Multiple Regression:

Assumptions

PEARLS BEFORE SWINE
BY STEPHAN PASTIS
WHAT ARE YOU DOING, RAT?
I'M PORING OVER PHYSICS BOOKS IN AN ATTEMPT TO UNIFY GENERAL RELATIVITY AND QUANTUM MECHANICS INTO ONE UNIFIED THEORY THAT GOVERNS OUR ENTIRE EXISTENCE.
WHAT HAVE YOU GOT SO FAR?
THIS.
Beer is good.
I DIDN'T KNOW IT WAS THAT SIMPLE.

Regression assumptions clarify the conditions under which multiple regression works well, ideally with unbiased and efficient estimates.

So, what do we mean by "regression assumptions"?

When we calculate a regression equation, we are attempting to use the independent variables (the *X*'s) to predict what the dependent variable (the Y) will be.

In the process of calculating the regression equation, we assume that certain conditions exist with regard to the data we are using.

These are the regression assumptions.

Perhaps a more technical definition is:

the assumptions made regarding how predicted values of Y are produced from the values of the *X*'s.

When the assumptions are met, we are more likely to have unbiased and efficient estimates.

Unbiased estimates are those that have no systematic tendency to be unreliable (either systematically too high or too low).

A biased estimate might consistently predict the estimate to be higher or lower than it actually is.

Efficient estimates have to do with how much variation there is around the true value (e.g., the standard error).

Efficient estimates have standard errors that are as small as possible.

Regression analysis is "robust" in that it will typically provide estimates that are reasonably unbiased and efficient even when one or more of the assumptions is not completely met.

However, a large violation of one or more assumptions will result in poor estimates and, consequently, the wrong conclusions being drawn.

The number of assumptions that should be considered varies from statistician to statistician.

This is because some "needed conditions" are treated as "assumptions" by some and not by others.

For example, "no multicollinearity" is described as an assumption by some and not by others, nevertheless, it is clearly a "needed condition" in order to interpret the individual effects of the independent variables.

We'll exam the most frequently cited assumptions.

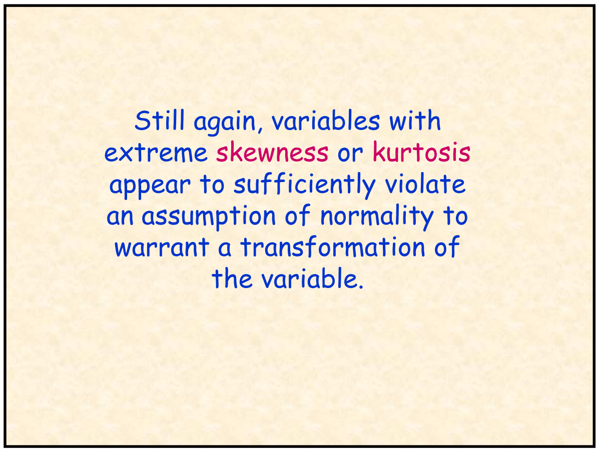
One assumption made by some statisticians is that the shape of the distribution of the continuous variables in the multiple regression correspond to a normal distribution.

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

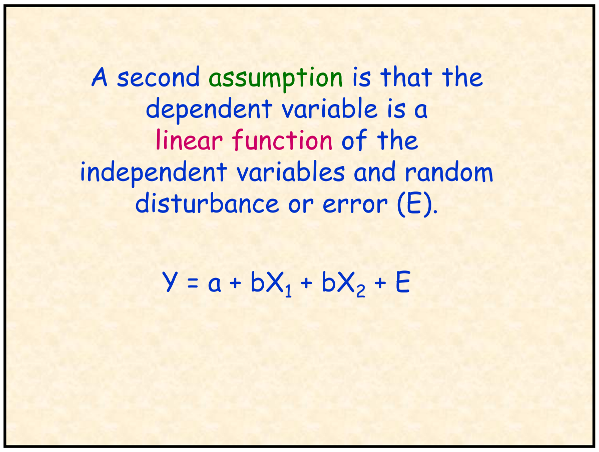
On the other hand, many statisticians explain that normalcy is <u>only</u> required of the error term in the regression equation.

