.2.2 of my lecture notes (Expected value of b).

Note that my derivation depends critically on assumption AX ($\mathrm{E}[\varepsilon|\boldsymbol{X}] = 0$), whereas the corresponding assumption in the video is A3 ($\mathrm{E}[\varepsilon] = 0$).

2:57 `variance-covariance matrix of b`:
This is covered in section 1.2.2 of my lecture notes (Covariance of b).

3:54-5:49 `estimate` σ^2 `[...] this estimator is unbiased [...] trace trick`:
This part can be skipped.

Lecture 2.4.2 on Multiple Regression: Evaluation - Statistical Tests

https://www.coursera.org/learn/erasmus-econometrics/lecture/vg4ra/lecture-2-4-2-on-multiple-regression-evaluation-statistical-tests

1:14-5:09 `test whether the j-th explanatory factor has an effect`:
This can be found in section 1.2.3 of my lecture notes. Note that the null hypothesis need not be $\beta_j = 0$; any other (justified) assumption about β_j can be made (for example, see section 1.2.5 on interest rate parity).

Lecture 2.5 on Multiple Regression: Application

https://www.coursera.org/learn/erasmus-econometrics/lecture/pbXxv/lecture-2-5-on-multiple-regression-application

0.34 `The resulting regression equation is shown on the slide`:
The slide also shows the standard error of regression (residual standard error in R) s=0.245. An interpretation of this result is not provided in the video but can be found in section 1.6.1 of my lecture notes. 0.245 is approximately the standard deviation of the percentage error $(y - \hat{y})/y$ or $(y - \hat{y})/\hat{y}$. In other words, the errors expressed as a percentage of observed or fitted wages have a standard deviation of about 24.5%.

1:20 `the gender dummy has a p-value of 0.097 and is not significant`:
Note that this conclusion is based on a two-sided test. For a one-sided test (based on the alternative hypothesis that women earn less than men on average) the p-value would be 0.097/2 (less than 5%). Based on a one-sided test, the coefficient –0.041 would be considered as significantly negative, and we would conclude that women earn significantly less than men (on average and ceteris paribus).

2:59 `in lecture 2.1 we found a total gender effect of minus 25%`:
Note that this total effect can be used for descriptive purposes. In this sample, the average wage of women (97.3) is, in fact, about 25% below the average wage of men (125.1). However, it is not correct to argue that the average wage is lower *because* of a gender effect. In a multiple regression we are getting *closer* to such an interpretation. Holding other effects (age, education, etc.) constant, we associate a difference of about 4% with a gender effect. Theoretically, if all relevant circumstances that determine wages have been observed and added to the regression, one could argue that the coefficient of 'female' is an estimate for a 'causal' (gender) effect. The lectures in week 4 on 'endogeneity' address this topic in more detail.

3:52 `regression using wage level`:
The slide also shows the standard error of regression (residual standard error in R) s=31.276. An interpretation of this result is not provided in the video. It is measured in the same (monetary) units as the wage level. A comparison with s=0.245 from the regression using the log of wages is possible but requires to take the log of fitted values and compute errors relative to the log of wages. The standard deviation of these errors is about 0.29, which implies that the fit of the model in levels is worse than the fit of the model using logs.

4:17 `the two models do not have the same dependent variable`:
This is an important, generally valid aspect. R-squared, the sum of squared residuals, the standard error of the regression etc. of models with different dependent variable, cannot be directly compared. A comparison is possible after the fitted values of one model have been converted to the same units of measurement of the dependent variable in the second (or other) model (as shown in my previous comment).

5:47 `model for log-wage contains the variable education`:
This aspect is covered in section 1.6.2 in my lecture notes.

8:43 `the largest wage effect is found`:
Note that each of the three coefficients indicates the percentage difference with respect to the reference category (i.e. education level 1) The effects of moving from one level to the next can be assessed on the basis of differences between coefficients (e.g. 0.38–0.17, 0.765–0.38).

