Regression explained in simple terms

A Vijay Gupta Publication

Regression explained

Published by VJBooks Inc.

You can reach the author at vgupta1000@aol.com.

Library of Congress Catalog No.: Pending
ISBN: Pending
First year of printing: 2000
Date of this copy: April 23, 2000

Publisher: VJBooks Inc.
Editor: Vijay Gupta
Author: Vijay Gupta

About the Author

Vijay Gupta has taught statistics and econometrics to graduate students at Georgetown University. A Georgetown University graduate with a Masters degree in economics, he has a vision of making the tools of econometrics and statistics easily accessible to professionals and graduate students.

In addition, he has assisted the World Bank and other organizations with statistical analysis, design of international investments, cost-benefit and sensitivity analysis, and training and troubleshooting in several areas.

He is currently working on:

- a package of SPSS Scripts "Making the Formatting of Output Easy"
- a manual on Word
- a manual for Excel
- a tutorial for E-Views
- an Excel add-in "Tools for Enriching Excel's Data Analysis Capacity"

Expect them to be available during fall 2000. Early versions can be downloaded from www.vgupta.com.

LINEAR REGRESSION

Interpretation of regression output is discussed in section 1[1]. Our approach might conflict with practices you have employed in the past, such as always looking at the R-square first. As a result of our vast experience in using and teaching econometrics, we are firm believers in our approach. You will find the presentation to be quite simple - everything is in one place and displayed in an orderly manner.

The acceptance (as being reliable/true) of regression results hinges on diagnostic checking for the breakdown of classical assumptions[2]. If there is a breakdown, then the estimation is unreliable, and thus the interpretation from section 1 is unreliable. The table in section 2 succinctly lists the various possible breakdowns and their implications for the reliability of the regression results[3].

Why is the result not acceptable unless the assumptions are met? The reason is that the strong statements inferred from a regression (i.e. - "an increase in one unit of the value of variable X causes an increase in the value of variable Y by 0.21 units") depend on the presumption that the variables used in a regression, and the residuals from the regression, satisfy certain statistical properties. These are expressed in the properties of the distribution of the residuals *(that explains why so many of the diagnostic tests shown in sections 3-4 and the corrective methods are based on the use of the residuals)*. If these properties are satisfied, then we can be confident in our interpretation of the results.

The above statements are based on complex formal mathematical proofs. Please check your textbook if you are curious about the formal foundations of the statements.

Section 3 provides a brief schema for checking for the breakdown of classical assumptions. The testing usually involves informal (graphical) and formal (distribution-based hypothesis tests like the F and T) testing, with the latter involving the running of other regressions and computing of variables.

1. Interpretation of regression results

[1] Even though interpretation precedes checking for the breakdown of classical assumptions, it is good practice to first check for the breakdown of classical assumptions (section 4), then to correct for the breakdowns, and then, finally, to interpret the results of a regression analysis.

[2] We will use the phrase "Classical Assumptions" often. Check your textbook for details about these assumptions. In simple terms, regression is a statistical method. The fact that this generic method can be used for so many different types of models and in so many different fields of study hinges on one area of commonality - the model rests on the bedrock of the solid foundations of well-established and proven statistical properties/theorems. If the specific regression model is in concordance with the certain assumptions required for the use of these properties/theorems, then the generic regression results can be inferred. The classical assumptions constitute these requirements.

[3] If you find any breakdown(s) of the classical assumptions, then you must correct for it by taking appropriate measures. Chapter 8 looks into these measures. After running the "corrected" model, you again must perform the full range of diagnostic checks for the breakdown of classical assumptions. This process will continue until you no longer have a serious breakdown problem, or the limitations of data compel you to stop.

Assume you want to run a regression of *wage* on *age*, *work experience*, *education*, *gender,* and a dummy for *sector of employment* (whether employed in the public sector).

wage = function(*age*, *work experience*, *education*, *gender*, *sector*)

or, as your textbook will have it,

$wage = \beta_1 + \beta_2 * age + \beta_3 * work\ experience + \beta_4 * education + \beta_5 * gender + \beta_6 * sector$

Always look at the model fit ("ANOVA") first. Do not make the mistake of looking at the R-square before checking the goodness of fit.

Significance of the model *("Did the model explain the deviations in the dependent variable")*
The last column shows the goodness of fit of the model. The lower this number, the better the fit. Typically, if "Sig" is greater than 0.05, we conclude that our model could not fit the data[4].

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	54514.39	5	10902.88	414.262	.000[b]
	Residual	52295.48	1987	26.319		
	Total	106809.9	1992			

a. Dependent Variable: WAGE

b. Independent Variables: (Constant), WORK_EX, EDUCATION, GENDER, PUB_SEC, AGE

[4] If Sig < .01, then the model is significant at 99%, if Sig < .05, then the model is significant at 95%, and if Sig <.1, the model is significant at 90%. Significance implies that we can accept the model. If Sig>.,1 then the model was not significant (a relationship could not be found) or "R-square is not significantly different from zero."