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

And, if the sample is randomly selected and sufficiently large (e.g., at least 120 cases), the Central Limit Theorem shows that the error terms will be normal

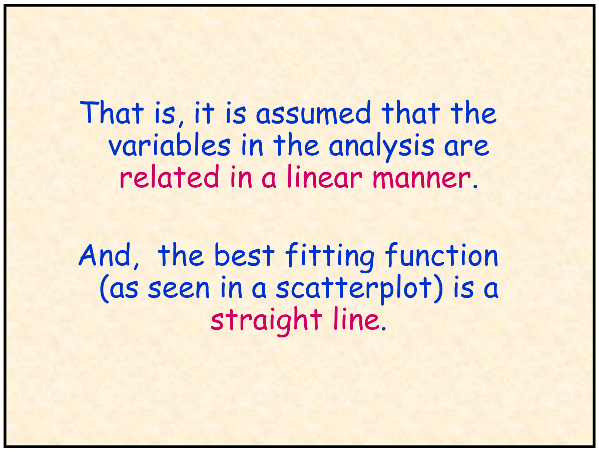


Still again, variables with extreme skewness or kurtosis appear to sufficiently violate an assumption of normality to warrant a transformation of the variable.



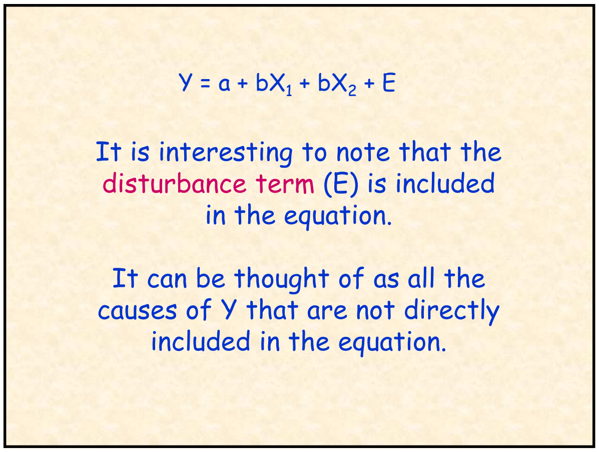
A second assumption is that the dependent variable is a linear function of the independent variables and random disturbance or error (E).

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



That is, it is assumed that the variables in the analysis are related in a linear manner.

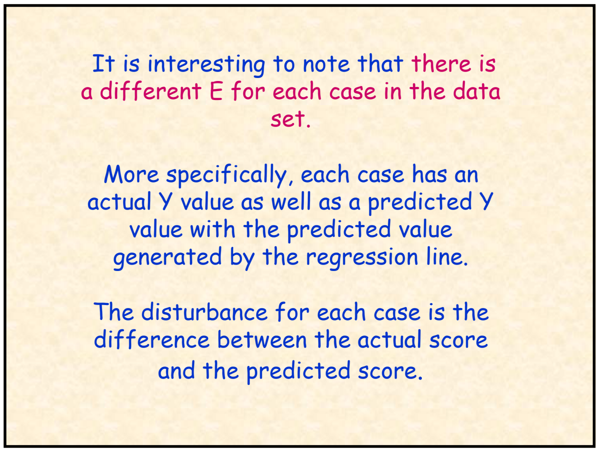
And, the best fitting function (as seen in a scatterplot) is a straight line.



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

It is interesting to note that the disturbance term (E) is included in the equation.

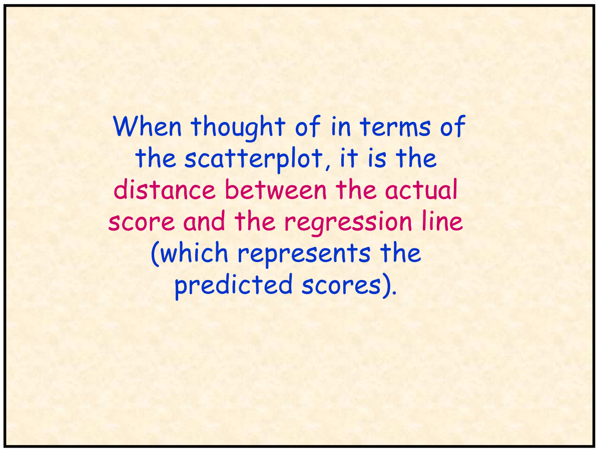
It can be thought of as all the causes of Y that are not directly included in the equation.



