Since the objective when conducting regression analysis is to draw inferences about the true β-coefficients from the $\widehat{\beta}$ estimates we need to make certain assumptions about the way in which Y_i, X_i, and u_i are generated. The reason for this is because Y_i depends on both X_i and u_i, so if we are not specific in how these terms are created it is difficult to make inferences. Thus, OLS (and regression analysis generally) rests on certain key assumptions, which are critical for the valid interpretation of estimates.

I. OLS Assumptions:

Below are the assumptions that OLS rests on (as taught by Wooldridge and Bartels):

- **MLR.1: Linear in the Parameters** - $E(Y|X_i) = \beta_0 + \beta_1 X_i$, this means that the parameter estimates $(\widehat{\beta})$ are linear , i.e., raised to the first power only. This may or may not be true in the X's
- **MLR.2: Random Sample -** we assume the data is drawn from a random sample of the population
- **MLR.3: No Perfect Collinearity -** none of the variables correlate perfectly with each other. Perfect collinearity would produce an undefined term - i.e., a denominator of zero.
- **MLR.4: Zero Conditional Mean -** $E(u_i|X_{i1}...X_{ik}) = 0$. If the independent variables are correlated with the error term then they are endogenous which would violate this assumption. This would mean that there is reciprocal causation between the independent variable(s) and the dependent variable, $X \leftrightarrow Y$. Independent variables that are uncorrelated with the error term are called exogenous.
- Up to this point, **MLR.1 - MLR.4** establish the ***formal statements of unbiasedness***. That being:
  - $E(Y|X_i) = \beta_0 + \beta_1 X_i$
  - $E(u_i|X_{i1}...X_{ik}) = 0$
  - However, we have only created a good unbiased estimator.
- The addition of **MLR.5** will produce the **best unbiased estimator**. That is:
- **MLR.5: Homoskedasticity -** The error variance is assumed to be uniform for all observations. Formally: $Var(u_i|X_{i1}...X_{ik}) = \sigma^2$
  - Note: The inclusion of **MLR.5** means that the variance of Y_i does ***not*** depends on the values of the independent variables. It should be noted that **violating homoskedasticity does not invalidate the regression model** - it simply weakens it. Therefore, we can still produce a good estimator even though MLR.5 has been violated.

  - The inclusion of **MLR.5** allows us to produce the ***Best Linear Unbiased Estimator (BLUE)***. That is finding the one and only regression line that minimizes the sum of squared residuals.
- **MLR.6: Normality of the error term -** $u_i \sim N(0,\ \sigma^2)$ we assume that the stochastic term (the errors) are **normally distributed with mean zero and a variance of** σ^2.
  - As mentioned above, **MLR.6** does not determine whether the estimates of $\widehat{\beta}$ are good or best. MLR.6 is necessary for deriving statistical inferences from the data.

Although these assumptions are all well and good, often times these assumptions are violated. The question then becomes: if any of these assumptions are violated can we still derive good (or potentially best) linear unbiased estimators? Well, yes and no. In the coming sections I will discuss certain violations to these assumptions, what it means for the estimates, and how (if at all possible) to deal statistically with these issues.

II. Multicollinearity: What Happens if the Explanatory Variables are Correlated?[1]

OLS assumes there is **no perfect collinearity** because: if multicollinearity is perfect, then the beta-coefficients are indeterminate and their standard errors are infinite. If multicollinearity is less than perfect, the beta-coefficients can be determined but the standard errors are very large, which lowers the ability to produce precise or accurate estimates.

To summarize the estimation problem:

$\widehat{\beta}_1....\widehat{\beta}_k$ gives the $\frac{\partial y}{\partial x}$ in $\overline{Y}$ as each $\widehat{\beta}$ changes by one unit holding the rest constant. If x_1 and x_2 are perfectly collinear x_2 cannot be held constant. As x_1 changes so does x_2, we therefore cannot estimate the effects of x_1 independently of x_2.

There are several **sources** of multicollinearity:

- The data collection method
- Constraints on the model or in the population being sampled
  - In other words biased data or heavily skewed data
- Model specification
- Overdetermined model
  - Garbage Can Regression
- Trending data - variables that increase or decrease in tandem.
  - For example GDP and personal income.