The F is comparing the two models below:

1. $wage = \beta_1 + \beta_2 * age + \beta_3 * work\ experience + \beta_4 * education + \beta_5 * gender + \beta_6 * sector$

2. $wage = \beta_1$

(In formal terms, the F is testiong the hypothesis: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$

If the F is not significant, then we cannot say that model 1 is any better than model 2. The implication is obvious--the use of the independent variables has not assisted in predicting the dependent variable.

Sum of squares

- The TSS (Total Sum of Squares) is the total deviations in the dependent variable. **The aim of the regression is to explain these deviations** (by finding the best betas that can minimize the sum of the squares of these deviations).

- The ESS (Explained Sum of Squares) is the amount of the TSS that could be explained by the model.

  The R-square, shown in the next table, is the ratio ESS/TSS. It captures the percent of deviation from the mean in the dependent variable that could be explained by the model.

- The RSS is the amount that could not be explained (TSS minus ESS).

In the previous table, the column "Sum of Squares" holds the values for TSS, ESS, and RSS. The row "Total" is TSS (106809.9 in the example), the row "Regression" is ESS (54514.39 in the example), and the row "Residual" contains the RSS (52295.48 in the example).

Model Summary [a,b]

Model	Variables Entered	Variables Removed	R Square	Adjusted R Square	Std. Error of the Estimate
1	*WORK_EX, EDUCATION, GENDER, PUB_SEC, AGE* [c,d]	.	*.510*	*.509*	*5.1302*

a. Dependent Variable: WAGE

b. Method: Enter

c. Independent Variables: (Constant), WORK_EX, EDUCATION, GENDER, PUB_SEC, AGE

d. All requested variables entered.

Adjusted R-square

Measures the proportion of the **variance** in the dependent variable (*wage*) that was explained by variations in the independent variables. In this example, the "**Adjusted** R-Square" shows that 50.9% of the variance was explained.

R-square

Measures the proportion of the **variation** in the dependent variable (*wage*) that was explained by variations in the independent variables. In this example, the "R-Square"' tells us that 51% of the variation (and not the variance) was explained.

Std Error of Estimate

Std error of the estimate measures the dispersion of the dependent variables estimate around its mean (in this example, the "Std. Error of the Estimate" is 5.13). Compare this to the mean of the "Predicted" values of the dependent variable. If the Std. Error is more than 10% of the mean, it is high.

The reliability of individual coefficients

The table "Coefficients" provides information on the confidence with which we can support the estimate for each such estimate (see the columns "T" and "Sig.".) If the value in "Sig." is less than 0.05, then we can assume that the estimate in column "B" can be asserted as true with a 95% level of confidence[5]. Always interpret the "Sig" value first. *If this value is more than 0 .1 then the coefficient estimate is not reliable because it has "too" much dispersion/variance.*

The individual coefficients

The table "Coefficients" provides information effect of individual variables (the "Estimated Coefficients" or "beta" --see column "B") on the dependent variable

Confidence Interval

[5] If the value is greater than 0.05 but less than 0.1, we can only assert the veracity of the value in "B" with a 90% level of confidence. If "Sig" is above 0.1, then the estimate in "B" is unreliable and is said to not be statistically significant. The confidence intervals provide a range of values within which we can assert with a 95% level of confidence that the estimated coefficient in "B" lies. For example, "The coefficient for *age* lies in the range .091 and .145 with a 95% level of confidence, while the coefficient for *gender* lies in the range -2.597 and -1.463 at a 95% level of confidence."

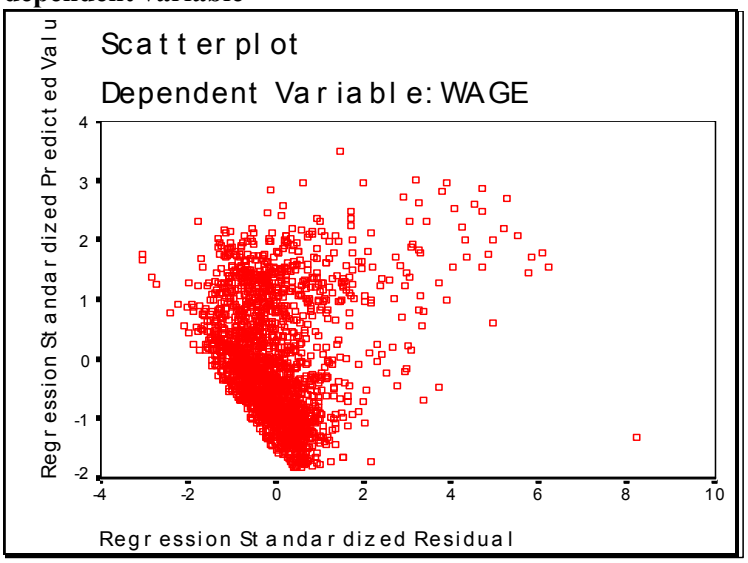
Coefficients[a]