3 Model Specification

Lecture 3.1 on Model Specification: Motivation

https://www.coursera.org/learn/erasmus-econometrics/lecture/a27La/lecture-3-1-on-model-specification-motivation

3:16 and 3:37 `on the same scale`:
It is a bit misleading to refer to 'the same scale'. First, comparable scales can always be obtained by shifting the mean and/or multiplying the variable with a constant (i.e. standardize using $z = (x - \bar{x})/s_x$ and subsequently scaling to any mean m_u and standard deviation s_u using $u = m_u + zs_u$). Note that standardizing/rescaling in this way does not make a variable normally distributed! This is a rather common misunderstanding!

In a cross-sectional regression putting variables on the same scale is not relevant at all (the coefficients account for any such changes; see 7:32 in my comments on lecture 1.4 in week 1). In a time series context, however, it is more important that the series have comparable 'stochastic' properties. In particular the should be stationary (this and related aspects will be covered in the lectures of week 6, and in sections 2 and 3 of my lecture notes). For the time being, it is enough to make sure that time-series regressions should not include a mixture of variables which have trends (like SP500) and others which have no trend (like book-to-market).

3:37 `the change in the log of the index from one period to the next`:
This transformation is the same as computing a log return (see section 2.1 of my lecture notes).

Lecture 3.2 on Model Specification: Specification

https://www.coursera.org/learn/erasmus-econometrics/lecture/gRZEw/lecture-3-2-on-model-specification-specification

0:37 `trade off`:
This trade off is covered in section 1.6.7 of my lecture notes.

2:36 `variance of the restricted estimator`:
My lecture notes do not contain this result. However, I have a clear perspective on the answer to the bias-efficiency trade off. It is stated in the last sentence of section 1.6.7: 'What is the point of estimating a parameter more precisely if it is potentially biased?'

5:59 `out of sample criteria`:
Note that RMSE and MAE are also used for in-sample comparisons.

A very powerful version of out-of-sample comparisons is to carry out cross-validations. In this case the entire sample is split repeatedly and randomly into two subsamples; one is used for estimation and the the other for out-of-sample prediction. RMSE, MAE or other measures are recorded in each repetition, and subsequently the resulting set of measures is analyzed to draw conclusions about the out-of-sample properties. Note that cross-validation techniques in a time-series context are more difficult than in a cross-sectional context.

6:22 `how to decide which variables to include`:
This sequence of steps, the main aspects, and my personal preferences are covered in section 1.6.8 of my lecture notes. Note that a frequently successful approach (not mentioned in the video or my lecture notes) is to combine forecasts or predictions from several models.

The idea is to retain a set of plausible, rather well-performing candidate models, compute predictions from each, and obtain the (weighted) average of all predictions. This works well if the individual models are not too similar.

Lecture 3.3 on Model Specification: Transformation

https://www.coursera.org/learn/erasmus-econometrics/lecture/TVbqK/lecture-3-3-on-model-specification-transformation

0:33, 0:44, 0.59 `level / growth rate / similar in nature`:
These considerations are mainly relevant in a time series context (see my comments at 3:16 in lecture 3.1)

1:53 `may affect the stability properties`:
These properties have been called 'stationary' above, and will be covered in the lectures of week 6, and in sections 2 and 3 of my lecture notes.

3:10 `nonlinear effects`:
Interactions are covered in section 1.6.3 of my lecture notes. Note that a quadratic term (i.e. squaring a variable) is the same as an interaction of the variable with itself. Furthermore, a ratio of two variables is an interaction of a variable with the reciprocal of another variable (without the need to explicitly compute that reciprocal as an additional variable).

As state at 4:04 none of these transformations actually makes the model nonlinear in the parameters β (despite such specifications allow for a rich set of nonlinear effects).

6:02 `dummy variables`:
Dummy variables are covered in section 1.6.2 of my lecture notes.

Note that an important version of working with dummy variables are diff-in-diff models (see section 1.6.4 of my lecture notes).