[1] http://psychologicalstatistics.blogspot.com/2013/11/multicollinearity-and-collinearity-in.html
https://www3.nd.edu/~rwilliam/stats2/l11.pdf

There are consequences of multicollinearity w.r.t. how it affects models:

- With high collinearity we encounter:
    - OLS estimators have **large variances** and **covariances**, making precise estimation difficult
    - Confidence intervals tend to be much wider
    - t-statistics of one or more of coefficients tend to be statistically insignificant
    - R^2 can be very high
    - OLS estimators and standard errors are sensitive to small changes in the data

Now that we know that multicollinearity can do, the question is how do we go about checking for it? There is no unique way to check for it, but there are some rules of thumb:

1. **High R^2 but few significant t-values.** If R^2 is high (generally in excess of 0.8), the F-test in most cases will be statistically insignificant (meaning that partial slope coefficients are not simultaneously equal to zero). Although a sensible approach, the problem is "it is too strong in the sense that multicollinearity is considered as harmful only when all of the influences of the explanatory variables on Y cannot be disentangled" (Kmenta 1986).
2. **High pairwise correlation among regressors.** Generally, when the correlation coefficient between any two regressors is large, then collinearity is a problem.[2] The problem with this rule however, is that high correlations may indicate collinearity but it is not necessary that they be high to have collinearity in any specific case. In other words, collinearity can exist with smaller correlation coefficient values. Basically, in multivariate regression models, the simple zero-order correlation will not provide perfect information because there are just too many variables. Alternatives to this are:
    a. Subset F-tests - look the relation between specific variables
        i. Problem here is cannot uncover multicollinearity between more than two variables
3. **Examination of partial correlations.** Farrar and Glauber (1967) suggested that one should look at the partial correlation coefficient instead of the zero-order correlations. Don't follow this - it's a thing but don't do it.
4. **Auxiliary regressions.** Regress each X on the other X variables and then look at the R^2. Klein's rule of thumb (1962) suggest that if the R^2 of one of these auxiliary regressions is greater than the regression of Y on the explanatory variables then multicollinearity is problem.
5. **Eigenvalues and condition index.** We can compute the Eigenvalues and then diagnose multicollinearity using the following equation:

[2] Textbooks seem to differ as to what is considered high collinearity between regressors. Most seem to hold by 0.8, however after conversations with some prominent methodologists, I've decided on 0.7 as a good threshold. Again, keep in mind this is just a rule of thumb and not a sacrosanct decree.

$$k \ = \ \frac{Maximum\ eigenvalue}{Minimum\ Eigenvalue}$$

One we have this we can then compute the condition index with is defined as:

$$Conditional\ Index \ = \ \sqrt{k}$$

$$\text{where, } k \ = \ \frac{Maximum\ eigenvalue}{Minimum\ Eigenvalue}$$

We then have a rule of thumb: if the CI is between 10 and 30 there is moderate to strong collinearity and if it exceeds 30 there is severe collinearity.

6. **Scatterplot.** Scatter your X-variables in a visual matrix and eyeball collinearity.
7. **Variance Inflating Factor (VIF).** The VIF is a value that indicates how much collinearity is influencing the sampling variance of the coefficients for each independent variable. The general numerical guidelines are VIF > 10, sqrt(VIF) > 2.

Remedial Measures:

There are two choices when dealing with multicollinearity:

- 1) do nothing
- 2) follow some rules of thumb.

- Do Nothing: sometimes there is nothing you can do with the data. So either get new data, give up the project, or estimate things but realize that your claims of causality are VERY questionable.

- Adjust your model. That is, respecify with different or new variables.
- Transformation of variables. This can be done in many different ways. Sometimes it is to account for nonlinearities in the data itself, sometimes we can add lagged variables to account for dependence of previous values. This should be dictated by other investigatory methods.
- Composite Variables:
  - Factor Analysis
  - Composite Index
- Increase your N. This is easier said than done.

III. Heteroskedasticity: Non-constant Error Variance

Recall the Gauss-Markov Theorem: Under the assumptions of **linearity, constant variance, and error independence**, the OLS estimator has **minimum variance** compared to all linear unbiased estimators (it gives you the best standard errors relative to other linear estimations)

If we were to have **heteroskedasticity**: Note that a violation of the constant variance assumption is an **efficiency problem**. Ultimately, these violations will affect the sampling variance of our estimators, and thus our standard errors which affect our statistical inferences.

But, even in the face of heteroskedasticity, OLS remains unbiased.