Model		Unstandardized Coefficients		t	Sig.	95% Confidence Interval for B	
		B	Std. Error			Lower Bound	Upper Bound
1	(Constant)	-1.820	.420	-4.339	.000	-2.643	-.997
	AGE	.118	.014	8.635	.000	.091	.145
	EDUCATION	.777	.025	31.622	.000	.729	.825
	GENDER	-2.030	.289	-7.023	.000	-2.597	-1.463
	PUB_SEC	1.741	.292	5.957	.000	1.168	2.314
	WORK_EX	.100	.017	5.854	.000	.067	.134

a. Dependent Variable: WAGE

Plot of residual versus predicted dependent variable

This is the plot for the standardized predicted variable and the standardized residuals. The pattern in this plot indicates the presence of mis-specification[6] and/or heteroskedasticity[7].

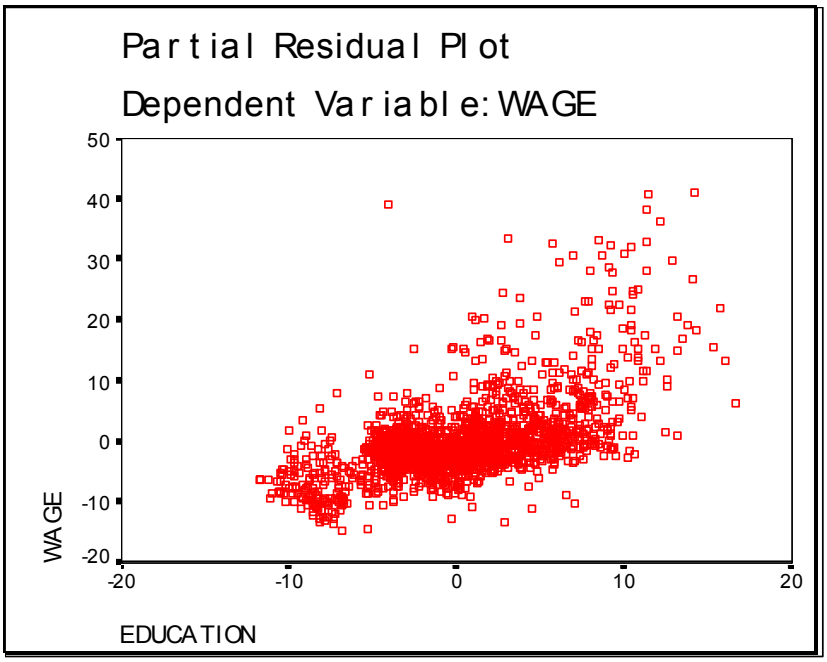


[6] This includes the problems of incorrect functional form, omitted variable, or a mis-measured independent variable. A formal test such as the RESET Test is required to conclusively prove the existence of mis-specification. Review your textbook for the step-by-step description of the RESET test.

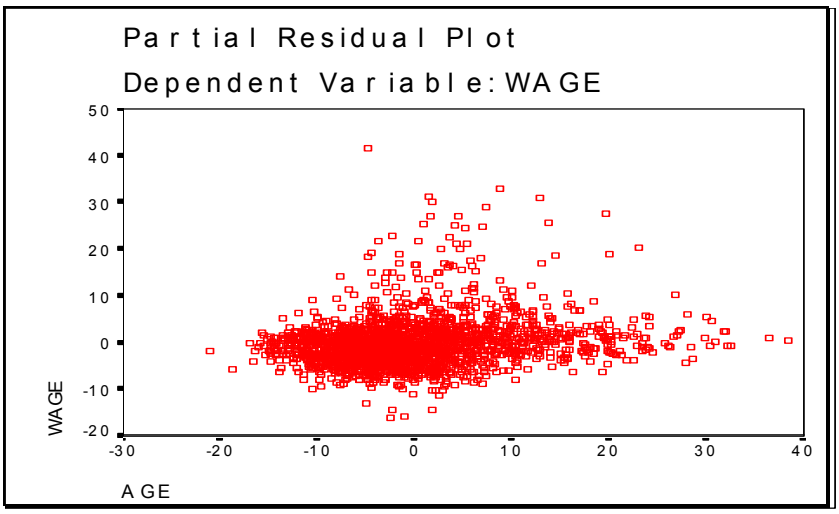
[7] A formal test like the White's Test is necessary to conclusively prove the existence of heteroskedasticity. Review your textbook for the step-by-step description of the RESET test.

Plot of residuals versus independent variables

The definite positive pattern indicates[8] the presence of heteroskedasticity caused, at least in part, by the variable education.



The plot of *age* and the residual has no pattern[9], which implies that no heteroskedasticity is caused by this variable.

