Multiple Regression:

Diagnostics
and
Solutions

Diagnostics for Regression Assumptions

I don't believe diagnostics are included on the PhD qualifying exam but knowledge of them will be useful to you whenever you want to use regression in the future.

Regression Diagnostics

There are a variety of statistical procedures that can be performed to determine whether the regression assumptions have been met. This is the purpose of diagnostics.

1. The first assumption was that the shape of the distribution of the continuous variables in the multiple regression correspond to a normal distribution.

That is, each variable's frequency distribution of values roughly approximates a bell-shaped curve.

There appears to be more controversy regarding this assumption than the others.

Many statisticians believe that only the "error term" needs to be normally distributed unless the sample is extremely small (below 100).

Violations to the assumption of normality (particularly with a small sample) can be identified by examining the skewness and kurtosis of the variables.

Positive skewness: a distribution's mean lies on the right side of the distribution
Negative skewness: a distribution's mean lies on the left side of the distribution
Positive kurtosis: there is an extreme peak in the center of the distribution (also called leptokurtosis)
Negative kurtosis: there is an extremely flat distribution (also called platykurtosis)

One approach for checking univariate normality is to examine histograms or stem-and-leaf plots for each variable.

However, this is imprecise and only provides an indication.

SPSS provides numerical measures of skewness and kurtosis. The closer to zero the more normal the variable (analyze, descriptive statistics, frequencies, statistics)

These can also be checked by examining the Shapiro-Wilk test (uses a .001 level to indicate significant normality violation) and the Kolmogorov-Smirnov test.

Solutions for non-normally distributed data from a small sample

Data transformation–SPSS changes every value of the non-normal variable or variables. The specific transformation to use varies but the log transformation is common.

Set the alpha level smaller (say, at .01 instead of .05) due to the rough approximation of the b coefficient resulting from the lack of normality

2. A second assumption is that the dependent variable is a linear function of the independent variables and random disturbance or error (E).

$$Y = a + bX_1 + bX_2 + E$$

Diagnostics for the linearity Assumption

One diagnostic for this assumption is to examine a scatterplot between the dependent variable and independent variable. Look to see if the relationship is a linear one.

Diagnostics for the linearity Assumption

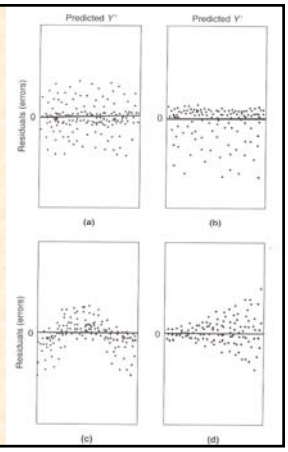
Another is to examine a scatterplot that shows the relationship between the predicted DV scores and the disturbance scores.

The disturbance scores (also referred to as the residuals or errors) are the differences between the obtained and predicted dependent variable (DV) scores. Each score reflects the distance from an actual value to the predicted value on the regression line.

SPSS can calculate and save the disturbance score for each case and can calculate and save the predicted value for each case. These can then be examined in a scatterplot to evaluate linearity (regression, linear, plot, move "zpred" into Y box and "zresid" into X box).

Diagnostics for the Linearity Assumption

Plots of predicted values of the DV against residuals.
(a) shows the assumptions met including linearity
(c) shows a failure of linearity.
(b) demonstrates a lack of normality; and
(d) shows heteroscedasticity.



Possible Solutions: How can a linear model represent nonlinear relationships?

We say the model is linear but more precisely, it is linear in the b coefficients.

We take each of the b's, multiply it by some number, and then add the results together.

$$Y = a + b_1X_1 + b_2X_2 + E$$

The model does not have to be linear in the X's.

We can make any mathematical transformation of the X's without causing any serious difficulty—except for interpretation.