Lecture 3.4 on Model Specification: Evaluation

https://www.coursera.org/learn/erasmus-econometrics/lecture/Phh8y/lecture-3-4-on-model-specification-evaluation

General remark on evaluating a model specification: This is a rather comprehensive and difficult subject. The Chow break test covered in this lecture is very well known, but certainly not enough to evaluate a model. Important aspects not mentioned in lecture 3.4 include checking the plausibility of the estimated coefficients (both in terms of sign and magnitude. Looking only at the 'significance' is not enough. A regressor which seems important from a theoretical/practical point of view should be kept in a model even though its coefficient is not significant at usual levels (e.g. I would take a p-value of 0.12 as sufficient evidence that the associated regressor does play a role, in particular if sign and magnitude are ok, and the regressor has sufficient ex-ante plausibility).

A constant concern of any modeling activity should be omitted regressors and/or endogeneity (covered in week 4). There exists no test which indicates that a regressor has been omitted (in particular, *which* regressor) and/or a regression is subject to an endogeneity problem, *unless* one has realized that such a problem may exist. A test for endogeneity requires so-called instruments. A test for an omitted regressor requires plausible ex-ante reasons for its relevance, and data on that regressor. After adding it to the model its coefficient can be tested (i.e. check its sign, magnitude and significance).

8:35 `normality is rejected`:
As stated in section 1.7.1 of my lecture notes 'a failure to obtain normal residuals in a regression may indicate missing regressors and/or other specification problems (although the specific kind of problem cannot be easily inferred)'.

Lecture 3.5 on Model Specification: Application

https://www.coursera.org/learn/erasmus-econometrics/lecture/Rhveo/lecture-3-5-on-model-specification-application

1:57 `is approximately equal to a growth rate`:
The chart correctly denotes this 'growth rate' as a log return.

3:10 `we regress the log equity premium on one of the variables`:
For the reasons stated in section 1.6.8 of my lecture notes I do not recommend this specific-to-general approach (as described starting at 7:16 of lecture 3.2). Each of the simple regressions potentially suffers from an omitted variable problem (in particular since it intended to subsequently use all regressors).

3:39 `book-to-market and the dividend/price ratio are significant`:
Note that such a conclusion does not account for the correlation among regressors. The whole point of multiple regressions is to make judgments about a single variable's significance *in the context of other relevant regressors.*

3:54 `apply the general-to-specific approach`:
This is the approach described (and preferred for the reasons stated) in section 1.6.8 of my lecture notes.

4:43 `we need to ensure that the disturbance term has mean zero`:
This is misleading because the disturbance is (the unobservable) ϵ. Its mean cannot be affected by the model specification. Adding a constant term only makes sure that the mean of the residuals e is zero!

5:04 `the dividend/price ratio has limited explanatory power when controlling for book-to-market and earnings/price effects`:
This example provides a key insight. The explanatory power (and significance) of a regressors should always be judged *in the context of other relevant regressors*.

5:29 `final model is a simple regression model`:
Note that this 'final result' of the specification search may be problematic. More specifically, after eliminating earnings/price from the model, the coefficient of book-to-market changes from –0.29 to –0.185. The latter coefficient can be viewed as suffering from an omitted variable bias (because of the lacking control from earnings/price). The p-value of 0.159 for earning/price (generally speaking: the p-values of plausible regressors which are not 'too large') indicates that earning/price may have a non-negligible effect *in fact* ()i.e. *in the data generating process*). The sample may be too small or measurement errors may be too strong to obtain a sufficiently precise (hence significant) estimate for its coefficient. Hence, omitting earning/price from the regression may unnecessarily reduce the explanatory power of the model and introduce a bias in the coefficients of other regressors (book-to-market in this case). As stated in the last sentence of section 1.6.7 of my lecture notes 'What is the point of estimating a parameter more precisely if it is potentially biased?'.

4 Endogeneity

Lecture 4.1 on Endogeneity: Motivation

https://www.coursera.org/learn/erasmus-econometrics/lecture/fXJ3X/lecture-4-1-on-endogeneity-motivation

1:58, 2:40, 3:24 `explanatory variables are exogenous and stochastic`:
Note that this assumption has not been mentioned in previous videos. In my lecture notes I have assumed from the very beginning (see assumptions in section 1.2.1) that $\boldsymbol{X}$ is stochastic, mainly for the reasons stated on the slide shown at 3:24.