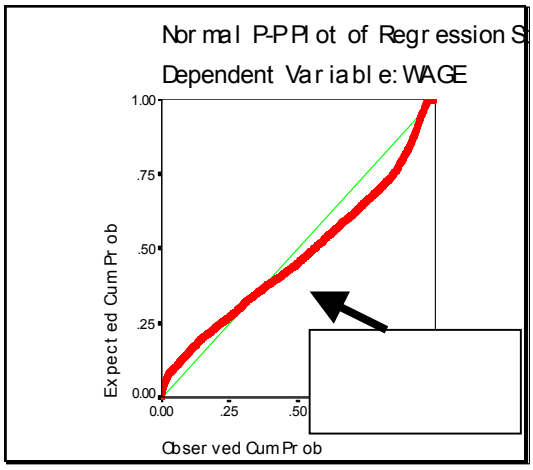


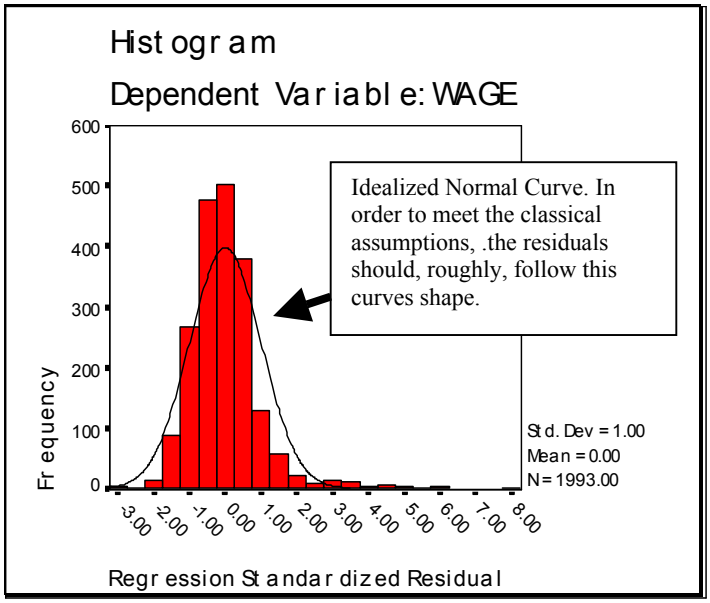
[8] A formal test like the White's Test is required to conclusively prove the existence and structure of heteroskedasticity .

[9] Sometimes these plots may not show a pattern. The reason may be the presence of extreme values that widen the scale of one or both of the axes, thereby "smoothing out" any patterns. If you suspect this has happened, as would be the case if most of the graph area were empty save for a few dots at the extreme ends of the graph, then rescale the axes using the methods. This is true for all scatter graphs.

Plots of the residuals

The histogram and the P-P plot of the residual suggest that the residual is probably normally distributed[10]. You can also use other tests to check for normality.





[10] The residuals should be distributed normally. If not, then some classical assumption has been violated.

Regression output interpretation guidelines

Name Of Statistic/ Chart	What Does It Measure Or Indicate?	Critical Values	Comment
Sig.-F	Whether the model as a whole is significant. It tests whether R-square is significantly different from zero	- below .01 for 99% confidence in the ability of the model to explain the dependent variable - below .05 for 95% confidence in the ability of the model to explain the dependent variable - below 0.1 for 90% confidence in the ability of the model to explain the dependent variable	**The first statistic to look for in SPSS output**. If Sig.-F is insignificant, then the regression as a whole has failed. No more interpretation is necessary (although some statisticians disagree on this point). You must conclude that the "Dependent variable cannot be explained by the independent/explanatory variables." The next steps could be rebuilding the model, using more data points, etc.
RSS, ESS & TSS	The main function of these values lies in calculating test statistics like the F-test, etc.	The ESS should be high compared to the TSS (the ratio equals the R-square). Note for interpreting the SPSS table, column "Sum of Squares": "Total" =TSS, "Regression" = ESS, and "Residual" = RSS	If the R-squares of two models are very similar or rounded off to zero or one, then you might prefer to use the F-test formula that uses RSS and ESS.

Name Of Statistic/ Chart	What Does It Measure Or Indicate?	Critical Values	Comment
SE of Regression	The standard error of the estimate predicted dependent variable	There is no critical value. Just compare the std. error to the mean of the predicted dependent variable. The former should be small (<10%) compared to the latter.	You may wish to comment on the SE, especially if it is too large or small relative to the mean of the predicted/estimated values of the dependent variable.
R-Square	Proportion of variation in the dependent variable that can be explained by the independent variables	Between 0 and 1. A higher value is better.	This often mis-used value should serve only as a summary measure of Goodness of Fit. Do not use it blindly as a criterion for model selection.
Adjusted R-square	Proportion of variance in the dependent variable that can be explained by the independent variables or R-square adjusted for # of independent variables	Below 1. A higher value is better	Another summary measure of Goodness of Fit. Superior to R-square because it is sensitive to the addition of irrelevant variables.
T-Ratios	The reliability of our estimate of the individual beta	Look at the p-value (in the column "Sig.") it must be low: - below .01 for 99% confidence in the value of the estimated coefficient - below .05 for 95% confidence in the value of the estimated coefficient - below .1 for 90% confidence in the value of the estimated coefficient	For a one-tailed test (at 95% confidence level), the critical value is (approximately) 1.65 for testing if the coefficient is greater than zero and (approximately) -1.65 for testing if it is below zero.