Therefore, possible Solutions to violations of the linearity assumption

Use a data transformation—for a curvilinear relationship (one bend), square the variable

For curvilinear relationships with more bends use additional data transformations

3. The third assumption was that the independent variables are unrelated to the random disturbance E.

$Y = a + bX_1 + bX_2 + E$

How can we determine if the independent variables are related to E?

Unfortunately, there are no easy ways to determine if this assumption has been violated or to what extent, other than to rely on one's knowledge of the phenomenon under study.

For example, if the researcher becomes aware of an independent variable, found in the literature to be important but not included in her/his regression model, then this variable should be added to the regression equation assuming it has been measured and is available.

If the data are experimental, there is much less opportunity for this assumption to be violated.

For example, the dependent variable will not affect the independent variable since, in an experimental design, it is measured prior to the treatment (independent variable) being introduced.

Possible Solutions to violations of the assumption of no relationship between the independent variables and U.

One solution for "omitted variables" is to add any relevant variables that have not been included in the regression equation. Of course, this requires becoming aware that a variable has been omitted and having a measure of it available.

A solution for "reverse causation" (i.e., the X's affect Y and Y affects one or more variables in the error term) is to use simultaneous equations methods used by economists. However, to use these methods requires meeting additional assumptions.

A solution for "measurement error" (i.e., the X's error is included in the error term so that X is related to E) is to obtain different indicators of the variable/concept of interest that have less measurement error.

In sum, if there is no measure for one or more variables included in the error term, there are no easy solutions for violation of this assumption.

Solutions tend to require either additional data and/or more complex methods of analysis.

4. Homoscedasticity was a fourth assumption of regression analysis.

Homoscedasticity suggests that the dependent variable has an equal level of variability for each of the values of the independent variables.

A picture helps to understand this:

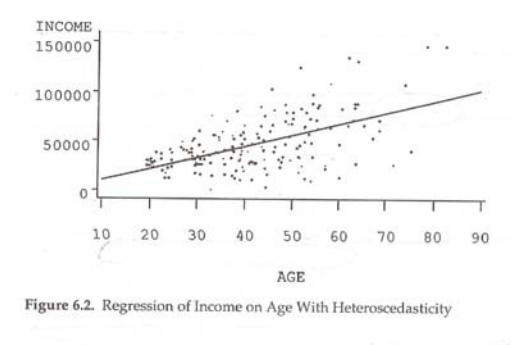


Figure 6.2. Regression of Income on Age With Heteroscedasticity

Here is a lack of homoscedasticity (referred to as heteroscedasticity)

Diagnostics for the homoscedasticity Assumption

Unlike violations to the assumption of no relationship between the IVs and E (third assumption), it is fairly easy to identify violations to homoscedasticity.

Diagnostics for the homoscedasticity Assumption

For bivariate relationships, examine scatterplots and look for variation within each value of the independent variable (as previously shown).

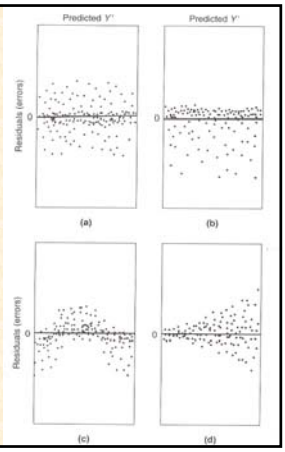
Diagnostics for the homoscedasticity Assumption

Similarly, for multiple regression comparable scatterplots can be produced by plotting the observed value Y on the vertical axis and the predicted values on the horizontal axis. These plots can reveal heteroscedasticity.

Diagnostics for the homoscedasticity Assumption

Plots of predicted values of the DV against residuals showing assumptions of homoscedasticity met (a) and assumption of homoscedasticity not met (d).

(regression, linear, plot, move "zpred" into Y box and "zresid" into X box).



