REGRESSION ANALYSIS: MECHANICS

These notes provide a review of ordinary least-squares (OLS) regression principles. OLS regression is used in every field imaginable, from anthropology to zoology. In finance, the most commonly used application is the estimation of "betas" or market risk factors. While understanding the regression model and its attendant assumptions is critical for an informed financial decision-maker, understanding the economics underlying the regression model specification and the idiosyncrasies of financial data are equally, if not more, important.

Objectives

After reviewing this note, you should be able to:

1. Estimate a simple OLS regression model.
    1.1. State and understand the four OLS regression assumptions.
    1.2. Estimate and interpret OLS regression results.
    1.3. Perform hypothesis tests and construct confidence intervals for individual regression coefficients.
    1.4. Assess goodness-of-fit using the R-squared.
    1.5. Estimate OLS regression through the origin.
2. Estimate a multiple OLS regression model.
    2.1. Restate OLS regression assumptions.
    2.2. Estimate and interpret OLS regression and ANOVA results.
    2.3. Perform hypothesis tests and construct confidence intervals for individual regression coefficients.
    2.4. Perform hypothesis tests for an entire OLS regression model.
    2.5. Calculate and interpret the adjusted R-squared and choose from a collection of models based on explanatory power and parsimony.
3. Recognize violations of model assumptions and understand the consequences.
    3.1. Specification errors
    3.2. Omitting relevant explanatory variables
    3.3. Including irrelevant explanatory variables
    3.4. Nonlinearities
    3.5. Multicollinearity
    3.6. Violations of the disturbance assumption
    3.7. Heteroscedasticity
    3.8. Serial correlation
    3.9. Non-normality

1. *Simple linear regression*

1.1 Assumptions

The goal of regression is to learn about the relationship between variables. Pay attention to the adjectives in front of the term "regression." They supply information about the structure of the model. In this first section, we focus on simple linear regression. The word *simple* refers to the fact that we have only *two* variables, *X* and *Y*, and the term *linear* refers to the fact that a line will represent the relation between the variables. In contrast to simple linear regression, *multiple* regression involves more than two variables, and *nonlinear* regression consists of a relationship between *X* and *Y* that is not a straight line.

In simple linear regression, *X* is the *independent variable* and appears on the right-hand side of the relation. Other names for it include *explanatory variable* and *predictor variable*. On the left-hand side is *Y*, the *dependent variable* or *response variable*. *X* is non-random, taking on values that the investigator sets. *Y* depends linearly on *X* but also has a random component ε. Thus, the relation between *X* and *Y* in a simple linear regression,

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad (1)$$

where β_0 represents the intercept of the regression line, β_1 represents the slope, and i means the i^{th} pair of observations of the variables *X* and *Y*. We have assumed that the relation between *X* and *Y* is linear and that the values of *X* are non-random or fixed.

The three remaining regression assumptions pertain to the error term, ε_i First, the expected value of ε_i is 0 and the variance of ε_i is constant across observations, that is, $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2$. Note that, if *X* is non-random, the error term will have constant variance if and only if the response variable *Y* has constant variance. The constant variance assumption is known as *homoscedasticity* and is the basis for *ordinary* least squares regression estimation. The term "ordinary" applies because every observation of Y_i has equal variance and weight in estimating the model. In contrast, a weighted least squares approach is appropriate if the response variable Y_i and the error term ε_i have non-constant variance (i.e., are *heteroscedastic*). This allows observations with smaller variances to have more weight than those with larger variances.

The second assumption governing the residual error term is that the errors are independent of one another, that is, $Cov(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$. Violation of this assumption induces *autocorrelation* or *serial correlation,* a problem often encountered in time-series data. Finally, the residual errors are normally distributed. Because the *X*'s are non-random, this assumption implies that the response variable *Y* is also normally distributed.

A single statement can summarize the four OLS regression assumptions:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2) . \tag{2}$$

The relation between X and Y is linear. The expression, $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$, means that the errors are independent and identically distributed (*iid*) with $N(0, \sigma^2)$, which signifies the distribution is normal with mean 0 and variance σ^2.

Caveat: The term "linear" refers to the fact that the regression equation is linear in the parameters, and not necessarily in the variables. Consider, for example, a nonlinear model such as

$$Y_i = e^{\beta_0 + \beta_1 X_i + \varepsilon_i} .$$

On face appearance, linear regression seems inappropriate. Such a model, however, is *inherently linear* in the sense that it may be re-written as the linear model,

$$\ln Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Nonlinear models that can be re-written into a linear form using only a transformation of the X or Y variables are linear.