Name Of Statistic/ Chart	What Does It Measure Or Indicate?	Critical Values	Comment
Confidence Interval for beta	The 95% confidence band for each beta estimate	The upper and lower values give the 95% confidence limits for the coefficient	Any value within the confidence interval cannot be rejected (as the true value) at 95% degree of confidence
Charts: Scatter of predicted dependent variable and residual	Mis-specification and/or heteroskedasticity	There should be no discernible pattern. If there is a discernible pattern, then do the RESET and/or DW test for mis-specification or the White's test for heteroskedasticity	Extremely useful for checking for breakdowns of the classical assumptions, i.e. - for problems like mis-specification and/or heteroskedasticity. At the top of this table, we mentioned that the F-statistic is the first output to interpret. Some may argue that the ZPRED-ZRESID plot is more important (their rationale will become apparent as you read through the rest of this chapter and chapter 8).
Charts: plots of residuals against independent variables	Heteroskedasticity	There should be no discernible pattern. If there is a discernible pattern, then perform White's test to formally check.	Common in cross-sectional data. If a partial plot has a pattern, then that variable is a likely candidate for the cause of heteroskedasticity.
Charts: Histograms of residuals	Provides an idea about the distribution of the residuals	The distribution should look like a normal distribution	A good way to observe the actual behavior of our residuals and to observe any severe problem in the residuals (which would indicate a breakdown of the classical assumptions)

Problems caused by breakdown of classical assumptions

The fact that we can make bold statements on causality from a regression hinges on the classical linear model. If its assumptions are violated, then we must re-specify our analysis and begin the regression anew. It is very unsettling to realize that a large number of institutions, journals, and faculties allow this fact to be overlooked.

When using the table below, remember the ordering of the severity of an impact.

- The worst impact is a bias in the F (then the model cant be trusted)
- A second disastrous impact is a bias in the betas (the coefficient estimates are unreliable)
- Compared to the above, biases in the standard errors and T are not so harmful (these biases only affect the reliability of our confidence about the variability of an estimate, not the reliability about the value of the estimate itself)

Summary of impact of a breakdown of a classical assumption on the reliability with which regression output can be interpreted

Violation ↓ **Impact** →	**F**	**R^2**	**β**	**Std error** (of estimate)	**Std error** (of β)	**T**	**Count of violations**
Measurement error in dependent variable	👍	👍	👍	👍	**X ↑**	**X ↓**	**2**
Measurement error in independent variable	**X**	**X**	**X**	**X**	**X**	**X**	**6**
Irrelevant variable	👍	👍	👍	👍	**X ↑**	**X ↓**	**2**
Omitted variable	**X**	**X**	**X**	**X**	**X**	**X**	**6**
Incorrect functional form	**X**	**X**	**X**	**X**	**X**	**X**	**6**

Violation ↓ **Impact** →	**F**	**R^2**	**β**	**Std error** (of estimate)	**Std error** (of β)	**T**	**Count of violations**
Heteroskedasticity	**X**	👍	👍	**X**	**X**	**X**	**3**
Collinearity	👍	👍	👍	👍	**X** ↑	**X** ↓	**2**
Simultaneity Bias	**X**	**X**	**X**	**X**	**X**	**X**	**6**

Legend for understanding the table

👍 The statistic is still reliable and unbiased.

X The statistic is biased, and thus cannot be relied upon.

↑ Upward bias in estimation

↓ Downward bias in estimation.

Diagnostics

This section lists some methods of detecting for breakdowns of the classical assumptions.

Why is the result not acceptable unless the assumptions are met? The reason is simple - the strong statements inferred from a regression (e.g. - "an increase in one unit of the value of variable X causes an increase of the value of variable Y by 0.21 units") depend on the presumption that the variables used in a regression, and the residuals from that regression, satisfy certain statistical properties. These are expressed in the properties of the distribution of the residuals. *That explains why so many of the diagnostic tests shown in sections 7.4-7.5 and their relevant corrective methods, shown in this chapter, are based on the use of the residuals.* If these properties are satisfied, then we can be confident in our interpretation of the results. The above statements are based on complex, formal mathematical proofs. Please refer to your textbook if you are curious about the formal foundations of the statements.

With experience, you should develop the habit of doing the diagnostics before interpreting the model's significance, explanatory power, and the significance and estimates of the regression coefficients. If the diagnostics show the presence of a problem, you must first correct the problem and then interpret the model. Remember that the power of a regression analysis (after all, it is extremely powerful to be able to say that "data shows that X causes Y by this slope factor") is based upon the fulfillment of certain conditions that are specified in what have been dubbed the "classical" assumptions.

Refer to your textbook for a comprehensive listing of methods and their detailed descriptions.