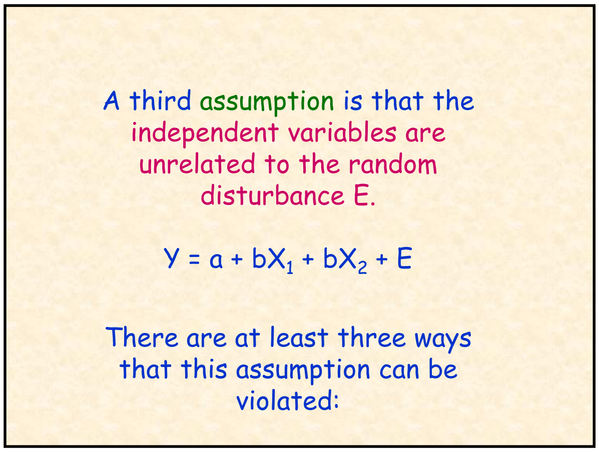
It is interesting to note that there is a different E for each case in the data set.

More specifically, each case has an actual Y value as well as a predicted Y value with the predicted value generated by the regression line.

The disturbance for each case is the difference between the actual score and the predicted score.



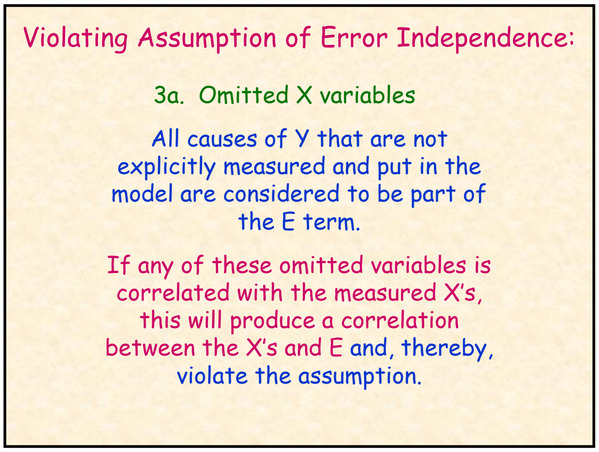
When thought of in terms of the scatterplot, it is the distance between the actual score and the regression line (which represents the predicted scores).



A third assumption is that the independent variables are unrelated to the random disturbance E.

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

There are at least three ways that this assumption can be violated:

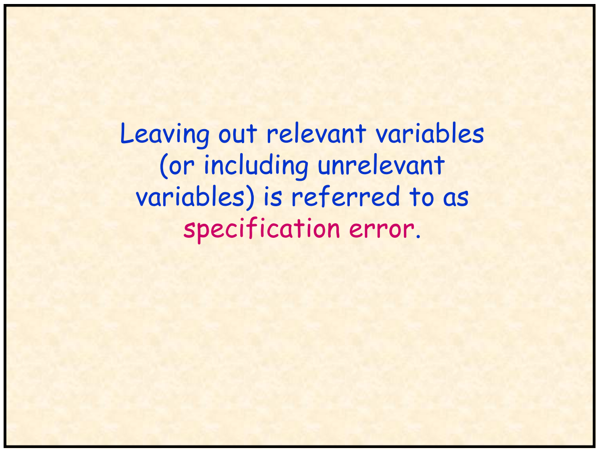


Violating Assumption of Error Independence:

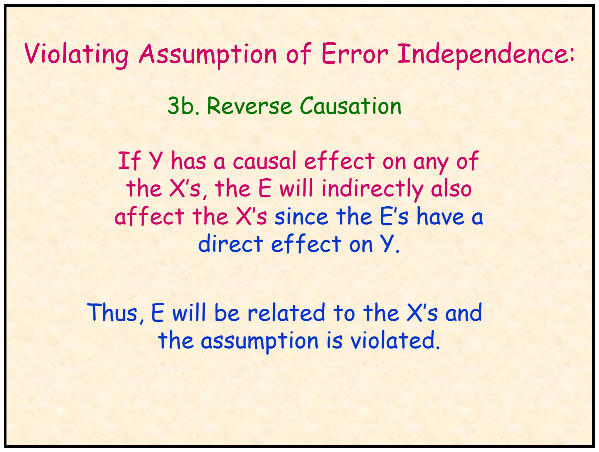
3a. Omitted X variables

All causes of Y that are not explicitly measured and put in the model are considered to be part of the E term.