3:44 `even with variables that are not included in the model`:
This possibility leads to the assumption/requirement AX: $\mathrm{E}[\boldsymbol{\epsilon}|\boldsymbol{X}] = \mathbf{0}$.

It is important to note that allowing $\boldsymbol{X}$ to be stochastic *does not lead* to endogeneity! This is only the case if AX is violated, which implies that $\boldsymbol{\beta}$ estimated by OLS is inconsistent.

Note that the video only requires $\boldsymbol{X}$ and $\boldsymbol{\epsilon}$ to be *uncorrelated*. My version of AX refers to *independence* between $\boldsymbol{X}$ and $\boldsymbol{\epsilon}$, which is a stronger assumption (as shown in the implications associated with AX on p.8). This stronger version allows to derive inconsistency of OLS in special cases.

4:56 `endogeneity is often due to an omitted variable`:
This aspect is covered in section 1.6.7 of my lecture notes.

Note that another important source of endogeneity is *simultaneity* which is not covered in this video. It is treated, however, in my lecture notes on p.59.

Lecture 4.2 on Endogeneity: Consequences

https://www.coursera.org/learn/erasmus-econometrics/lecture/NgOl4/lecture-4-2-on-endogeneity-consequences

0:43 `measurement error`:
This example is very instructive and informative about the consequences (i.e. inconsistency) of endogeneity (as generated by measurement error). Note that inconsistency also arises as a consequence of omitted variables or simultaneity. You may want to consider doing experiments (simulations) on your own, and investigate the consequences of omitted variables in the same way as here for the case of measurement errors.

6:20 `consistency of OLS`:
This aspect is covered in section 1.3.1 of my lecture notes. Note that the results shown here are derived under the assumption of exogeneity (i.e. endogeneity is not violated).

7:07 `Q inverse must exist`:
This condition is stated in my lecture notes after equation (11) by stating/requiring that $\boldsymbol{Q}$ is a positive definite matrix.

Lecture 4.3 on Endogeneity: Estimation

https://www.coursera.org/learn/erasmus-econometrics/lecture/Muofe/lecture-4-3-on-endogeneity-estimation

0.50 `unexplained factors`:
The term 'unexplained factor' may be misleading. It actually refers to components of the disturbances $\boldsymbol{\epsilon}$ which are unobserved/unobservable, and which *do affect* $\boldsymbol{y}$ (via $\boldsymbol{\epsilon}$). They

are 'unexplained' in the sense that we have no information to do so. To the extent that $\boldsymbol{X}$ is correlated with $\boldsymbol{\epsilon}$ this would be possible, but at the same time this makes our estimates of coefficients relating $\boldsymbol{y}$ and $\boldsymbol{X}$ biased and inconsistent.

1:01 `instruments`:
These aspects are covered in secction 1.9.2 of my lecture notes. It is important to also highlight the '*exclusion condition*' (not mentioned in the video.) It makes clear that the search for suitable instruments has to account for the fact that an instrument *must not* appear in the original regression. In other words, it must not have been part of our initial choice of potential regressors required to properly explain $\boldsymbol{y}$ (from a theoretical, practical, or intuitive viewpoint).

2:33 `This solves endogeneity as the unexplained part of X is by construction uncorrelated with the explained part / so X explained is now exogenous`:
Note that the 'unexplained part' of $\boldsymbol{X}$ are the residuals from the first-stage regression (denoted by v in my lecture notes). The 'explained part' is denoted by $\hat{x}$ (the fitted values from the first-stage regression). v and $\hat{x}$ are uncorrelated 'by construction' because of the normal equation (2) (implied by the LS principle; see p.3). $\hat{x}$ (the explained part) is exogeneous if/since it is only determined by the exogenous explanatory regressors and the exogeneous instrument.

2:46 `matrix of instruments`:
Note that the matrix of instruments $\boldsymbol{Z}$ consists of those (original) regressors which are (or can be safely) assumed to be exogeneous and the instrument(s) (see lecture 4.4 at 1:39).