OLS regression assumptions

1) The relation between X and Y is linear.
2) The error term ε is independent, identically (normally) distributed with mean 0 and constant variance σ^2. (3 assumptions)

1.2 Estimation

We can estimate the model based on our data satisfying the four assumptions of OLS regression. First, we denote the estimated regression line by

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

where $\hat{\beta}_0$ represents our best guess for the true intercept β_0, $\hat{\beta}_1$ is our best guess for the population slope β_1, and $\hat{Y}$ is the predicted value of Y that falls along the regression line. To calculate this line, we need to choose values of $\hat{\beta}_0$ and $\hat{\beta}_1$. This is done here using a method known as *ordinary least square* (OLS) estimation. As noted earlier, the term "ordinary" arises because every observation is assumed to have equal variance and is therefore given equal weight in estimating the model. The term "least squares" is used because we will choose the line that minimizes the

squared distances between the observed and the predicted response variables. Defining the sample residual as $e_i = Y_i - \hat{Y}_i$, the OLS objective function is:

$$\underset{\{\hat{\beta}_0, \hat{\beta}_1\}}{Min} \sum_{i=1}^{n} e_i^2$$

Among other things, minimizing the sum of squares errors implies that the sum of the regression errors (and the average error, for that matter) will equal zero. This means that the regression can be re-expressed in deviations from the mean form. That is, if the mean in the regression model (2) is 0, the mean of *Y* is:

$$\bar{Y} = \beta_0 + \beta_1 \bar{X}\,, \tag{3}$$

where $\bar{X}$ is the mean of *X*. Taking the difference between the expressions,

$$\begin{aligned} Y_i - \bar{Y} &= \beta_0 + \beta_1 X_i + e_i - \beta_0 - \beta_1 \bar{X} \\ &= \beta_1 (X_i - \bar{X}) + e_i \end{aligned}$$

Expressing the deviations from the mean as $y_i = Y_i - \bar{Y}$ and $x_i = X_i - \bar{X}$, the regression equation becomes

$$y_i = \beta_1 x_i + e_i\,. \tag{4}$$

It also means that the least squares *estimators* of β_0 and β_1 are identified. Write the sum of squared errors, $\sum_{i=1}^{n} e_i^2$, as

$$\begin{aligned} \sum_{i=1}^{n} e_i^2 &= \sum_{i=1}^{n} (y_i - \beta_1 x_i)^2 \\ &= \sum_{i=1}^{n} (y_i^2 + \beta_1^2 x_i^2 - 2\beta_1 x_i y_i) \end{aligned} \tag{5}$$

Differentiating (5) with respect to β_1,

$$\frac{d \sum_{i=1}^{n} e_i^2}{d\beta_1} = \sum_{i=1}^{n} (2\beta_1 x_i^2 - 2x_i y_i)\,. \tag{6}$$

Setting (6) equal to 0, simplifying, and re-arranging provides the least squares estimator of the slope coefficient, that is,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}\,. \tag{7}$$

Because the mean residual error is zero, the estimator for the intercept follows from (3), that is,

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} \,. \tag{8}$$

1.3 Hypothesis tests for individual regression parameters

After a model has been fitted, we may assess whether the independent variable has a significant effect on the response variable. To do so, we test the hypothesis that the parameter β_1 equals zero. If the slope of the regression line is zero, this implies that there is no linear relation between *X* and *Y*; in this case, the "true" regression line is $E(Y \mid X) = \beta_0$, and we are just as well off using the sample mean, $\overline{Y}$, to predict future *Y*'s. A population slope not equal to zero implies that the variables are somehow linearly correlated. A regression line with a slope greater than zero means that *X* has a positive effect on *Y*, and one with a downward slope implies a negative relation between the two variables. We can test for any of these relations using the data collected in our sample.

Suppose we are interested in learning about the relation between two variables. The variables are believed to be linearly related, but it is not known whether the relation is negative or positive. A two-tailed hypothesis test can be used to check whether the population slope is equal to or unequal to zero. (If we had an idea that the slope was either positive or negative, then we would use a one-tailed test). But before any testing is done, we must preset a desired level of the test, denoted by α. In our case, α represents the probability of incorrectly concluding that the two variables are related in some linear manner ($\beta_1 \neq 0$), when in fact they are not ($\beta_1 = 0$). A researcher formalizes these possibilities by specifying two different hypotheses: a null hypothesis, denoted H_0, and an alternative hypothesis, denoted H_1. The alternative hypothesis is usually what the researcher is trying to prove, for example, that the variables are related, and the null hypothesis usually refers to the status quo. So, in our example, we would choose the null hypothesis to be $H_0 : \beta_1 = 0$ and the alternative as $H_1 : \beta_1 \neq 0$. Once we have stated our null and alternative hypotheses, specified the level of the test, and collected the data, we can formally test our beliefs.

Based on our sample, the best guess for the "true" slope of the regression line is $\hat{\beta}_1$. There is, however, error associated with this estimate. This inaccuracy can be quantified by the standard error of the estimate $s_{\hat{\beta}_1}$, which is defined as

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{\sum_{i=1}^{n} x_i^2}} \tag{9}$$

where

$$s = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}} \tag{10}$$

is an estimate for σ the standard deviation of the error. Using this measure of uncertainty, we can standardize our parameter estimate to get the test statistic $t = \hat{\beta}_1 / s_{\hat{\beta}_1}$, which follows a *t*-distribution with $n-2$ degrees of freedom. If the absolute value of $\hat{\beta}_1$ is much larger than its standard error, then *t* will also grow large in absolute value, indicating that β_1 it is different than zero. A large positive *t*-ratio is evidence of a positive relation, and a large one a negative relation. Because the test statistic has a *t*-distribution, we can set cutoffs, or critical values, for rejecting the null hypothesis for any specified level of significance. The table below gives the rejection rules for three types of hypothesis tests of the regression line slope. Given a probability α that a *t*-distributed random variable (with $n-2$ degrees of freedom) is greater than some *critical value* t_α, we have the following rules.

