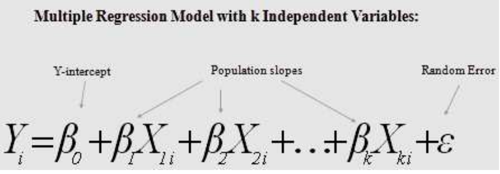
2. MULTIPLE LINEAR REGRESSION

Multiple linear regression is a method used to model the linear relationship between a dependent variable and more than one independent (explanatory or regressors) variables. A multiple linear regression model has the following general form:



$$Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+..+\beta_kX_{ki}+\varepsilon$$

where,

Y_i = ith observation of dependent variable Y
X_{ki} = ith observation of kth independent variable X
β_0 = intercept term
β_k = slope coefficient of kth independent variable
ε_i = error term of ith observation
n = number of observations
k = total number of independent variables

- ***A slope coefficient, β_j is known as partial regression coefficients or partial slope coefficients.*** It measures how much the dependent variable, Y, changes when the independent variable, X_j, changes by one unit, ***holding all other independent variables constant.***
- ***The intercept term (β_0)*** is the value of the dependent variable when the independent variables are all equal to zero.
- A regression equation has k slope coefficients and k + 1 regression coefficients.

Practice: Example 1
Volume 1, Reading 10.

Simple vs. Multiple Regression

Simple Regression	Multiple Regression
1. One dependent variable Y predicted from one independent variable X 2. One regression coefficient 3. **r^2**: proportion of variation in dependent variable Y predictable from X	1. One dependent variable Y predicted from a set of independent variables (X_1, X_2 …. X_k) 2. One regression coefficient for each independent variable 3. **R^2**: proportion of variation in dependent variable Y predictable by set of independent variables (X's)

2.1 Assumptions of the Multiple Linear Regression Model

The Multiple linear regression model is based on following six assumptions. When these assumptions hold, the regression estimators are ***unbiased, efficient*** and ***consistent***.

NOTE:

- Unbiased means that the expected value of the estimator is equal to the true value of the parameter.
- Efficient means that the estimator has a smaller variance than any other estimator.
- Consistent means that the biasness and variance of the estimator approach zero as the sample size increases.

Assumptions:

1. The relationship between the dependent variable, Y, and the independent variables, X_1, X_2, . . . ,X_k, is linear.
2. The independent variables (X_1, X_2, . . . ,X_k) are not random. Also, no exact linear relation exists between two or more of the independent variables.
3. The expected value of the error term, conditional on the independent variables, is 0: $E(\varepsilon|\ X_1, X_2, . . . , X_k) = 0$.
4. The variance of the error term is constant for all observations i.e. errors are ***Homoskedastic***.
5. The error term is uncorrelated across observations (i.e. ***no serial correlation***).
6. The error term is normally distributed.

NOTE:

- Linear regression can't be estimated when an exact linear relationship exists between two or more independent variables. But when two or more independent variables are highly correlated, although there is no exact relationship, it leads to multicollinearity problem. (Discussed later in detail).
- Even if independent variable is random but uncorrelated with the error term, regression results are reliable.

Practice: Example 2 & 3
Volume 1, Reading 10.

2.2 Predicting the Dependent Variable in a Multiple Regression Model

The process of calculating the predicted value of dependent variable is the same as we did in Reading 11.

Prediction equation

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_{1i} + \hat{b}_2 X_{2i} + \ldots + \hat{b}_k X_{ki}$$

where,

$\hat{Y}_i$: Estimated or predicted value of Y
b_0: Estimated intercept
b_1, b_2,… & b_k: Estimated slope coefficients

Assumptions of the regression model must hold in order to have reliable prediction results.

Practice: Example 4
Volume 1, Reading 10.

Sources of uncertainity when using regression model & estimated parameters:

1. Uncertainity in error term.
2. Uncertainity in the estimated parameters of the model.

2.3 Testing Whether All Population Regression Coefficients Equal Zero

To test the significance of the regression as a whole, we test the null hypothesis that all the slope coefficients in a regression are simultaneously equal to 0.