4:30 and 5:11:
These steps can be skipped.

5:26 `how can we obtain instruments`:
Note that the consistency of 2SLS results from *assuming* to have instruments with the required (ideal!) properties. However, the main challenge in empirical work is to *identify* and *obtain data* on suitable instruments.

8:29 `if X is endogenous only 2SLS will be consistent`:
Note that the consistency of 2SLS is only obtained if suitable instruments are used. If instruments are weak (insufficient explanatory power in the first-stage regression) or invalid (not exogenous and not uncorrelated with $\boldsymbol{\epsilon}$), 2SLS can result in even worse estimates than OLS.

Lecture 4.4 on Endogeneity: Testing

https://www.coursera.org/learn/erasmus-econometrics/lecture/MRfe4/lecture-4-4-on-endogeneity-testing

2:58 `Sargan test`:
The Sargan test is not explicitly mentioned as such in my lecture notes. However, the required steps of this test are described in the third paragraph on p.64 (the second paragraph provides an intuition).

Note that this test requires more instruments than endogeneous variables ($m>0$ in my lecture notes and $m>k$ in the video at 4:34). In empirical research it is usually very hard to find suitable instruments in the first place. Finding even more instruments than required is still harder. Finally, note that the presence of more than one endogenous regressors is associated with even stronger requirements than just increasing the number

of instruments (as stated in the last sentence of the first paragraph on p.64).

5:57 `exogenous by construction`:
Note that regressors are hardly ever exogenous 'by construction'. I'd prefer to state that one part of the regressors can be (safely) assumed to be exogenous.

5 Binary Choice

Lecture 5.1 on Binary Choice: Motivation

https://www.coursera.org/learn/erasmus-econometrics/lecture/SzZIo/lecture-5-1-on-binary-choice-motivation

4:04 `estimate for beta_2 turns out to be -0.86 divided by 1,000`:
Note that dividing by 1000 is not necessary, and affects the coefficient in a deterministic way. If price is not divided by 1000 the estimated coefficient is given by –0.861/1000.

5:16 `the slope of the regression line is the same for every value of price`:
Note that this is a *general* feature of linear models (as already pointed out in the context of choosing a log-linear model; see section 1.6.1 of my lecture notes).

Lecture 5.2 on Binary Choice: Representation

https://www.coursera.org/learn/erasmus-econometrics/lecture/JqZTD/lecture-5-2-on-binary-choice-representation

4:16 `the shape of the logit function stays the same`:
Note that the chart *may give the impression* that the shape of the function does not 'stay the same'. In fact, you can check the statement by comparing the value of the black logistic function at any x with that of the purple function at $x-2$; you will find the two values to be identical. Accordingly, the derivatives of the two functions at x and $x-2$ are the same. This is a general feature which does not only hold for a logistic function, but for any function (see https://en.wikipedia.org/wiki/Horizontal_translation).

6:22 `size of the effect`:
The 'size of the effect' refers to the change in the probability when x changes.

Lecture 5.3 on Binary Choice: Estimation

https://www.coursera.org/learn/erasmus-econometrics/lecture/BJ98y/lecture-5-3-on-binary-choice-estimation

1.06 `The corresponding estimation technique is called maximum likelihood`:
This method is described in section 1.4 of my lecture notes. I recommend reading section 1.4 and/or the next two paragraphs below before watching the rest of this video.

The key to understanding the maximum likelihood principle rests on the following aspect. In the standard case, the parameters of a distribution are known, and we can compute probabilities. In maximum likelihood estimation, the situation is reversed. We know the relative frequencies ('probabilities') of observations, and we compute (estimate) the parameters. This can be described a bit more accurately. For a given probability density/distribution function with *known* parameters, the probability of observing specific values of the associated random variable can be computed.[1]

Suppose we have a sample of observations, but we do not know the parameters of the associated distribution (which has to be assumed to be known). Maximum likelihood estimation rests on *knowing* the probabilities (actually the relative frequencies) of observations, and looks for the *unknown* parameters. Suppose we need to choose from two pairs of parameters of a normal distribution: (μ_1, σ_1) and (μ_2, σ_2). One of these two pairs will fit better to the histogram (i.e. the relative frequencies) of the data. Maximum likelihood suggests choosing that pair which fits better/best because it is more/most likely that the sample comes from a population having those parameters.