If a formal[11] diagnostic test confirms the breakdown of an assumption, then you must attempt to correct for it. This correction usually involves running another regression on a transformed version of the original model, with the exact nature of the transformation being a function of the classical regression assumption that has been violated[12].

Collinearity[13]

Collinearity between variables is always present. A problem occurs if the degree of collinearity is high enough to bias the estimates.

Note: Collinearity means that two or more of the independent/explanatory variables in a regression have a linear relationship. This causes a problem in the interpretation of the regression results. If the variables have a close linear relationship, then the estimated regression coefficients and T-statistics may not be able to properly isolate the unique effect/role of each variable and the confidence with which we can presume these effects to be true. The close relationship of the

[11] Usually, a "formal" test uses a hypothesis testing approach. This involves the use of testing against distributions like the T, F, or Chi-Square. An "informal' test typically refers to a graphical test.

[12] Don't worry if this line confuses you at present - its meaning and relevance will become apparent as you read through this chapter.

[13] Also called Multicollinearity.

variables makes this isolation difficult. Our explanation may not satisfy a statistician, but we hope it conveys the fundamental principle of collinearity.

Summary measures for testing and detecting collinearity include:

- Running bivariate and partial correlations (see section 5.3). A bivariate or partial correlation coefficient greater than 0.8 (in absolute terms) between two variables indicates the presence of significant collinearity between them.
- Collinearity is indicated if the R-square is high (greater than 0.75[14]) and only a few T-values are significant.
- Check your textbook for more on collinearity diagnostics.

Mis-specification

Mis-specification of the regression model is the most severe problem that can befall an econometric analysis. Unfortunately, it is also the most difficult to detect and correct.

Note: Mis-specification covers a list of problems. These problems can cause moderate or severe damage to the regression analysis. Of graver importance is the fact that most of these problems are caused not by the nature of the data/issue, but by the modeling work done by the researcher. It is of the utmost importance that every researcher realise that the responsibility of correctly specifying an econometric model lies solely on them. A proper specification includes determining curvature (linear or not), functional form (whether to use logs, exponentials, or squared variables), and the accuracy of measurement of each variable, etc.

Mis-specification can be of several types: incorrect functional form, omission of a relevant independent variable, and/or measurement error in the variables. Sections 7.4.c to 7.4.f list a few summary methods for detecting mis-specification. Refer to your textbook for a comprehensive listing of methods and their detailed descriptions.

Simultaneity bias

Simultaneity bias may be seen as a type of mis-specification. This bias occurs if one or more of the independent variables is actually dependent on other variables in the equation. For example, we are using a model that claims that income can be explained by investment and education. However, we might believe that investment, in turn, is explained by income. If we were to use a simple model in which income (the dependent variable) is regressed on investment and education (the independent variables), then the specification would be incorrect because investment would not really be "independent" to the model - it is affected by income. Intuitively, this is a problem because the simultaneity implies that the residual will have some relation with the variable that has been incorrectly specified as "independent" - the residual is capturing (more in a metaphysical than formal mathematical sense) some of the unmodeled reverse relation between the "dependent" and "independent" variables.

[14] Some books advise using 0.8.

Incorrect functional form

If the correct relation between the variables is non-linear but you use a linear model and do not transform the variables, then the results will be biased.

Why should an incorrect functional form lead to severe problems? Regression is based on finding coefficients that minimize the "sum of squared residuals." Each residual is the difference between the predicted value (the regression line) of the dependent variable versus the realized value in the data. If the functional form is incorrect, then each point on the regression "line" is incorrect because the line is based on an incorrect functional form. A simple example: assume Y has a log relation with X (a log curve represents their scatter plot) but a linear relation with "Log X." If we regress Y on X (and not on "Log X"), then the estimated regression line will have a systemic tendency for a bias because we are fitting a straight line on what should be a curve. The residuals will be calculated from the incorrect "straight" line and will be wrong. If they are wrong, then the entire analysis will be biased because everything hinges on the use of the residuals.

Listed below are methods of detecting incorrect functional forms:

- Perform a preliminary visual test. Any pattern in a plot of the predicted variable and the residuals plot implies mis-specification (and/or heteroskedasticity) due to the use of an incorrect functional form or due to omission of a relevant variable.
- If the visual test indicates a problem, perform a formal diagnostic test like the RESET test or the DW test.
- Check the mathematical derivation (if any) of the model.
- Determine whether any of the scatter plots have a non-linear pattern. If so, is the pattern log, square, etc?
- The nature of the distribution of a variable may provide some indication of the transformation that should be applied to it. For example, section 3.2 showed that *wage* is non-normal but that its log is normal. This suggests re-specifying the model by using the log of *wage* instead of *wage*.
- Check your textbook for more methods.