Common hypothesis tests for the slope of a regression line

Null hypothesis	*Alternative hypothesis*	*Rejection rule*
$H_0 : \beta_1 = 0$	$H_1 : \beta_1 \neq 0$	Reject H_0 if $\lvert t \rvert > t_{\alpha/2}$
$H_0 : \beta_1 = 0$	$H_1 : \beta_1 > 0$	Reject H_0 if $t > t_\alpha$
$H_0 : \beta_1 = 0$	$H_1 : \beta_1 < 0$	Reject H_0 if $t < -t_\alpha$

Note that the first row of the table is a "two-tailed" hypothesis test. Since the alternative hypothesis $\beta_1 \neq 0$ does not specify whether β_1 it is greater than or less than 0, we compare the absolute value of the *t*-statistic with the critical *t*-value corresponding to a probability $\alpha/2$, that is, $\alpha/2$ in each tail of the two tails of the distribution.

We can use a similar procedure to perform hypothesis tests on the intercept. We specify the null and alternative hypotheses involving β_0, select the level of significance, and calculate $\hat{\beta}_0$ according to the formula given above and its standard error,

$$s_{\hat{\beta}_0} = s\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} x_i^2}}\,. \tag{11}$$

Then $t = (\hat{\beta}_0 - c) / s_{\hat{\beta}_0}$ is the test statistic for the null hypothesis $H_0 : \beta_0 = c$ that the intercept is equal to some value c. (This is more general than the decision rules for the slope. For the slope, an investigator usually wants to evaluate whether it is equal to something other than zero. To avoid confusion, the tests of β_1 assume that c is always equal to zero.) It should also be noted that hypothesis tests for the slope are far more common in a simple regression setting than for the intercept.

Common hypothesis tests for the intercept of a regression line

Null hypothesis	*Alternative hypothesis*	*Rejection rule*
$H_0 : \beta_0 = c$	$H_1 : \beta_0 \neq c$	Reject H_0 if $\lvert t \rvert > t_{\alpha/2}$
$H_0 : \beta_0 = c$	$H_1 : \beta_0 > c$	Reject H_0 if $t > t_{\alpha}$
$H_0 : \beta_0 = c$	$H_1 : \beta_0 < c$	Reject H_0 if $t < -t_{\alpha}$

Confidence intervals

The idea of constructing confidence intervals is related to that of hypothesis testing. Again, we are interested in finding whether the regression's independent variable significantly affects the dependent variable. But this time, instead of using a test statistic and critical value, we construct intervals and try to "pin down" the true value of a parameter and base our inferences on that. As shown above, the tools used in constructing confidence intervals are identical to those used in hypothesis testing.

To calculate a confidence interval for the slope of the regression line, we need only three things: the point estimate for the parameter, the standard error of the estimate, and the confidence coefficient that is taken from a *t*-distribution with $n-2$ degrees of freedom. The interval itself is just the estimate of the parameter plus or minus a margin related to the estimate's standard error and the selected confidence level.

Confidence intervals for the intercept and slope of a regression line

Parameter	*Interval size*	*Confidence interval*
Intercept	$(1-\alpha)\%$	$\hat{\beta}_0 \pm t_{\alpha/2} s_{\hat{\beta}_0}$
Slope	$(1-\alpha)\%$	$\hat{\beta}_1 \pm t_{\alpha/2} s_{\hat{\beta}_1}$

The confidence coefficient, α identical to the critical value used in hypothesis testing, is based on a *t*-distribution with $n-2$ degrees of freedom and is chosen by the modeler to give a specified level of confidence. Clearly, a larger interval will yield a higher level of confidence and vice versa. The most common confidence widths are 90%, 95%, and 99%.

After the formulas above are used to construct confidence intervals for a population parameter, one can check to see if a certain value of interest falls in the interval. If we are testing that the independent variable affects *Y*, for example, we should construct a confidence interval for the slope. If zero is contained in the interval, then the data does not give sufficient evidence that *X* affects *Y*. Conversely, if the interval does not hold zero, then there is evidence that the two variables are related at the α level of confidence. The conclusions obtained from building confidence intervals will give the exact same results as using hypothesis tests.

Prediction

Another reason for using ordinary least squares regression is to predict future observations. Forecasting sales as a function of marketing expenses is an example. Prediction can be summarized in one paragraph in the following way. Once a linear model has been fitted using previous data, our best guess of *Y* (call it $\hat{Y}_p$) for a specified value of *X* (name it X_p) is $\hat{Y}_p = \hat{\beta}_0 + \hat{\beta}_1 X_p$. We can also quantify our uncertainty about the estimate. A $(1-\alpha)\%$ confidence interval for a new observation Y_p is $\hat{Y}_p \pm t_{\alpha/2} s_p$ where the standard error of the prediction is

$$s_p = s\sqrt{1+\frac{1}{n}+\frac{(X_p-\bar{X})^2}{\sum_{i=1}^{n} x_i^2}}. \qquad (12)$$

Caveat: A warning about prediction. The estimates and intervals for new values of *Y* are only valid if X_p falls within the range of the *X* values used in the

regression. Extrapolation with *X*'s outside its range is silly. It falls outside the range of experiences.