1:39 `likelihood contribution of observation i`:
In case of a continuous dependent variable the 'contribution to the likelihood' is not specified in terms of a probability for observing $y_i = 0$ or $y_i = 1$, but in terms of the probability of observing (the real-valued) y_i. Note that MLE requires assuming an (appropriate) probability density/distribution function for y_i. In section 1.4 I show that MLE in a regression context requires assuming a density for the residuals ϵ.

6:10 `the maximum likelihood estimator does not exist`:
Another (intuitive) way of explaining this result is based on the fact that the distribution of y in such a case ('all values are the same') is degenerate. Since y has no variance it is impossible to find out which value(s) of the unknown β make it more or less likely to observe y.

6:22 `provided of course that the model is correctly specified`:
Refering to 'correct specification' is essential here. For example, the consistency of MLE does not hold in case of endogeneity, or if there are omitted regressors.

8:18 `Likelihood Ratio Test`:
The likelihood-ratio test is covered in section 1.5 of my lecture notes. Note that this test is a very general test which does not only apply in case of logistic regressions.

9:19 `you reject the restrictions`:
'Rejecting the restriction' implies that the additional regressors of the unrestricted model supply additional explanatory power such that their parameters cannot be considered to be jointly insignificant (i.e. the additional regressors are important/relevant). Hence, the unrestricted model is preferred over the restricted model.

[1] In case of a discrete random variable we can compute the probability for specific values, or for a range of values (below, above, between) for a continuous variable.

Lecture 5.4 on Binary Choice: Evaluation

https://www.coursera.org/learn/erasmus-econometrics/lecture/glnyt/lecture-5-4-on-binary-choice-evaluation

2:19 `obtains the value very close to [0][1]`:
Note that Richard Paap says 'very close to *zero*' whereas the transcript below the video states 'very close to *one*'. Since the likelihood is (in this particular case) a product of values close to one, it equals a value close to one. However, the log-likelihood will be close to 0.

Lecture 5.5 on Binary Choice: Application

https://www.coursera.org/learn/erasmus-econometrics/lecture/jJGTY/lecture-5-5-on-binary-choice-application

3:32 `The exact logit specification is given on the slide`:
Note that dividing age by 10 in the squared term is not necessary; this only makes the associated coefficient β_4 larger.

4:21 `The McFadden R-squared is low as is usual for logit models. The Nagelkerke R-squared is much higher.`:
Comparing these two R-squared values is misleading! Stating that Nagelkerke R-squared is 'much higher' is irrelevant. This is somewhat clarified in the following statements.

7:26 `marginal effect of age`:
Note that reproducing this graph is a bit tricky. You need to account for the fact that age appears twice in the function. In addition, in the present example, age has been divided by 10 before squaring.

7:51 `a change in age`:
Arguing in terms of 'changing age' is usually problematic. Statements refering to 'changing age' make more sense if we are comparing two customers (with otherwise identical features) whose age differs (e.g. by one).

6 Time Series

Lecture 6.1 on Time Series: Motivation

https://www.coursera.org/learn/erasmus-econometrics/lecture/UjOZl/lecture-6-1-on-time-series-motivation

1:57 `spurious conclusions`:
The issue of *spurious regression* is not thoroughly covered in my lecture notes. It is only mentioned in sections 1.8.3 and 3.2.4. It is remarkable that this online course treats this important aspect very early.

2:30 `whereas the variables X and Y are completely uncorrelated`:
This statement may be misleading or confusing because we can 'see' (or measure) that the two series have a non-zero (positive) correlation. It should be noted at the outset that this problem requires x and y to be *highly autocorrelated*. Autocorrelation is treated in lecture 6.2 (at 6:14). For the time being it is sufficient to know that the coefficients of lagged x and y have to be close to one (around or above 0.9 as in this example).