H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ (no linear relationship)
H_1: at least one $\beta_i \neq 0$ (at least one independent variable affects Y)

In multiple regression, the F-statistic is used to test whether at least one independent variable, in a set of independent variables, explains a significant portion of variation of the dependent variable. The F statistic is calculated as the ratio of the mean regression sum to squares of the mean squared error,

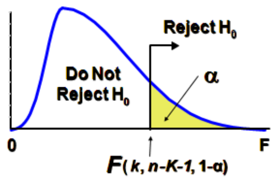
$$\frac{MSR}{MSE} = \frac{RSS/_{k}}{SSE/_{n-k-1}}$$

df numerator = k
df denominator = n – k – 1

Note: F-test is always a one-tailed test.

Decision Rule: Reject H_0 if F>F-critical.

Reject H_0 in favor of H_a if f_{calc} falls in colored area



- Reject H_0 for H_a if P-value = P(F>f_{calc})<α

NOTE:

When independent variable in a regression model does not explain any variation in the dependent variable, then the predicted value of y is equal to mean of y. Thus, RSS = 0 and F-statistic is 0.

- Larger R^2 produces larger values of F.
- Larger sample sizes also tend to produce larger values of F.
- *The lower the p-value, the stronger the evidence against that null hypothesis.*

Example:

k = 2
n = 1,819
df = 1,819 – 2 – 1 = 1,816
SSE = 2,236.2820
RSS = 2,681.6482
α = 5%
F-statistic = $\frac{MSR}{MSE}$= (2,681.6482/2) / (2,236.2820/1,816) = 1,088.8325

F-critical with numerator df = 2 and denominator df = 1,816 is 3.00.

Since F-statistic > F-critical, Reject H_0 that coefficients of both independent variables equal 0.

2.4 Adjusted R^2

In multiple linear regression model, R^2 is less appropriate as a measure to test the "goodness of fit" of the model because R^2 always increases when the number of independent variables increases. It is important to keep in mind that a high R^2does not imply causation.

The ***adjusted R^2*** is used to deal with this artificial increase in accuracy. Adjusted R^2 does not automatically increase when another variable is added to a regression; it is adjusted for degrees of freedom. The *adjusted R^2* is given by

$$\bar{R}^2 = 1 - \left(\frac{n-1}{n-k-1}\right)(1 - R^2)$$

where,

n = sample size,
k = number of independent variables

- When k ≥ 1, then R^2 is strictly > Adjusted R^2.
- Adjusted R^2 decreases if the new variable added does not have any significant explanatory power.

- Adjusted R^2 can be negative as well but R^2 is always positive.
- Adjusted R^2 is always ≤ R^2.

NOTE:

When Adjusted R^2 is used to compare regression models, both the dependent variable definition and sample size must be same for each model.

3. USING DUMMY VARIABLES IN REGRESSIONS

Dummy variable is a qualitative variable that takes on a value of 1 if a particular condition is true and 0 if that condition is false. It is used to account for qualitative variables such as male or female, month of the year effects, etc.

Suppose we want to test whether total returns of one small-stock index, the Russell 2000 Index, differ by months. We can use dummy variables to estimate the following regression,

$$Returns_t = b_0 + b_1 jan_t + b_2 Feb_t + \ldots + b_{11} Nov_t + \varepsilon_t$$

- If we want to distinguish among n categories, we need n -1 dummy variables e.g. in above regression model we will need 12 – 1 = 11 dummy variables. If we take 12 dummy variables, Assumption 2 is violated.
- b_0 represents average return for stocks in December.
- $b_1, b_2, b_3, \ldots, b_{11}$ represent difference between returns in that month and returns for December i.e.
  - Average stock returns in Dec = b_0
  - Average stock returns in Jan = $b_0 + b_1$
  - Average stock returns in Feb = $b_0 + b_2$
  - Average stock returns in Nov = $b_0 + b_{11}$

As with all multiple regression results, the F-statistic for the set of coefficients and the R^2 are evaluated to determine if the months, individually or collectively, contribute to the explanation of monthly return. We can also test whether the average stock return in each of the months is equal to the stock return in Dec (the omitted month) by testing the individual slope coefficient using the following null hypotheses:

H_0: $b_1 = 0$ (i.e. stock return in Dec = stock return in Jan)
H_0: $b_2 = 0$ (i.e. stock return in Dec = stock return in Feb)
and so on….

Practice: Example 5
Volume 1, Reading 10.