1.4 Goodness of fit

Another critical aspect of regression analysis is model testing. This can be used when trying to assess a model's predictive power or when choosing between two or more models. A common way to measure goodness of fit is by decomposing the sum of squares of the data into the amount explained by the model and the amount left unexplained. The higher the amount explained by the model, the better the model. The decomposition is done as follows. For a single variable *Y* with n observations, the *total sum of squares* is given by $SST = \sum_{i=1}^{n} y_i^2$. Once we have fitted a model, we obtain the fitted values $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ for each observation of *X*. The squared distances between these predictions and the overall mean $\overline{Y}$ give the *regression* or *explained sum of squares*, $SSR = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2 = \sum_{i=1}^{n} \hat{y}_i^2$ which is the amount of the *SST* explained by the model. Finally, the part left unexplained by the model, called the *error* or *residual sum of squares* is just $SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2$. By Pythagorus' Theorem, the regression and error sum of squares must add up to the total sum of squares. This type of decomposition is closely related to analysis of variance (ANOVA) and is often used to assess how well an estimated model fits the data.

To illustrate, consider now a model that perfectly predicts all the data points, that is, $\hat{Y}_i = Y_i$ for all *i*. In this case, the regression sum of squares equals the total sum of squares, and the error sum of squares equals zero—a perfect fit. On the other hand, a model with an estimated slope $\hat{\beta}_1 = 0$ will have no predictive power at all (because $\hat{Y}_i = \hat{\beta}_0$ for all *i*), and therefore the total sum of squares will equal the error sum of squares, leaving the variation explained by the model as zero.

<u>R-squared</u>

A commonly used indicator of regression goodness of fit is the R^2 statistic. It is also referred to as the *coefficient of determination* and is the proportion of the total variation that is explained by the model. In the case of simple linear regression, R^2 is the square of the correlation between *X* and *Y*. The R^2 is simply the ratio of the regression sum of squares (*SSR*) to the total sum of squares (*SST*):

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n} \hat{y}_i^2}{\sum_{i=1}^{n} y_i^2}.$$

Since the range of *SSR* is 0 to *SST*, the range R^2 is from 0 to 1. A perfect model fit will yield an R^2 of 1, and a model with no explanatory power whatsoever gives $R^2 = 0$. In general, a model with a high R^2 is preferred to one with a low one.

1.5 OLS regression through origin

On occasion, it is necessary to consider a simple regression whose intercept term, for economic reasons, equals zero, that is,

$$Y_i = \beta_1 X_i + \varepsilon_i . \tag{13}$$

As before, the error term ε_i is assumed to be independent, identically (normally) distributed with mean zero and constant variance. The least squares estimator of β_1 is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}, \tag{14}$$

which is like (B) except the levels of X_i and Y_i are used rather than their deviations from their respective means. The standard error of the estimate, $s(\hat{\beta}_1)$ IS DEFINED AS

$$s(\hat{\beta}_1) = \frac{s}{\sqrt{\sum_{i=1}^{n} X_i^2}} \tag{15}$$

where

$$s = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-1}} \tag{16}$$

The standard error of the prediction is

$$s_p = s \sqrt{1 + \frac{X_p^2}{\sum_{i=1}^{n} X_i^2}} . \tag{17}$$

2. *Multiple linear regression*

Analogous to simple linear regression models, we can fit a model using two or more explanatory variables of the form,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \varepsilon_i \qquad (18)$$

where k denotes the number of independent variables in the model, Y is the dependent variable, $X_1, \cdots, X_k$ are the independent variables, and $\beta_0, \beta_1, \cdots, \beta_k$ are the $k+1$ regression coefficients. This is referred to as a *multiple linear regression model* because the equation is linear in the parameters. Note that $X_2, \cdots, X_k$ could all be functions of X_1 such as X_1^2 or $\ln X_1$, and this would still be considered a linear model. The idea of multiple regression is that more than one independent variable may be needed to predict Y effectively. Added variables may supply more explanatory power.

2.1 Assumptions

The assumptions for multiple linear regression are the same as the five we used in the simple regression model. First, the relation between X and Y is assumed to be linear except that now there are multiple Xs, as shown in (18). Second, the values of the Xs are non-random. In addition, we require that there is no exact linear relation among any of the independent variables. Finally, the error term ε is assumed to be independent and identically (normally) distributed with mean 0 and constant variance σ^2.

2.2 Estimation

Like simple regression, multiple linear regression models are fitted using ordinary least squares. Again, the requirement is that the sum of squared errors (*SSE*) is minimized. Estimation of the parameters in the case $k > 1$ involves using linear algebra and matrix inversion, so the calculations are tedious to perform.