The statement rests *on knowing* that the disturbances of x and y are uncorrelated by construction. Hence, x and y may be correlated (from a purely observational perspective) but that must not be taken as evidence that the two series have anything in common on a 'deeper' level. In other words, the 'spurious regression problem' highlights the need to 'look closer' in order to avoid drawing inappropriate conclusions about the 'actual' relation between two series (i.e. to avoid that 'pictures can sometimes fool us' as stated at 4:45).

3:20 `include the Y variable one period lagged`:
Note that the motivation for adding lagged values of y (and x at 3:50) is explained in section 1.8.2 of my lecture notes.

6:37 `use both time series to estimate the common trend`:
Note that proceeding in this way requires that the two series do *actually* have something 'in common' (ideally, justified from a subject-matter/economic point of view), and we are not 'fooled' by spurious correlations.

Lecture 6.2 on Time Series: Representation

https://www.coursera.org/learn/erasmus-econometrics/lecture/LIRdq/lecture-6-2-on-time-series-representation

1:24 `autocovariances`:
Autocovariances and autocorrelations are covered in sections 2.1.4 and 2.2 of my lecture notes.

1:40 `the past carries no predictive value for the future`:
It is very important to note that here 'the past' only refers to the past of that same time series. Zero autocorrelation of a series (then called white-noise) *does not* preclude the possibility that this series is affected by past values (lags) of *other* time series!

1:58 `a white noise time series cannot be predicted from its own past`:
It must be pointed out again that the key term is '*own* past'. The slide adds 'by linear models' which refers to the possibility that *nonlinear* models may be able to capture important features of a white-noise series (e.g. the variance of the series as covered in section 2.5 of my lecture notes). The reference to 'linear models' must not be mistaken to imply linear models with other series as regressors (see previous comment).

2:06 `model is deemed successful if [...] the residuals are white noise`:
This is a key aspect of time series analysis, summarizing the main requirement of any modeling attempt.

7:45 `slowly decaying pattern`:
The patterns in (P)ACF are not used here in order to judge which of the candidate time series models (AR, MA, ARMA) is appropriate. This 'identification' is treated in section 2.2 of my lecture notes (summarized in Table 1). However, while associating different versions of ARMA models with (P)ACF patterns has a clear theoretical basis, it may be rather hard to find clear associations empirically (i.e. based on observed series and estimated (P)ACF).

8:10 `Several trend models are available`:
These models are covered in section 2.3 of my lecture notes.

Lecture 6.3 on Time Series: Specification and Estimation

https://www.coursera.org/learn/erasmus-econometrics/lecture/mHGkR/lecture-6-3-on-time-series-specification-and-estimation

0:29 `use the autocorrelation and partial autocorrelation functions to specify a first-guess model`:
Such a first guess can be based on the theoretical patterns described in Table 1 of my lecture notes (section 2.2.3, p.106). Note that sample variation can lead to estimated patterns which only roughly correspond to those theoretical patterns. Therefore, it may be difficult to find a clear association between sample (P)ACF and theoretical (P)ACF. It is quite common to entertain several 'first guess' models choose on the basis of in-sample fit and/or (simulated) out-of-sample evidence.

2:25 `test question`:
We need to distinguish the disturbances ϵ_t of an AR(p) model which are white-noise *by assumption*, and the residuals e_t from an estimated AR(p) model. The estimated residuals of an AR(p) model are uncorrelated with the 'regressors' y_{t-1}, y_{t-2}, ..., y_{t-p} as an implication of the LS principle (i.e. the first order condition leading to the normal equation). One can view the LS estimation in this case as an attempt to minimize the (in-sample) squared forecast errors. In other words, the estimates of the AR coefficients f_1, f_2,...,f_p are chosen such that

$$\sum_{t=p+1}^{n} e_t^2 \longrightarrow \min \qquad e_t = y_t - (f_1 y_{t-1} + f_2 y_{t-2} + \cdots + f_p y_{t-p}).$$

These residuals – provided that the (type of ARMA) model and the number of lags are correctly chosen – are orthogonal to y_{t-1}, y_{t-2}, ..., y_{t-p} (i.e. $\text{corr}[e_t, y_r] = 0$ for $r < t$). Since

$$e_s = y_s - (f_1 y_{s-1} + f_2 y_{s-2} + \cdots + f_p y_{s-p}) \qquad r \leq s < t$$

and e_t is uncorrelated with e_s (see video at 3:00).