Omitted variable

Not including a variable that actually plays a role in explaining the dependent variable can bias the regression results. Methods of detection [15] include:

- Any pattern in this plot implies mis-specification (and/or heteroskedasticity) due to the use of an incorrect functional form or due to the omission of a relevant variable.
- If the visual test indicates a problem, perform a formal diagnostic test such as the RESET test.
- Apply your intuition, previous research, hints from preliminary bivariate analysis, etc. For example, in the model we ran, we believe that there may be an omitted variable bias because of the absence of two crucial variables for wage determination - whether the labor is unionized and the professional sector of work (medicine, finance, retail, etc.).
- Check your textbook for more methods.

[15] The first three tests are similar to those for Incorrect Functional form.

Inclusion of an irrelevant variable

This mis-specification occurs when a variable that is not actually relevant to the model is included[16]. To detect the presence of irrelevant variables:

- Examine the significance of the T-statistics. If the T-statistic is not significant at the 10% level (usually if T< 1.64 in absolute terms), then the variable may be irrelevant to the model.

Measurement error

This is not a very severe problem if it only afflicts the dependent variable, but it may bias the T-statistics. Methods of detecting this problem include:

- Knowledge about problems/mistakes in data collection
- There may be a measurement error if the variable you are using is a proxy for the actual variable you intended to use. In our example, the wage variable includes the monetized values of the benefits received by the respondent. But this is a subjective monetization of respondents and is probably undervalued. As such, we can guess that there is probably some measurement error.
- Check your textbook for more methods

Measurement errors causing problems can be easily understood. Omitted variable bias is a bit more complex. Think of it this way - the deviations in the dependent variable are in reality explained by the variable that has been omitted. Because the variable has been omitted, the algorithm will, mistakenly, apportion what should have been explained by that variable to the other variables, thus creating the error(s). Remember: our explanations are too informal and probably incorrect by strict mathematical proof for use in an exam. We include them here to help you understand the problems a bit better.

Heteroskedasticity

Heteroskedasticity implies that the variances (i.e. - the dispersion around the expected mean of zero) of the residuals are not constant, but that they are different for different observations. This causes a problem: if the variances are unequal, then the relative reliability of each observation (used in the regression analysis) is unequal. The larger the variance, the lower should be the importance (or weight) attached to that observation. As you will see in section 8.2, the correction for this problem involves the downgrading in relative importance of those observations with higher variance. The problem is more apparent when the value of the variance has some relation to one or more of the independent variables. *Intuitively, this is a problem because the distribution of the residuals should have no relation with any of the variables (a basic assumption of the classical model).*

Detection involves two steps:

- Looking for patterns in the plot of the predicted dependent variable and the residual

[16] By dropping it, we improve the reliability of the T-statistics of the other variables (which are relevant to the model). But, we may be causing a far more serious problem - an omitted variable! An insignificant T is not necessarily a bad thing - it is the result of a "true" model. Trying to remove variables to obtain only significant T-statistics is bad practice.

- If the graphical inspection hints at heteroskedasticity, you must conduct a formal test like the White's test. Section 7.5 teaches you how to conduct a White's test[17].

Checking formally for heteroskedasticity: White's test

The White's test is usually used as a test for heteroskedasticity. In this test, a regression of the squares of the residuals is run on the variables suspected of causing the heteroskedasticity, their squares, and cross products.

$$(\text{residuals})^2 = b_0 + b_1\, educ + b_2\, work\_ex + b_3\,(educ)^2 + b_4\,(work\_ex)^2 + b_5\,(educ{*}work\_ex)$$

Model Summary [a]

Variables Entered	R Square	Adjusted R Square	Std. Error of the Estimate
SQ_WORK, SQ_EDUC, EDU_WORK, Work Experience, EDUCATION	*.037*	*.035*	*.2102*

a. Dependent Variable: SQ_RES

White's Test

- Calculate $n*R^2$ → $R^2 = 0.037$, n=2016 → Thus, $n*R^2 = .037*2016 = 74.6$.
- Compare this value with $\chi^2(n)$, i.e. with $\chi^2(2016)$ (χ^2 is the symbol for the Chi-Square distribution)

$\chi^2(2016) = 124$ obtained from χ^2 table. (For 955 confidence) As $n*R^2 < \chi^2$, heteroskedasticity can not be confirmed.

Note: Please refer to your textbook for further information regarding the interpretation of the White's test. If you have not encountered the Chi-Square distribution/test before, there is no need to panic! The same rules apply for testing using any distribution - the T, F, Z, or Chi-Square. First, calculate the required value from your results. Here the required value is the sample size ("n") multiplied by the R-square. You must determine whether this value is higher than that in the standard table for the relevant distribution (here the Chi-Square) at the recommended level of confidence (usually 95%) for the appropriate degrees of freedom (for the White's test, this equals the sample size "n") in the table for the distribution (which you will find in the back of most econometrics/statistics textbooks). If the former is higher, then the hypothesis is rejected. Usually the rejection implies that the test could not find a problem[18].

[17] Other tests: Park, Glejser, Goldfelt-Quandt. Refer to your text book for a comprehensive listing of methods and their detailed descriptions.