2.3 Hypothesis tests for individual parameters

Once we have fitted a multiple regression model, we may wish to find out whether a particular independent variable has a significant effect on Y. We do this by testing the hypothesis that the corresponding parameter value is equal to zero. The procedure is almost identical to the case of simple regression models, except that, in a more general sense, the test statistic $t = \hat{\beta}_i / s_{\hat{\beta}_i}$ has a *t*-distribution with $n-k-1$ degrees of freedom. Most regression software packages automatically display the parameter estimates, standard errors, *t*-ratios, and *p*-values when performing a linear regression.

Confidence intervals

Similarly to hypothesis testing, we can also learn about the effects of individual explanatory variables by constructing confidence intervals. For each of

the $k+1$ parameters in the model, we can obtain interval estimates using the regression output from software packages in the following manner:

$$\text{A } (1-\alpha)\% \text{ confidence interval for } \beta_i \text{ is: } \hat{\beta}_i \pm t_{\alpha/2} s_{\hat{\beta}_i}$$

To see whether the i^{th} explanatory variable affects Y, we check for the presence of zero in the confidence interval for β_i.

2.4 Hypothesis tests for significant overall regression

In addition to testing whether individual explanatory variables influence Y, we may also want to check whether the model has significant predictive power. We can do this using an *F*-test, where the null hypothesis under question is that all coefficients are equal to zero: $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$. This is compared with the alternative hypothesis that at least one of the coefficients is non-zero. The test statistic here is denoted by F, which can be regarded as a *signal-to-noise* ratio. The signal refers to the part of the variation explained by the model, and the noise relates to the amount left unexplained. The *F*-test can also be understood by examining the following generic analysis of variance (ANOVA) table resulting from a linear regression procedure.

ANOVA Table

Source	*Sum of squares*	*Degrees of freedom*	*Mean squared error*	*F-ratio*
Regression	*SSR*	k	SSR/k	$\dfrac{SSR/k}{SSE/(n-k-1)}$
Error	*SSE*	$n-k-1$	$SSE/(n-k-1)$	
Total	*SST*	$n-1$	$SST/(n-1)$	

An analysis of variance (ANOVA) table like the one above is standard output from most software packages when a regression procedure is performed. The first column shows how the total variation of Y is decomposed: *Regression*, the part explained by the model; *Error*, the part unexplained by the model; and, *Total*, both parts put together. In the *Sum of squares* column, *SSR*, *SSE*, and *SST* stand for regression, error, and total sum of squares, respectively. The *degrees of freedom* (*df*) are divided up as follows: total degrees of freedom is the number of observations minus one (which is lost because the overall mean $\overline{Y}$ is estimated); degrees of freedom for the regression is equal to the number of explanatory variables in the model; and the difference between the two gives the degrees of freedom of the error. The error *df* is also the degrees of freedom on which *t*he t-test for individual parameter significance is based. The *mean squared error* column gives essentially

the average sum of squares for each source. Note that the mean squared total is the unbiased estimate for the variance of *Y*. The *F*-statistic in column five can be thought of as the signal-to-noise ratio: the regression mean squared, SSR/k, being the signal, and the error mean squared, $SSE/(n-k-1)$, as the noise. If the *F*-statistic gets high, the regression explains a substantial proportion of the variance; if it gets low, much is left unexplained. Therefore, we reject the null hypothesis of a useless model if F it is large enough. And, for any given significance level, we can find the critical value of *F* using an *F*-distribution with k and $n-k-1$ degrees of freedom.

Prediction

For any combination of independent variables lying in the range of the observed X's, we can obtain point estimates and prediction intervals by using a formula like the one given for the simple regression case. The only difference is that the standard error, which depends on the distance from the prediction point X_p and the mean $\bar{X}$, must be measured in *k*-dimensional space. However, most statistical software packages provide this output, so no matrix algebra is necessary.

2.5. Model selection and goodness of fit

Adjusted R-squared

As with simple regression, goodness-of-fit analysis for multiple regression is based on the sum of squares decomposition. The coefficient of determination, R^2 is widely used because of its ease of interpretation. R^2 can be misleading, however, because it increases monotonically with *k*. In other words, with each independent variable added to the model, the coefficient of determination must either increase or remain the same. This becomes a problem when extraneous variables are included in the model solely to boost the value of R^2, ignoring scientific or statistical indications that they do not belong. Parsimony, or economy of explanation, is valuable because it eases the burden of understanding and interpreting the model. The adjusted R^2 (or $\bar{R}^2$) accounts for the number of independent variables when assessing goodness of fit in multiple regression. Its formula is

$$\bar{R}^2 = 1 - \frac{Var(e)}{Var(Y)} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} = 1 - (1-R^2)\left(\frac{n-1}{n-k-1}\right), \qquad (19)$$

where *k* is the number of independent variables in the model. We can see that the $\bar{R}^2$ is always less than or equal to R^2. (It only equals R^2 when the model has only an intercept.) It is also possible for $\bar{R}^2$ to be negative. The advantage $\bar{R}^2$ is that it penalizes the model for including variables that do not provide information about *Y*. Note that, for two models with the same R^2 but different *k*, the $\bar{R}^2$y will be larger for the smaller model. This is intended to ensure that the issue of parsimony is addressed in the model selection process.