If any of these omitted variables is correlated with the measured X's, this will produce a correlation between the X's and E and, thereby, violate the assumption.



Leaving out relevant variables (or including unrelevant variables) is referred to as specification error.

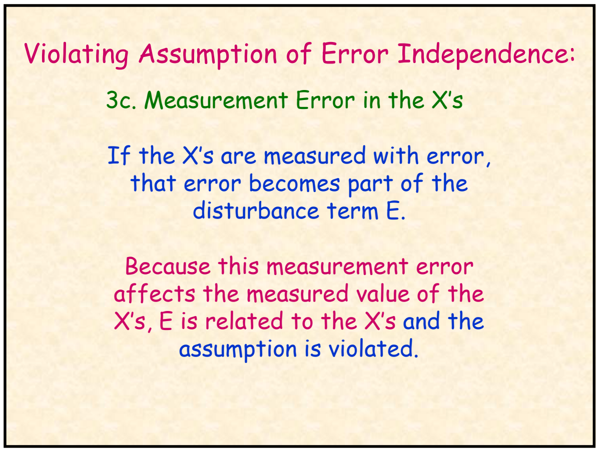


Violating Assumption of Error Independence:

3b. Reverse Causation

If Y has a causal effect on any of the X's, the E will indirectly also affect the X's since the E's have a direct effect on Y.

Thus, E will be related to the X's and the assumption is violated.

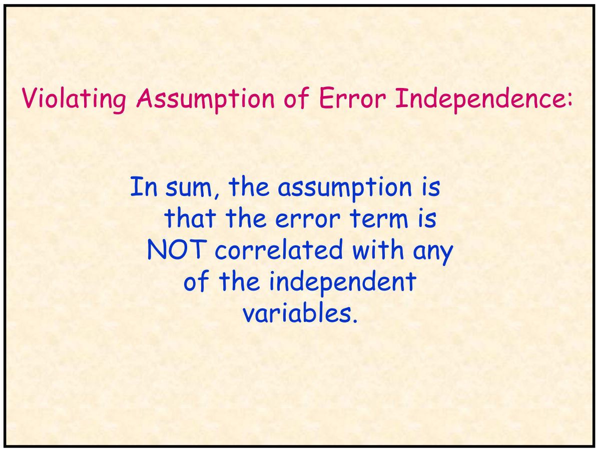


Violating Assumption of Error Independence:

3c. Measurement Error in the X's

If the X's are measured with error, that error becomes part of the disturbance term E.

Because this measurement error affects the measured value of the X's, E is related to the X's and the assumption is violated.



Violating Assumption of Error Independence:

In sum, the assumption is that the error term is NOT correlated with any of the independent variables.

Homoscedasticity is 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:

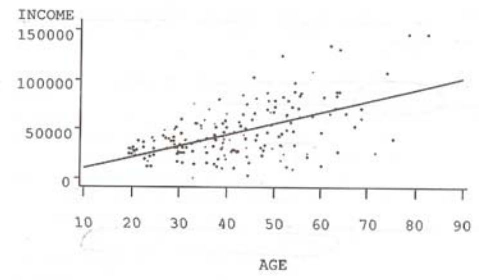


Figure 6.2. Regression of Income on Age With Heteroscedasticity

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

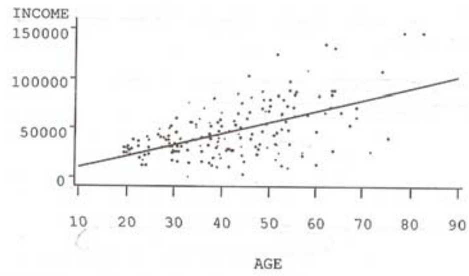


Figure 6.2. Regression of Income on Age With Heteroscedasticity

Notice that the variance of the disturbance terms is small at age 20 but is very large at the oldest ages—the variance is not equal across the values of age.