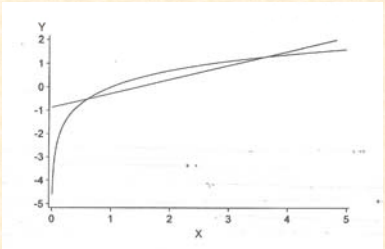
Diagnostics for the homoscedasticity Assumption

Another method is to do a Levine's test, rejection of the null hypothesis indicates that the assumption was violated

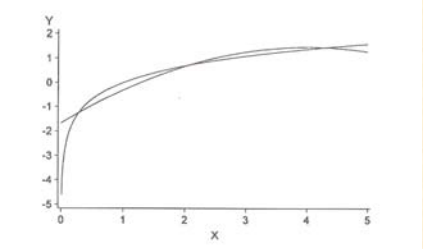
Possible Solutions to violations of homoscedasticity assumption

One solution is to do a data transformation —calculate the natural logarithm of the value to be predicted. A better solution may be to use a square root transformation, and in other cases "weighted least squares".

The quadratic polynomial appears to be the best choice. Here, we compare a linear to a polynomial approximation of a logarithmic function. Below is the linear approximation. The R^2 is .77.

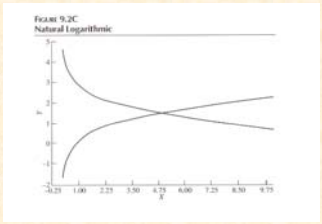


Here is a quadratic approximation to a logarithmic function. The R^2 is .91.



A second possibility to a logarithmic transformation of the independent variable.

Logarithmic transformations are infrequently used.



In these logarithmic functions, successively larger values of X predict increasingly smaller changes in Y. The rapidity with which the curve flattens out depends on the size of b.

Why use a polynomial when we could use the logarithmic transformation?

1. The polynomial can approximate a wider variety of different curves than the logarithmic transformation.

2. Polynomials can be used with interval level data such as indices whereas the logarithmic transformation requires ratio variables that have an absolute zero.

Still another solution is to use "robust standard errors" that are available as an option in some regression applications. According to one statistician (Allison), this procedure doesn't solve for inefficiency but does give reasonably accurate p values.

5. A fifth assumption is that the disturbances are uncorrelated with one another.

If two cases in the data set are in some way related to one another then their error terms will also be related and the assumption will not be met.

For example:

In our study of nursing homes, we surveyed nurse aides working in 11 NHs. Those NAs working in the same NH are more likely to have unmeasured factors in common (e.g., management style). To the extent that this is true, the assumption of uncorrelated disturbance was not met.

Diagnostics for the assumption of uncorrelated disturbances

There are no convenient ways of diagnosing the correlation of the disturbances.

One approach is to calculate the residuals for all respondents and then examine correlations between the residuals of suspected groups of respondents.

For example, in our study of NHs, we could first calculate the residuals from the regression equation for all nurse aides. Then we would compute the correlations between the residuals for groups of NAs based on which NH they worked in. We would look for any substantive correlations.

Another diagnostic approach, for more general forms of clustering, is the examination of the intra-class correlation coefficient, but there are few statistical packages that will calculate it without special programming.

More generally, the issue of correlated disturbances is strongly affected by the sampling design.

If we have a simple random sample from a large population, it's unlikely that correlated disturbances will be a problem.

On the other hand, if the sampling method involves any kind of clustering, where people are chosen in groups rather than as individuals, the possibility of correlated disturbances should be seriously considered.

Possible Solutions to violations of uncorrelated disturbances assumption

If the clusters of cases are recognizable, then the clusters can be included within the regression equation as controls.

For example, when using the nursing home data, each NH can be included within the regression to control for correlated disturbances.

Other more sophisticated statistics are also available.

6. A sixth assumption is that the error terms are normally distributed

We assume that the shape of the distribution of the disturbance term, E, is a normal distribution (e.g., a bell shaped curve).