The $\bar{R}^2$ percentage of the variation of Y around its mean is measured, which is explained by the regression equation and adjusted for degrees of freedom. An increase in $\bar{R}^2$ says that the marginal benefit of adding a variable outweighs the cost while a decrease says that marginal cost outweighs the benefit. The $\bar{R}^2$ is usually preferred as a measure of best fit over the R^2 because it can be used to compare the fits of equations with the same dependent variable and different numbers of independent variables. However, it is essential to note that quality of fit of the regression equation is only one measure of the regression quality and that maximizing the $\bar{R}^2$ is not always the best way to maximize the quality of an equation. Theoretical justification is also imperative.

3. *Violations of OLS assumptions*

3.1 Specification errors

Implicit in the specification of the multiple regression model (18) are the assumptions that we know the identity of all the *k* relevant explanatory variables and that their relationship with the dependent variable is linear. But, since regression is, by nature, exploratory, we need to be concerned about the "correctness" of our specification. What impact does failing to include a relevant explanatory variable have on estimation? Along the same line, what is the effect of including an explanatory variable that does not belong in the regression model? Finally, what is the impact of estimating a linear relation when the actual relation is nonlinear? We address each of these issues in turn.

3.2. Omitting relevant explanatory variables

Not including a relevant explanatory variable can have profound implications. To see this, assume that the "true" model is

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \,, \tag{20}$$

where the variables are expressed as deviations from their means. Now, suppose that, instead of estimating (20), we estimate

$$y_i = \beta_1^* x_{1i} + \varepsilon_i^* \ . \tag{21}$$

What are the implications?

We know from our discussion of simple linear regression that the estimated slope in (21) is:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}. \tag{22}$$

Substituting (20) for y_i, we obtain

$$\begin{aligned} \hat{\beta}_1^* &= \frac{\beta_1 \sum_{i=1}^{n} x_{1i}^2 + \beta_2 \sum_{i=1}^{n} x_{1i} x_{2i} + \sum_{i=1}^{n} x_{1i} \varepsilon_i}{\sum_{i=1}^{n} x_{1i}^2} \\ &= \beta_1 + \frac{\beta_2 \sum_{i=1}^{n} x_{1i} x_{2i}}{\sum_{i=1}^{n} x_{1i}^2} + \frac{\sum_{i=1}^{n} x_{1i} \varepsilon_i}{\sum_{i=1}^{n} x_{1i}^2}. \end{aligned} \tag{23}$$

Since X_1 is fixed and the expected value of the error is 0, the expected value of the estimated slope is

$$E(\hat{\beta}_1^*) = \beta_1 + \beta_2 \frac{\sum_{i=1}^{n} x_{1i} x_{2i}}{\sum_{i=1}^{n} x_{1i}^2}. \tag{24}$$

This means that, in general, the estimated slope parameter in (21) will be biased. The direction of the bias depends on the product of β_2 and the covariance between the independent variables. If β_2 and the covariance are both positive, the estimated slope will be upward biased. The intuition is that the estimated slope picks up the co-variation of y_i with x_{1i} and some of the co-variation of y_i with x_{2i}. Only if the correlation between the independent variables is zero will the estimated slope be unbiased.

The standard error of the estimate will also be biased. In the case of estimating (21) when (14) is the correct model, the standard error of $\hat{\beta}_1^*$ will be less than the standard error of $\hat{\beta}_1$. We thereby run the risk of rejecting the null hypothesis that the slope parameter is 0 when it is. If we omit an important variable from the equation, then we violate the classical assumption that the explanatory variables are independent of the error term. This is because the stochastic error term includes the effects of any omitted variables. For example, if you are given:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i ,$$

In addition, the stochastic error term is redefined as:

$$\varepsilon_i^* = e_i + \beta_{22} X_i^2$$

Thus, the presence of omitted variable bias violates one of the classical assumptions of OLS and the Gauss-Markov Theorem (which says that OLS is the minimum variance unbiased estimator).

3.3 Including irrelevant explanatory variables

The effects of including an irrelevant variable are much less severe. This is because the estimated parameters of the relevant explanatory variables stay unbiased. The only cost, so to speak, is that the standard errors of the estimates will be larger than they should be, making it more difficult to reject the null hypothesis of a zero-slope parameter. Thus, if you reject the null in the presence of irrelevant variables, you can be quite confident of your decision.

3.4 Nonlinearities

A separate discussion of the effects of nonlinearity is unwarranted. Fitting a linear regression to a nonlinear relation is a particular case of the omitted variables discussed above. We can expect bias in the estimated coefficients and the standard errors to be smaller than appropriate.

3.5 Multicollinearity

Multicollinearity arises when two or more variables are highly correlated with each other. While it is possible to obtain least squares estimates of the regression coefficients, the interpretation of the coefficients is difficult. Recall that the interpretation of a regression coefficient is the change in Y with respect to a change in X_1, *holding other factors constant*. The presence of multicollinearity means that other factors are not being held constant. If X_1 and X_2 are highly correlated, a change in X_1 implies a change in X_2, and vice versa.