What this looks like graphically:

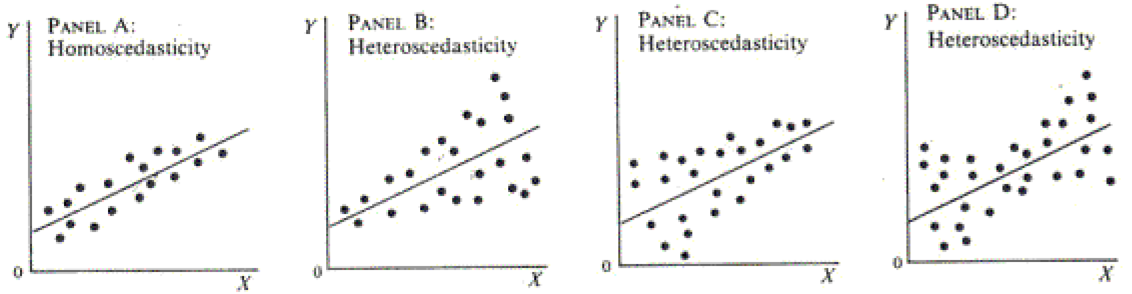
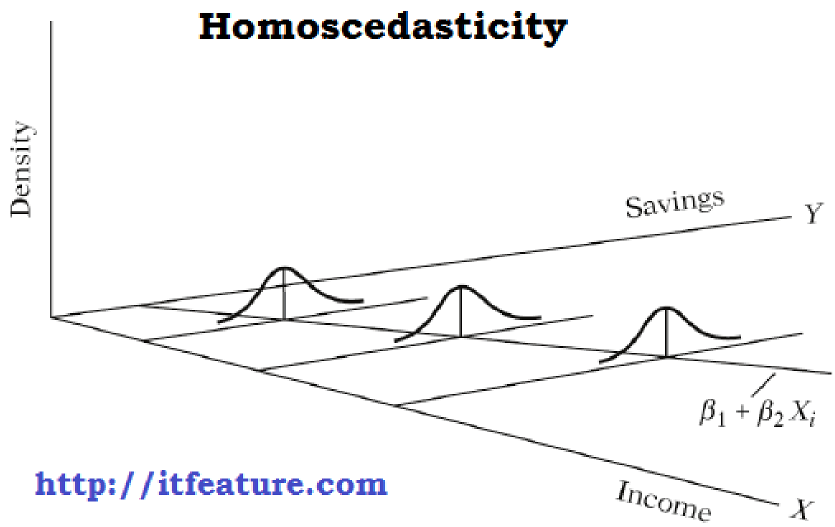
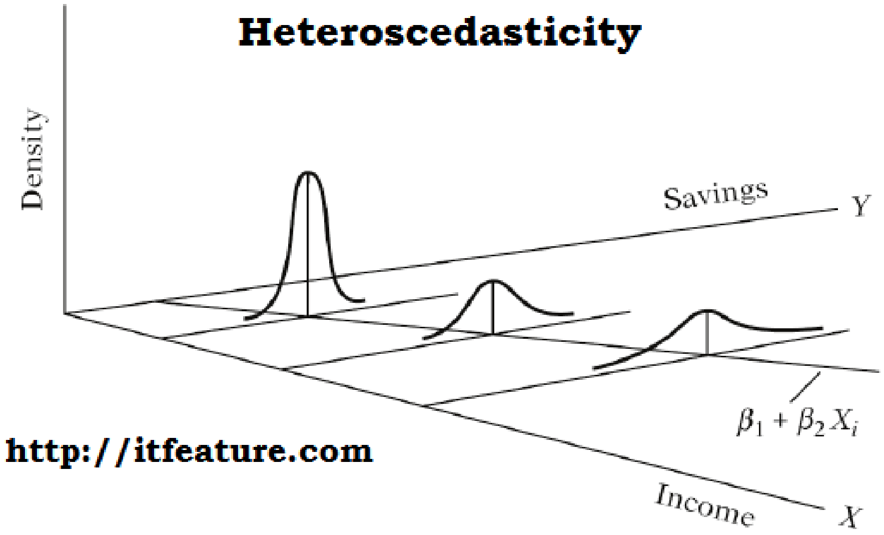


Fig. 9-1

If we are meeting OLS assumptions and **homoskedasticity** you'd see **equal variances** across the range of y-hat (fitted values, along x-axis, and residuals along y). Conceptually, the error variances around each x-value would be uniform, as shown below.



Under **heteroskedasticity**, the error distributions around the x-values would be nonuniform, as seen in the figure below.