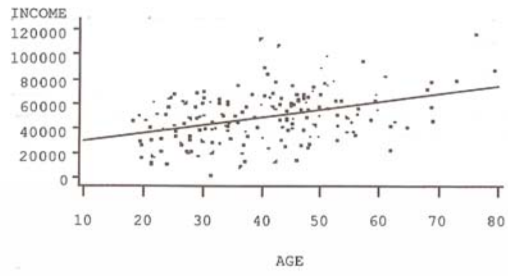


Figure 6.1. Regression of Income on Age With Homoscedasticity

Here the assumption of homoscedasticity is met—the variances at each value of age is close to being equal.

Homoscedasticity produces efficient estimates.

It's also worth noting that when the error terms vary depending on the value of X (i.e., heteroscedasticity exists), then the error term is related to X. This violates the assumption of error independence reviewed earlier.

A fifth assumption is that the disturbances of one case are uncorrelated with those of another case.

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

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.

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.

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).

This should not be confused with the X's or Y variables.

While there is a lack of agreement on the need for X's to be normally distributed, all statisticians agree that the error term must be normally distributed.

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.

A seventh condition of the data that is often referred to as an assumption of regression analysis is the lack of collinearity or of multicollinearity among the independent variables

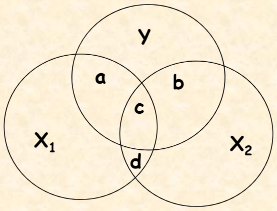
A high level of collinearity or of multicollinearity creates biased estimates between the variables involved.

Collinearity exists when two predictors correlate very strongly.

Multicollinearity exists when more than two predictors correlate very strongly.

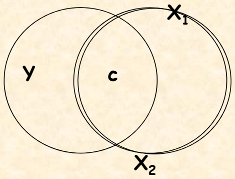
Collinearity

If collinearity between two independent variables is relatively small then the b coefficients will be minimally effected.



Collinearity

If collinearity between two independent variables is extreme then the b coefficients will be drastically effected or not calculated at all.

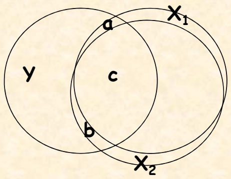


Collinearity

If collinearity between two independent variables is "near-extreme" then the b coefficients will be calculated but can be drastically effected with the effect potentially going unnoticed.

X_1 = a+c

X_2 = b+c

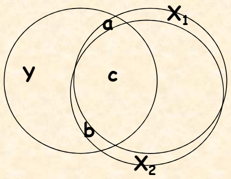


Collinearity

The squared semi-partial correlation should help in recognizing this problem.

X_1 = a+c

X_2 = b+c



Three additional effects of Multicollinearity:

1. One of the independent variables will be assigned all the variance associated with Y and the second independent variable will be treated as if it has none.

2. None of the variables will be treated as having an effect because all the explained variance is "held constant" by the other independent variable(s).

Other effects of Multicollinearity:

3. The software will not calculate a b coefficient for one of the independent variables.

A final, eighth, "assumption" to be considered here is: there is a lack of outliers.

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 Final Word on Assumptions:

Remember, regression analysis is "robust" in that it will typically provide estimates that are reasonably unbiased and efficient even when one or more of the assumptions is not completely met.

The End.

PEARLS BEFORE SWINE
BY STEPHAN PASTIS
WHAT ARE YOU DOING, RAT?
I'M PORING OVER PHYSICS BOOKS IN AN ATTEMPT TO UNIFY GENERAL RELATIVITY AND QUANTUM MECHANICS INTO ONE UNIFIED THEORY THAT GOVERNS OUR ENTIRE EXISTENCE.
WHAT HAVE YOU GOT SO FAR?
THIS.
Beer is good.
I DIDN'T KNOW IT WAS THAT SIMPLE.