A rule of thumb states that multicollinearity is likely to be a problem if a simple correlation between the two variables is larger than the correlation of either or both variables with the dependent variable. Such a rule may be reasonable when there are only two independent variables in the regression. However, simple correlations with more than two independent variables will not detect a more complicated linear relation among variables. The simplest way to see if multicollinearity is a problem is to examine the standard errors of the coefficients. If several coefficients have high standard errors dro,pping one or more variables from the equation lowers the standard errors of the remaining variables. In this case, multicollinearity will usually be the source of the problem.

Consequences of multicollinearity

1. Estimates will remain unbiased.
2. The variances and standard errors of the estimates will increase.
3. The computed t-scores will fall.
4. Estimates will become sensitive to changes in specification.
5. The overall fit of the equation and the estimation of the coefficients of non-multicollinear variables will be largely unaffected.

Remedies for multicollinearity

1. Do nothing. Action need only be taken if the consequences cause insignificant t-scores or unreliable estimated coefficients.
2. Drop a redundant variable.
3. Increase the size of the sample. This reduces the variance of the estimated beta coefficients, which aids in minimizing the impact of multicollinearity in the regression.

It is essential to recognize that multicollinearity exists in every regression equation. Thus, the focus is not to decide if multicollinearity exists in your regression but rather its degree. While there is no perfect statistical test, there are tools researchers used to detect it.

One method is by evaluating the correlation coefficient (r) between the explanatory variables. If the absolute value of r is closer to 1, then this indicates a strong correlation between the explanatory variables and that multicollinearity is a potential problem.

3.6 Violations of disturbance assumption

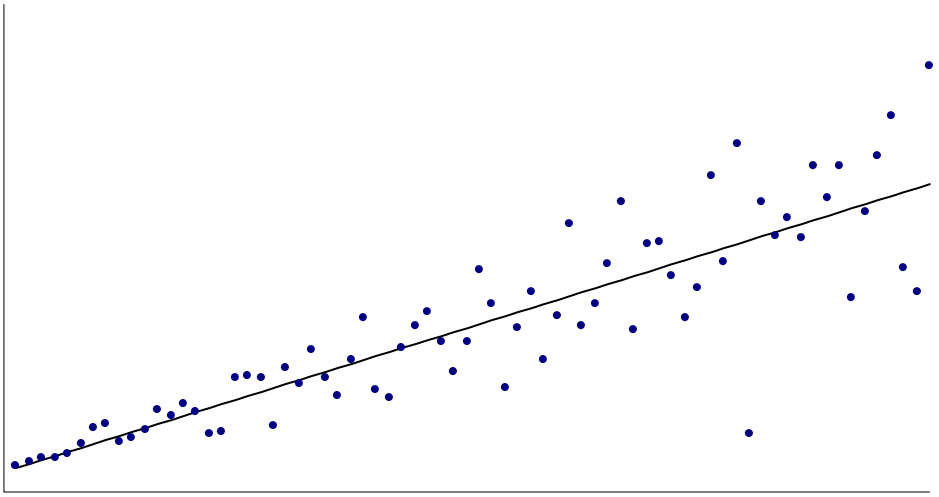
OLS regression assumes that the relation between dependent and independent variables is linear and that the error term ε is independent and identically (normally) distributed with mean 0 and constant variance σ^2. However, these assumptions can be violated in four ways: (a) the relation is nonlinear, (b) the error variance may not be constant, (c) the residual errors may be correlated, and (d) the errors may not be normally distributed. Below, we discuss how to detect such violations, explain the consequences of each violation, and suggest remedies to fix or at least mitigate the effects of the violation.

3.7 Heteroscedasticity

A plot of the residuals around the fitted values of Y can help detect a violation of the constant variance (or homoscedastic) error assumption. If the residuals have a constant variance, the error term is homoscedastic. If the residuals appear fan-shaped, as shown in the figure below, heteroscedasticity may be a problem. When looking at the graphed plot of your regression, if you see a

pattern/ a clustering in the residuals then this is an indication of heteroscedasticity. As the scatter should be random and unpredictable for homoscedasticity to be present. A popular test for the presence of heteroscedasticity is the Goldfeld-Quandt (1965) test. The steps are as follows:

a) Order the data by the X_i observations.
b) Omit the c central observations.[1]
c) Fit separate regressions to the first $(n-c)/2$ and the last $(n-c)/2$ observations. Naturally, $(n-c)/2$ must exceed the number of parameters to be estimated.
d) Compute the ratio $R = SSE_2 / SSE_1$, where SSE_1 and SSE_2 are the sum of squared errors from the first and second regressions, respectively.[2] Under the assumption of homoscedasticity, R has an F distribution with $(n-c)/2$ $(n-c)/2$ degrees of freedom in the numerator and the denominator, respectively.



With a heteroscedastic error term, ordinary least squares estimation places greater weight on observations with large error variances than those with small errors. This implicit weighting occurs because the sum of the squared residuals associated with the large variance error terms is greater than that of the squared residuals associated with small error variances. While the parameter estimates stay unbiased, the standard errors are biased.

Correcting for heteroscedasticity is possible using *weighted least squares* estimation. To illustrate, suppose that the error variances in the regression,

[1] The power of the test depends on the choice of c. The greater the value of c, the lower the power of the test. On the other hand, the lower the value of c, the greater the power, however, the more likely the residual variances will move closer together.