Identifying heteroskedasticity:

1. Plot residuals against y-hats.
2. Plot the residuals against the x's

Tests for heteroskedasticity:

- **Breusch-Pagan test:**
    A. Test for whether error variance changes as a function of the independent variables

B. One can do a global test OR a covariate-specific test

- In this case our **null hypothesis** is homoskedasticity – ***so if you reject H_0 there's TROUBLE.*** You do **NOT** want to reject the null. Large p-values are "good," in that they suggest you are not violating the homoskedasticity assumption

Correcting for Heteroskedasticity:

- We know that if there is significant heteroskedasticity it is affecting our standard errors, which will be biased. So how do we correct for this?
    - The most common correction is robust standard errors, which are also called "heteroskedastic-consistent standard errors" or "White standard errors" after Albert White - This methods corrects for heteroskedasticity of an unknown form.
    - When variance is known, can use GLS or weighted least squares.
        - Usually we do not know the variance, so GLS is not helpful. We can however estimate the variance, when we do this we use fGLS (or feasible Generalized Least Squares).
        - Note: GLS and fGLS require us to make a lot of assumptions:
            - GLS requires that we know the variance
            - fGLS allows us to estimate the variance, but again this requires that we assume some model for the heteroskedasticity of the data
            - If we are unwilling to make these assumptions we can run OLS and then if we suspect heteroskedasticity we can correct with RSEs.[3]

[3] In light of King and Roberts and O'Hara and Parmeter it seems that the best approach would be to ALWAYS test for heteroskedasticity of the data and then determine whether we need to fix it with RSEs. Making a priori assumptions about the nature of the data seems unwise.

IV. Model Specification and Diagnostic Testing:

Multicollinearity and heteroskedasticity are indeed major issues with data analysis, but there are other problems that can arise. If a model is misspecified then a researcher opens themselves up to potential problems - inefficiency of the estimates, questionable confidence intervals. Model misspecification unlike multicollinearity and heteroskedasticity is not solely a "data issue" it is both a theory and data issue. Any model we estimate is precisely that, an estimate and we must make choices when we model. Some of these are driven by the data we have (the kind of DV we have, what we are looking to make claims about), yet simultaneously the choice of variables are dictated by theory and expectations.

There are a few key things that often lead to specification errors:

1. OVB - omitted variable bias
2. Garbage can regression
3. Wrong functional form
4. Measurement Errors
5. Incorrect Specification of the Error Term
6. Assumption that the errors are normally distributed

Consequences of Model Specification Errors:

Underfitting a Model (OVB) -
Suppose the true model is:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u_i$$

but instead we model

$$y_i = \beta_0 + \beta_1 x_1 + u_i$$

The consequences of omitting x_3 are as follows:

1. If the left-out variable is correlated with one that is included in the model, the R^2 is **nonzero** between the two and the **parameter estimates are inconsistent**.
2. If the variables are uncorrelated the **intercept is biased** but $\widehat{\beta}_2$ is unbiased
3. The **error variance is incorrectly estimated**
4. The variance of $\widehat{\beta}_2$ is a biased estimate of the true β_2.
    a. As a consequence the **confidence intervals are misleading**

The Consequences of Garbage Can Regression:

1. The OLS estimators of the overspecified model are unbiased and consistent

2. The error variance is correctly estimated
3. Confidence intervals are all good
4. **The beta coefficients are inefficient**. That is their variances will be generally larger than those of the true betas.

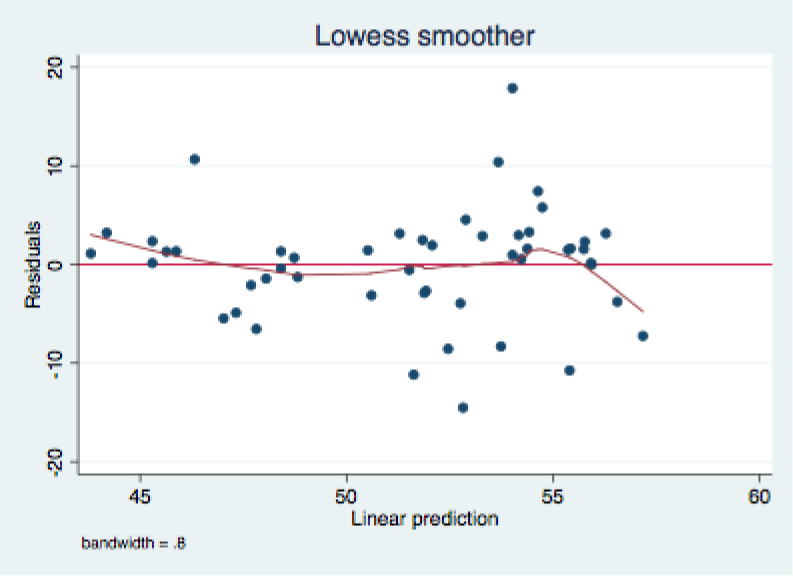
In such a case parsimony is the best approach.