However, if the model is not correctly specified and lags are missing, this leads not only to an omitted variable bias but also to autocorrelated residuals. For example, if the data

generating process (DGP) is indeed AR(2) but an AR(1) model is fitted, the residuals *are not* white-noise! This can be derived as follows:

$$\text{DGP AR(2)}: y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t \qquad \epsilon_t \sim \text{white-noise},$$

$$\text{fitted AR(1)}: y_t = \varphi_1 y_{t-1} + \eta_t \qquad \varphi_1 = \phi_1 + \phi_2 \frac{\gamma_1}{\gamma_0}.$$

The autocovariance of η_t (from the misspecified AR(1) model) is given by

$$\mathrm{E}[\eta_t \eta_{t-1}] = \mathrm{E}[(y_t - \varphi_1 y_{t-1})(y_{t-1} - \varphi_1 y_{t-2})] = (1 + \varphi_1^2)\gamma_1 - \varphi_1\gamma_2 - \varphi_1\gamma_0,$$

where γ_k are the autocovariances of y_t (note that $\gamma_k{=}\gamma_{-k}$).

3:28 `we have to resort to the method of maximum likelihood in case of ARMA models`:
While it is correct that 'moving average models include also lagged forecast errors that are still unknown' this does not require to use maximum likelihood. It is also true that LS estimation cannot be applied in terms of the usual formula $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, since in case of MA models $\boldsymbol{X}$ consists of lagged residuals which are not known and cannot be used to define $\boldsymbol{X}$. However, ARMA models can be estimated using a recursive version of LS which requires numerical optimization to minimize the sum of squared errors.

5:53 `proper statistical analysis`:
Some more specific theoretical foundations can be found in sections 1.3 and 1.3.3 of my lecture notes.

7:33 `The answer is shown on the slide`:
An alternative way to derive this result can be found in footnote 98 on p.121 in my lecture notes.

8:53 `first you perform an ADF test`:
I do not recommend starting with an ADF test as a general rule (see the discussion at the bottom of p.122 in my lecture notes). I recommend to do ADF tests only if you are in doubt about the stationarity of a time series. Doubts about stationarity can be raised if a graphical inspection of a series shows slow mean reversion and/or rather long-term (stochastic) trends, or autocorrelations decay slowly and rather linearly (instead of exponentially). Admittedly, these guidelines require some experience. I recommend to simulate time series which are close to being non-stationary, inspect time series plots and autocorrelations, and apply ADF tests.

Lecture 6.4 on Time Series: Evaluation and Illustration

https://www.coursera.org/learn/erasmus-econometrics/lecture/eKJgO/lecture-6-4-on-time-series-evaluation-and-illustration

0:40 `Any application of time series modeling should start with examining whether the time series is stationary`:
See my comments at 8:53 from lecture 6.3.

Lecture 6.5 on Time Series: Application

https://www.coursera.org/learn/erasmus-econometrics/lecture/MThTF/lecture-6-5-on-time-series-application

6:20 `we see that the residuals are very similar`:
This 'test' is an informal way of checking again the results from the previous F-test (are three lags enough or do we need to add lags until twelve?). To check whether the model with three lags is appropriate one should test whether autocorrelations of the residuals are (individually and/or jointly) significant. If this test shows significant residual autocorrelation the model is not correctly specified despite the similarity of the residuals of the two models.

9:41 `we simply add up the monthly growth rates`:
This result is shown in terms of multi-period returns in section 2.1 and used in terms of long-horizon returns in example 19 (section 1.8.3) of my lecture notes.