[2] Place the largest *SSE* in the numerator. Hence, the notation implicitly assumes $SSE_2 > SSE_1$.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \varepsilon_i , \tag{25}$$

vary directly with one of the explanatory variables, say, $X_{i,1}$, that is,

$$Var(\varepsilon_i) = CX_{1i}^2 ,$$

where *C* is a non-zero constant. To correct for heteroscedasticity, we multiply the terms of the regression by the inverse of X_{1i}, and run the regression,

$$\frac{Y_i}{X_{1i}} = \beta_0 \frac{1}{X_{1i}} + \beta_1 + \cdots + \beta_k \frac{X_{ki}}{X_{1i}} + \frac{\varepsilon_i}{X_{1i}} . \tag{26}$$

The error term in the transformed regression model, $\varepsilon_i^* = \frac{\varepsilon_i}{X_{1i}}$, now has constant variance, that is,

$$Var\left(\varepsilon_i^*\right) = Var\left(\frac{\varepsilon_i}{X_{1i}}\right) = \frac{1}{X_{1i}} Var\left(\varepsilon_i\right) = C . \tag{27}$$

Uncovering the form of heteroscedasticity is sometimes tricky since the error variance may be a nonlinear function of one of the independent variables, or it may be a function of other variables, *Z*, not included in the regression model. Standard econometric textbooks offer guidance on appropriate correction procedures. After the structure of the error variance is determined, however, variable transformations in a weighted least squares framework rectifies the problem.

3.8. Serial correlation

Pure serial correlation violates the classical assumption that assumes uncorrelated observations of the error term and breaks the Gauss-Markev Theorem of minimum variance. First-order serial correlation occurs when the current value of the error term is a function of the previous value of the error term. Serial correlation is seen in time series data sets denoted as Y_t.

The presence of serial correlation does not bias the parameter estimates. It does, however, affect the standard errors of the estimates. In the presence of positive serial correlation, the standard errors will be smaller than they should be, potentially causing us to reject the null when we should not. To understand possible correction procedures, consider the nature of the problem. Under the assumption that the error term is serially correlated, the regression model is

$$Y_t = \beta_0 + \beta_1 X_{1t} + \cdots + \beta_k X_{kt} + \varepsilon_t , \tag{28}$$

where

$$\varepsilon_t = \rho \varepsilon_{t-1} + \upsilon_t , \tag{29}$$

ρ is the first-order serial correlation, and υ_t is normally distributed with zero mean and constant variance and is independent of other errors through time. Since equation (28) holds for all time periods, we can write

$$Y_{t-1} = \beta_0 + \beta_1 X_{1t-1} + \cdots + \beta_k X_{kt-1} + \varepsilon_{t-1}. \tag{30}$$

Multiplying (30) by ρ and subtracting it from (28), we get

$$Y_t^* = \beta_0 (1-\rho) + \beta_1 X_{1t}^* + \cdots + \beta_k X_{kt}^* + \upsilon_t, \tag{31}$$

where the asterisks denote *generalized differences*. The variable X_{1t}^*, for example, is defined as $X_{1t}^* = X_{1t} - \rho X_{1t-1}$. Since the error term in (31) is independent through time, the standard errors of the regression model (31) will be unbiased.

Positive serial correlation, in which ρ is positive, implies that the error term tends to have the same sign from one time period to the next. For example, a massive external shock to an economy, such as Covid-19, may have lingering effects that last over several years. As it will take multiple time periods to heal from such an economic shock.

In the case of negative serial correlation, the negative ρ value implies that the error term has the tendency to switch signs from negative to positive consecutively. Negative serial correlation suggests a cyclical pattern is present. However negative serial correlation is much less likely to occur than positive serial correlation.

3.9. Non-normality

A violation of the normality assumption is severe. The reason is simple. Since parameter estimation is based on the minimization of the sum of *squared error*s, a few extreme observations can disproportionately influence the parameter estimates and their standard errors. One way to test for normally distributed errors is to use a *normal probability plot* of the residuals. A normal probability plot is a plot of the fractals of error distribution versus those of a normal distribution with the same mean and variance. If the distribution is normal, the points on this plot should fall close to the diagonal line. A "bow-shaped" pattern of deviations from the diagonal indicates that the residuals have excessive skewness (i.e., they are not symmetrically distributed, with too many significant errors in the same direction). An "S-shaped" pattern of deviations indicates that the residuals have excessive kurtosis (i.e., there are either too many or too few significant errors in both directions).

Normality violations often arise because (a) the distributions of the dependent or independent variables are non-normal or (b) the linearity assumption is violated. In such cases, a nonlinear transformation of variables might cure both problems. In some cases, the problem with the residual distribution is due to one or two huge errors called *outliers*. Scrutinize outliers

closely. If they are merely errors or unique events not likely to be repeated, you may have cause to remove them.

4. Summary

Summarizing the content of these notes in a paragraph is impossible. If any of the concepts need to be clarified, return to the ***Objectives*** listed on the first page and examine the big picture. On the other hand, you can proceed by reading the next set of notes in which regression analysis is applied to an important financial markets problem.