Tests of Specification Errors:

Detecting Garbage Can regression: take a look at the F-statistic or the adjusted R^2

Testing for OVB and Incorrect Functional Form:

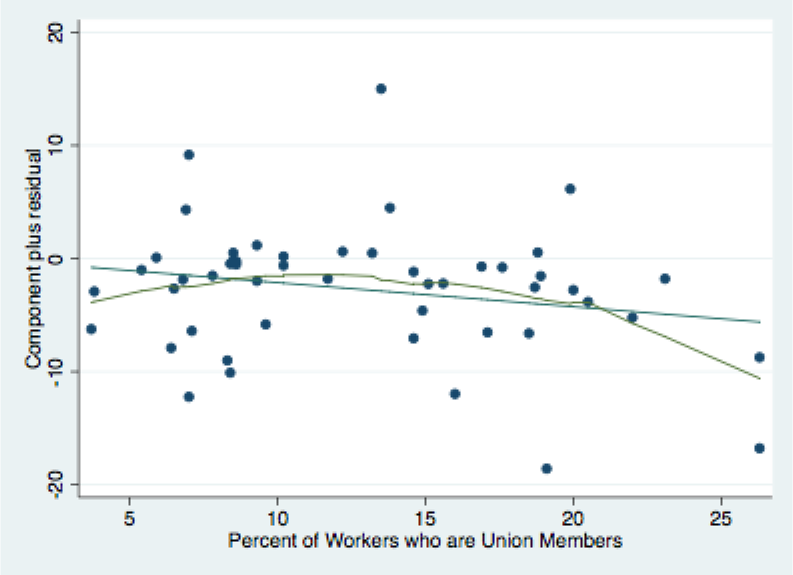
1. Examination of the residuals: Plot residuals against y-hat and the X's, if there are noticeable patterns then we need to investigate models misspecification
2. For functional form:
   a. Plot the residuals against the y-hats and overlay a lowess line. For example see below:



This graph shows that a linear functional form is not appropriate for this dataset. MLR.4 states: E(u|x) = 0, meaning we are assuming a zero conditional mean. If the expectation of the error term given the independent variables is equal to zero then the independent variables are uncorrelated with the error term (i.e., the independent variables are exogenous). There are two things that can induce a violation of MLR.4; omitted variable bias or specifying an incorrect functional form.

The graph above plots the predicted values for the response variable against the residual for each response term. The curvature line (the Lowess smoother) indicates that a linear model is not appropriate for this dataset.

a. Produce partial residual plots for each independent variable (“cprplot” in stata):



As noted in the example above When plotting the component plus residual for union, we see a concave quadratic functional form suggesting that we should make a change in the model to account for the data.

Measurement Error:

- Error in the Dependent Variable:
    - Measurement error in the DV will still produce unbiased beta-estimates, in other words, measurement error in the DV does not destroy the unbiased property of OLS. HOWEVER - the variances and standard errors of the beta-estimates will be unreliable. The estimated variances are now LARGER than in a case without measurement error.
- Error in the Explanatory Variables:
    - This is “a whole nother ball game”. Measurement error in the explanatory variables means we can no longer assume:

$$E(u_i|X_{i1}...X_{ik}) = 0.$$

    - That is, we can no longer assume that the stochastic error is independent of the explanatory variable. Is the error and the explanatory variable(s) are correlated this violated MLR.4 and the OLS estimators are not only biased but also inconsistent - that is they are asymptotically biased. Measurement error in the explanatory variables make consistent estimation of the parameters impossible

- Solution: One possible solution is the use of instrumental variables. Using IVs and 2SLS we can obtain consistent estimates of the beta-coefficients. However, this comes with its own can of worms.

Review of things that influence the size of standard errors:

- Discuss the four factors that influence the size of standard errors (of partial slope estimates) in multiple regression. Make sure to indicate the direction of influence (e.g., an increase in factor A inflates standard errors)
  - The standard error increases (i.e., worsens) as error variance increases. In other words, the greater the variance in the data from the population the greater our standard error will be since this is a measure of population variance.
  - The standard error decreases (improves) as the variance in X increases. Meaning, the greater the dispersion of the independent variable X the smaller the standard error.
  - The standard error decreases (improves) as the sample size increases. This is the case because as N increases the sample mean approaches the population mean (i.e., the sample mean becomes a better estimate of the population mean) and there will be less variability between the two.
  - Multicollinearity can increase standard error. That is, as a particular beta-estimator correlates with the other independent variables the standard error increases. This happens because part of the denominator for determining $SE(\beta_j)$ is a regression of X_1 on the remainder of the independent variables. Since this R^2x_1,x_2 is bounded by -1; 1 as the X correlate they will approach 1 (or equal 1) and will not decrease the overall error variance when divided.