4. VIOLATIONS OF REGRESSION ASSUMPTIONS

4.1 Heteroskedasticity

Heteroskedasticity occurs when the variance of the errors differs across observations i.e. variances are not constant.

Types of Heteroskedasticity:

1. Unconditional Heteroskedasticity: It occurs when Heteroskedasticity of the error variance does not systematically increase or decrease with changes in the value of the independent variable. Although it violates Assumption 4, but it creates no serious problems with regression.

2. Conditional Heteroskedasticity: Conditional heteroskedasticity exists when Heteroskedasticity of the error variance increases as the value of independent variable increases. It is more problematic than unconditional hetroscadasticity.

4.1.1) Consequences of (Conditional) Heteroskedasticity:

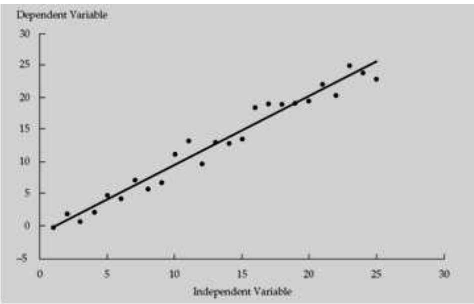
- It does not affect consistency but it can lead to wrong inferences.
- Coefficient estimates are not affected.
- It causes the F-test for the overall significance to be unreliable.
- It introduces biasness into estimators of the standard error of regression coefficients; thus, t-tests for the significance of individual regression coefficients are unreliable.

When Heteroskedasticity results in underestimated standard errors, t-statistics are inflated and probability of Type-I error increases. The opposite will be true if standard errors are overestimated.

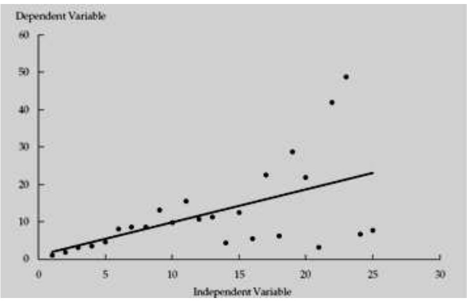
4.1.2) Testing for Heteroskedasticity:

1. ***Plotting residuals:*** A scatter plot of the residuals versus one or more of the independent variables can describe patterns among observations (as shown below).

Regressions with Homoskedasticity



Regressions with Heteroskedasticity



2. Using ***Breusch–Pagan test***: The Breusch–Pagan test involves regressing the squared residuals from the estimated regression equation on the independent variables in the regression.

H_0 = No conditional Heteroskedasticity exists
H_A = Conditional Heteroskedasticity exists

Test statistic = $n \times R^2_{residuals}$

where,

$R^2_{residuals}$ = R^2 from a second regression of the squared residuals from the first regression on the independent variables
n = number of observations

- Critical value is calculated from χ^2 distribution table with df = k.
- It is a one-tailed test since we are concerned only with large values of the test statistic.

Decision Rule: When test statistic > critical value, Reject H_0 and conclude that error terms in the regression model are conditionally Heteroskedastic.

- If no conditional heteroskedasticity exists, the independent variables will not explain much of the variation in the squared residuals.
- If conditional heteroskedasticity is present in the original regression, the independent variables will explain a significant portion of the variation in the squared residuals.

Practice: Example 8
Volume 1, Reading 10.

4.1.3) Correcting for Heteroskedasticity:

Two different methods to correct the effects of conditional heteroskedasticity are:

1. Computing ***robust standard errors*** (heteroskedasticity-consistent standard errors or white-corrected standard errors), corrects the standard errors of the linear regression model's estimated coefficients to deal with conditional heteroskedasticity.

2. ***Generalized least squares*** (GLS) method is used to modify the original equation in order to eliminate the heteroskedasticity.

4.2 Serial Correlation

When regression errors are correlated across observations, then errors are serially correlated (or auto correlated). Serial correlation most typically arises in time-series regressions.

Types of Serial Correlation:

1. ***Positive serial correlation*** is a serial correlation in which a positive (negative) error for one observation increases the probability of a positive (negative) error for another observation.

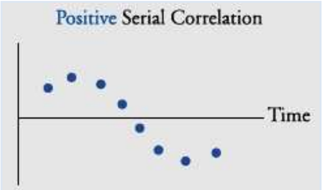
2. ***Negative serial correlation*** is a serial correlation in which a positive (negative) error for one observation increases the probability of a negative (positive) error for another observation.