[18] We use the phraseology "Confidence Level of "95%." Many professors may frown upon this, instead preferring to use "Significance Level of 5%." Also, our explanation is simplistic. Do not use it in an exam! Instead, refer to the chapter on "Hypothesis Testing" or "Confidence Intervals" in your textbook. A clear understanding of these concepts is essential.

ANSWERS TO CONCEPTUAL QUESTIONS ON REGRESSION ANALYSIS

1. Why is the regression method you use called 'Least Squares'? Can you justify the use of such a method?
   Ans: The method minimises the squares of the residuals. The formulas for obtaining the estimates of the beta coefficients, std errors, etc. are all based on this principle.
   Yes, we can justify the use of such a method: the aim is to minimise the error in our prediction of the dependent variable, and by minimising the residuals we are doing just that. By using the "squares" we are precluding the problem of signs thereby giving positive and negative prediction errors the same importance.

2. The classical assumptions mostly hinge on the properties of the residuals. Why should it be so?

   Ans: this is linked to question 1. The estimation method is based on minimising the sum of the squared residuals. As such, all the powerful inferences we draw from the results [like R^2, betas, T, F, etc.] are based on assumed properties of the residuals. Any deviations from these assumptions can cause major problems.

3. Prior to running a regression, you have to create a model. What are the important steps and considerations in creating such a model?

   Ans: the most important consideration is the theory you want to test/support/refute using the estimation. The theory may be based on theoretical/analytical research and derivations, previous work by others, intuition, etc..
   The limitations of data may constrain the model.
   Scatter plots may provide an indication of the transformations needed.
   Correlations may tell you about the possibility of collinearity and the modeling ramifications thereof.

4. What role may correlations, scatter plots and other bivariate and multivariate analysis play in the specification of a regression model?

   Ans: prior to any regression analysis, it is essential to run some descriptives and basic bivariate and multivariate analysis. Based on the inferences from these, you may want to construct a model which can answer the questions raised by the initial analysis and/or can incorporate the insights from the initial analysis.

5. After running a regression, in what order should you interpret the results and why?

   Ans: first, check for the breakdown of classical assumptions [collinearity, heteroskedasticity, etc..]. Then, you are sure that no major problem is present, interpret the results in roughly the following order: Sig F, Adj R^2 , Std error of estimate, Sig-T, beta, Confidence interval of beta

6. In the regression results, are we confident about the coefficient estimates? If not, what additional information [statistic] are we using to capture our degree of un-confidence about the estimate we obtain?

   Ans: no, we are not confident about the estimated coefficient. The std error is being used to

capture this. The T scales the level of confidence for all variables on to a uniform scale.

7. Each estimated coefficient has a distribution of its own. What kind of distribution does each beta have? In this distribution, where do the estimate and the std error fit in?

Ans: Each beta is distributed as a T-distribution with mean equal to the estimated beta coefficient and std error equal to the estimated std error of the estimated beta.

8. How is a T-statistic related to what you learned about Z-scores? Why cannot the z-score be used instead of the T-statistic?

Ans: the T is used because we have an estimated std error and not the actual std error. If we had the actual std error for a beta, then we could use the z-score.

9. In the regression results, what possible problems does a pattern in a plot of the predicted values and the residuals tell you about?

Ans: omitted variable, incorrect functional form or/and heteroskedasticity.

10. In the regression results, what possible problem does a pattern in a plot of an independent variable and the residuals tell you about?

Ans: heteroskedasticity and the possible transformation needed to correct for it.

11. Intuitively, why should a scatter plot of the residuals and a variable X indicate the presence of heteroskedasticity if there is a pattern in the scatter plot?

Ans: The residuals are assumed to be randomly distributed with a zero mean and constant variance. If you have heteroskedasticity, but are still assuming a zero mean, the implication is that "as the values of X increase, the residual changes in a certain manner'. This implies that the variance of the residuals is changing.

12. In plain English, what are the numerator and denominator of the F-statistic in terms of the sum of squares?

Ans: the numerator measures the increase in the explanatory power of the model as a result of adding more variables. The denominator measures the percentage of deviation in Y that cannot be explained by the model. So the F gives you "the impact of additional variables on explaining the unexplained deviations in Y".

13. In a F-test, the degrees of freedom of the numerator equals the number of restrictions. Intuitively, why should it be so?

Ans: The number of restrictions equals the difference in number of explanatory variables across the two models being compared. The degrees of freedom reflect the fact that the difference in the explanatory power is what is being captured by the numerator.

14. The classical linear breakdown problems are:
    - heteroskedasticity,
    - collinearity,
    - autocorrelation,

- measurement errors,
- misspecification of functional form,
- omitted variable and irrelevant variable.

What problems does each breakdown cause?

Ans: The problems are biases in the estimates of :

- the beta estimates: omitted variable, incorrect functional form, measurement error in an independent variable
- the std errors of the betas: all
- the T-statistics: all
- the std error of the estimate: all except heteroskedasticity
- the Sig-F-statistic: except heteroskedasticity, collinearity, irrelevant variable

FLOW DIAGRAM FOR REGRESSION ANALYSIS