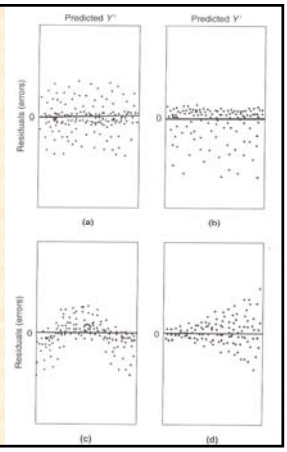
While many statisticians argue that the independent and dependent variables don't need to be normally distributed, they all appear to agree that the error term must be normally distributed.

Diagnostics for the Normality Assumption

An examination of the residuals around the regression (prediction) line can show the extent of departure from normality.

Diagnostics for the Normality Assumption

Plots of predicted values of the DV against residuals showing assumptions of normality met (a) and assumptions of normality not met (b).



Solutions: fortunately, if we have a probability sample that is sufficiently large this assumption will be met.

The Central Limit Theorum proves for us that a sufficiently large "probability sample" will result in a normal distribution of error terms.

7. A seventh assumption is a lack of multicollinearity

How can we detect multicollinearity?

1. examine zero-order correlations for any independent variables that are highly correlated (.80 or higher is considered highly susceptable to multicollinearity).

2. Regress each independent variable on the remaining independent variables. If the R^2 is .80 or higher, there is a good chance of multicollinearity.

3. Remove or add an independent variable to the regression equation and examine whether it causes one or more of the b coefficients to change drastically.

4. Examine the tolerance and the variance inflation factor.

Tolerance is the amount of a predictor's variance not accounted for by the other predictors. Lower tolerance indicates a stronger relationship (increasing the chances of obtaining multicollinearity).

Tolerance

Tolerance is calculated by:

$1 - R^2$

for each independent variable.

A tolerance below .40 indicates a possible multicollinearity problem.

Variance Inflation Factor

VIF is calculated by:

1 / tolerance

for each independent variable.

A VIF greater than 2.50 indicates a possible multicollinearity problem.

Where is Multicollinearity most likely to occur?

time-series data

panel data

Solutions for Multicollinearity when the independent variables are measuring similar concepts:

1. Delete one or more variables from the regression model.

2. Combine the collinear variables into an index.

3. Estimate a latent variable model (e.g., structural equation modeling)

Solutions for Multicollinearity when the independent variables are distinct concepts

4. Increase the number of cases in hopes of obtaining greater variability in the effected variables

5. Stratify the data in hopes that those in the specific strata show more variability in the effected variables

8. The eighth assumption was a lack of outliers

How can we detect outliers?

Diagnostics for Outliers

- box and whisker plots
- scatterplots of each variable
- examine each case's Mahalanobis distance—cases that are significant can be considered outliers
- convert all scores to Z scores and then save the Z scores and observe them. Any score greater than 2.5 is a candidate for an outlier.

Solutions of Handling Outliers

Outliers reflect mixed opportunities.

On the one hand, they may be new and exciting patterns within a data set.

On the other, they may signal anomalies within the data that may need to be addressed before any analyses are done.

Solutions When Detecting an Outlier

1. One strategy: run the desired regression analysis with and without the outlier(s). If there is no effect of the outlier then, typically, don't remove it.

2. If there is an effect of the outlier(s), consider removal of it/them. The greater the effect of a single case, the greater the reason for removing it since no single case should have considerable effect on the results.

Are there other assumptions beyond these five to consider?

For the qualifying exam, no. However, there are several additional conditions that you should be aware of. These are important in order for regression to run well and include multicollinearity and the awareness of outliers.

As noted in a previous lesson, multicollinearity exists when independent variables are correlated.

A high level of multicollinearity creates biased estimates between the variables involved.

Outliers are cases that have very extreme or unusual values.

Outliers can be caused by coding errors, extraordinary circumstances for a specific case, or may perhaps reflect an emerging pattern.

Outliers can result in a single or few cases having too much impact on the regression solution relative to the other cases.

The End.