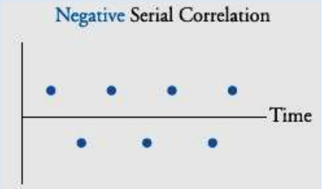
4.2.1) Consequences of Serial Correlation:

- The principal problem caused by serial correlation in a linear regression is an incorrect estimate of the regression coefficient standard errors.
- When one of the independent variables is a lagged value of the dependent variable, then serial correlation causes all the parameter estimates to be inconsistent and invalid. Otherwise, serial correlation does not affect the consistency of the estimated regression coefficients.
- Serial correlation leads to wrong inferences.
- In case of positive (negative) serial correlation: *Standard errors are underestimated (overestimated) → T-statistics (& F-statistics) are inflated (understated) →Type-I (Type-II) error increases.*

4.2.2) Testing for Serial Correlation:

1. ***Plotting residuals*** i.e. a scatter plot of residuals versus time (as shown below).





2. ***Using Durbin-Watson Test:*** The Durbin Watson statistic is used to test for serial correlation

$$DW = \frac{\sum_{t=2}^{T}(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^{T}\hat{\varepsilon}_t^{\,2}}$$

where,

$\hat{\varepsilon}_t$ is the regression residual for period t.
The DW statistic tests the null hypothesis of no autocorrelation against the alternative hypothesis of positive (or negative) autocorrelation. In case of large sample size, Durbin-Watson statistic (d) is approximately equal to

$$\mathbf{d} \approx 2\,(1 - r)$$

where,

r = sample correlation b/w regression residuals from one period and from the previous period.

The above equation implies that

- d = 2, if no autocorrelation (r = 0)
- d = 0, if autocorrelation is +1.0
- d = 4, if autocorrelation is –1.0

Decision Rule:

A. *For positive autocorrelation, the decision rule is:*

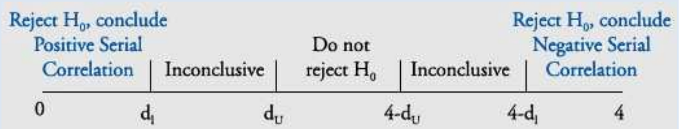
H_0; no positive auto correlation
H_a; positive auto correlation

- If $d < d_l$ → Reject H_0
- If $d > d_u$ → Do not reject H_0
- If $d_l \le d \le d_u$ → Inconclusive

B. *For negative autocorrelation, the decision rule is:*

H_0; no negative auto correlation
H_a; negative auto correlation

- If $d > 4 - d_l$ → Reject H_0
- If $d < 4 - d_u$ → Do not reject H_0
- If $4 – d_u \le d \le 4 - d_l$ → Inconclusive


Reject H_0, conclude Positive Serial Correlation
Inconclusive
Do not reject H_0
Inconclusive
Reject H_0, conclude Negative Serial Correlation
0
d_l
d_U
4-d_U
4-d_l
4

4.2.3) Correcting for Serial Correlation:

The two different methods to correct effects of serial correlation are:

1. **Adjust** the coefficient standard errors for the linear regression parameter estimates to account for the serial correlation e.g. using **Hansen's method**. Hansen's method also simultaneously corrects for conditional heteroskedasticity. (Mostly this method is recommended).
2. **Modify** the regression equation itself to eliminate serial correlation.

4.3 Multicollinearity

Multicollinearity occurs when two or more independent variables (or combinations of independent variables) are highly (but not perfectly) correlated with each other.

4.3.1) Consequences of Multicollinearity:

- A high degree of multicollinearity can make it difficult to detect significant relationships.
- Multicollinearity does not affect the consistency of the estimates of the regression coefficients but estimates become extremely imprecise and unreliable.
- It does not affect F-statistic.
- The multicollinearity problem does not result in biased coefficient estimates; however, standard errors of regression coefficients can increase, causing insignificant t-tests and wide confidence intervals i.e. Type-II error increases.

4.3.2) Detecting Multicollinearity

- High pairwise correlations among independent variables do not necessarily indicate presence of multicollinearity while a low pairwise correlation among independent variables is not an evidence that multicollinearity does not exist. Correlation between independent variables is useful as an indicator of multicollinearity only in case of two independent variables.
- The classic symptom of multicollinearity is a high R^2 (and significant F-statistic) even though the t-statistics on the estimated slope coefficients are not significant.

4.3.3) Correcting for Multicollinearity

The problem of multicollinearity can be corrected by excluding one or more of the regression variables.

4.4 Summarizing the Issues

Problem	How to detect	Consequences	Possible Corrections
(Conditional) Heteroskedasticity i.e. Errors are correlated with earlier X	Plot residuals or use Breusch–Pagan test	Wrong inferences; incorrect standard errors	Use robust standard errors or GLS
Serial correlation i.e. Errors are correlated with earlier errors	Durbin-Watson Test	Wrong inferences; incorrect standard errors	Use robust standard errors (Hansen's method) or modifying equation
Multicollinearity i.e. independent variables are strongly correlated with each other	High R^2 and significant F-statistic but low t-statistic	Wrong inferences;	Omit variable

5. MODEL SPECIFICATION AND ERRORS IN SPECIFICATION

Model specification refers to the set of variables included in the regression and the regression equation's functional form. Incorrect model specification can result in biased & inconsistent parameter estimates and violations of other assumptions.

5.1 Principles of Model Specification

1. The model should be based on logical economic reasoning.
2. The functional form chosen for the variables in the regression should be compatible with the nature of the variables.
3. The model should be *parsimonious* (i.e. economical both in terms of time & cost).
4. The model should be examined for any violation of regression assumptions before being accepted.
5. The model should be tested for its validity & usefulness out of sample before being accepted.

Types of misspecifications:

1. Misspecified Functional Form:

a) *Omitted variables bias*: One or more important variables are omitted from regression.

- When relevant variables are excluded, result can be biased & inconsistent parameter estimates (unless the omitted variable is uncorrelated with the included ones).
- When irrelevant variables are included, standard errors are overestimated.

b) One or more of the regression variables may need to be transformed (for example, by taking the natural logarithm of the variable) before estimating the regression.

c) The regression model pools data from different samples that should not be pooled.

2. *Independent variables are correlated with the error term.* This is a violation of Regression Assumption 3, that the error term has a mean of 0, and causes the estimated regression coefficients to be biased and inconsistent. Three common problems that cause this type of time-series misspecification are:

a) Including lagged dependent variables as independent variables in regressions (with serially correlated errors) e.g. $Y_t = b_0 + b_1X_t + b_2 Y_{t-1} + \varepsilon_t$

b) Including a function of dependent variables as an independent variable i.e. forecasting "past" instead of future e.g. $EPS_t = b_0 + b_1BV_t + \varepsilon_t$; we should rather use $Y_t = b_0 + b_1 BV_{t-1} + \varepsilon_t$

c) Independent variables that are measured with error i.e. due to use of wrong proxy variable. When this problem exists in a single independent variable regression, the estimated slope coefficient on that variable will be biased toward 0.

3. *Other types of Time-series Misspecification* e.g. nonstationarity problem, which results in non-constant mean and variance over time. (Discussed in detail in Reading 13)

6. MODEL WITH QUALITATIVE DEPENDENT VARIABLES

Qualitative dependent variables are dummy variables used as dependent variables instead of independent variables.

- The ***probit model*** is based on the normal distribution and estimates the probability that Y = 1 (a condition is fulfilled) given the value of the independent variable X.
- The ***logit model*** is identical to probit model, except that it is based on the logistic distribution rather than the normal distribution.
- ***Discriminant analysis*** is based on a linear function, similar to a regression equation, which is used to create an overall score. Based on the score, an observation can be classified into categories such as bankrupt or not bankrupt.

Economic meaning of the results of multiple regression analysis and criticism of a regression model and its results:

1. The validity of a regression model is based on its assumptions. When these assumptions do not hold, regression estimates and results are inaccurate and invalid.
2. Regression does not prove causality between variables; it only discovers correlations between variables.
3. Regression Analysis focuses on its use for statistical inference only. A relationship may be statistically significant but has no economic significance e.g. a regression model may identify a statistically significant abnormal return after the dividend announcement, but these returns may prove unprofitable when transactions costs are taken into account.

Practice: End of Chapter Practice Problems for Reading 10 &FinQuiz Item-set ID# 11514, 15830 & 16534.