36-401 Class Notes and Exercise for 10/17/00

Assumptions, continued

Howard Seltman

- Principles of Understanding the Effects of Assumption Violation on OLS Estimates

  - ◇ Least Squares coefficient estimates are linear combinations of the Y's. Under all assumptions, the estimates are unbiased, the standard error estimates are unbiased, and the t-tests, F-tests, and confidence intervals are correct, even for small samples. In addition, the coefficient are the most efficient (lowest variance) of all unbiased estimators.
  - ◇ If the covariates are measured with error, the coefficients are not even consistent; for any sample size, they are biased towards zero. (Predictions are OK.)
  - ◇ Assuming the model has the correct structural form, i.e. $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$,

    $$E(\hat{\boldsymbol{\beta}}) = E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[X\boldsymbol{\beta}] = \boldsymbol{\beta}$$

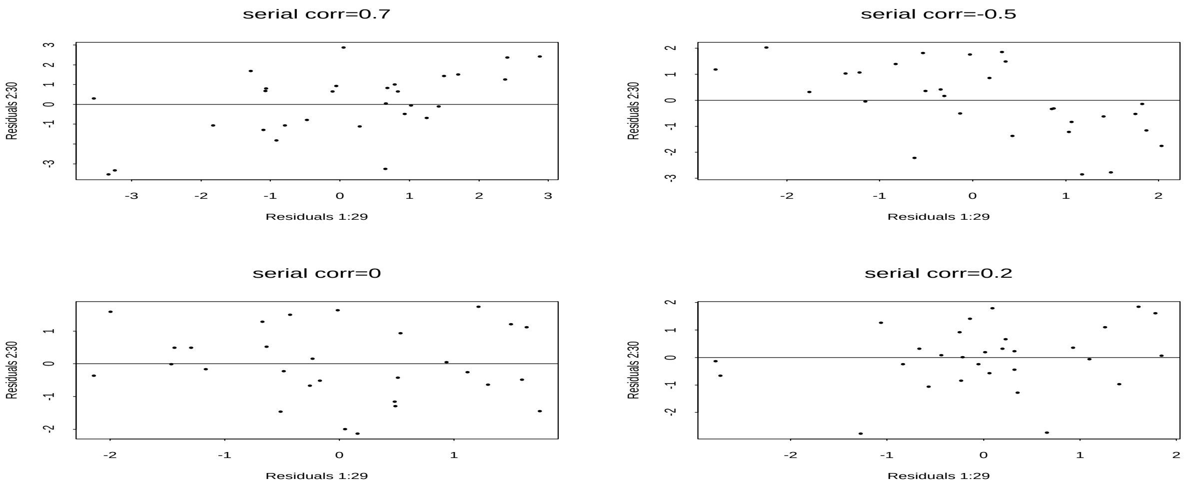
    and similarly

    $$V(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{Var(Y)}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

    which equals $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ if $\mathrm{Var(Y)} = \sigma^2\mathbf{I}$. So, if $E(\epsilon_i) = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$, $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$, but the errors are not normally distributed, then the coefficient and standard error estimates are unbiased, In this case, OLS estimates are Best Linear Unbiased Estimators (though they are not best among *all* unbiased estimators). Without normality, t-tests, F-tests and confidence interval coverage are lost, but they are approximately correct for large n (i.e. they are asymptotically correct) through the central limit theorem.
  - ◇ If $\epsilon \sim N(\mathbf{0}, \Sigma \neq \sigma^2\mathbf{I})$, least squares estimates of coefficients are unbiased, but the variances will tend to be large (i.e. the estimates are inefficient), and the standard errors are wrong.

Problem	**Biased b**	**Biased SE**	**Invalid t & F tests**	**High Var(b) (Inefficient)**
Nonlinear relationship	yes	yes	yes	–
Omit "relevant" X	yes	yes	yes	–
Include irrelevant X	no	no	no	yes
X measured with error	yes	yes	yes	–
Heteroscedasticity	no	yes	yes	yes
Autocorrelation	no	yes	yes	yes
Non-normal ϵ	no	no	yes	yes
Multicollinearity	no	no	no	yes

- Serially Correlated Errors



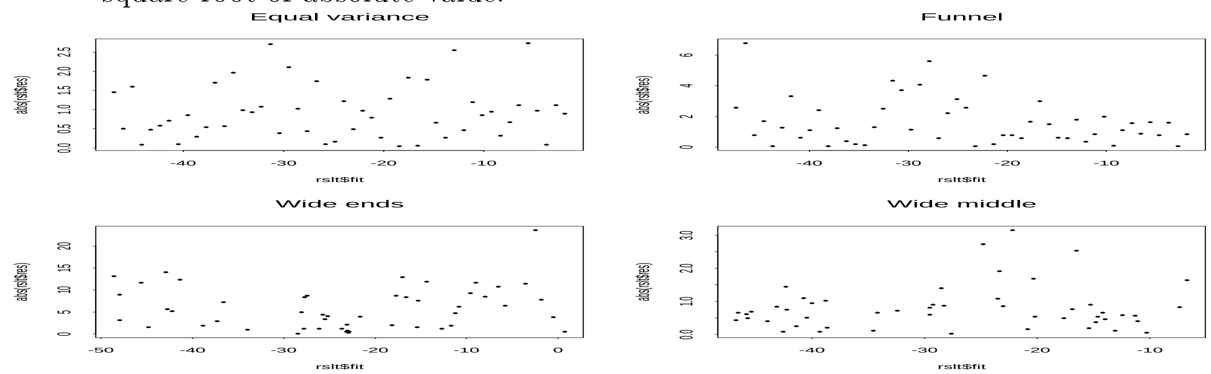
$$\text{Durbin-Watson test: } d = \frac{\sum_{i=2}^{n}(e_i - e_{(i-1)})^2}{\sum_{i=1}^{n} e_i^2}$$

Above graphs have true serial residual correlations of 0.7, -0.5, 0, 0.2 and using DW(rslt$res) from DW.q, d=0.89, 2.92, 2.23, 1.38 with d_L=1.35 and d_U=1.49 (from table in back of RwG book). H_0 is that the errors are uncorrelated. This is rejected for "A" only. Values between d_L and d_U are considered inconclusive. Values of d greater than d_U allow acceptance of H_0.

For negative correlation, $d > 2$ and we need to look up 4-d instead. For "B" we use d=4-2.92=1.08 which is significant (< 1.35).

- Heteroscedasticity

A good way to look at the equal variance assumption is to plot fitted values vs the *absolute value* of the residuals. Alternatives to absolute value include square or square root of absolute value.



- Generalized Least Squares (not Generalized Linear Model)

  More general model:
  $$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},\ \ \boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}_n, \boldsymbol{\Sigma})$$

  New estimator:
  $$\mathbf{b} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{Y}$$

- Partial Regression Plots to Check Structural Form

  These are an excellent way to see how each covariate is contributing to prediction of E(Y) in a multivariate regression model. They take into account multicollinearity in the following way: Out of the $K-1$ covariates in the model, choose one, say X_k. Regress X_k on all of the other covariates and call the residual "x". "x" is a measure of the information content of X_k after removing any common content with the other covariates. Now regress Y on all of the X variables except X_k. Call the residual "y". "y" is a measure of the variation in Y that is NOT explained by the other X's. The plot of "x" vs "y", with a regression line, is the "partial regression plot". It shows how much adding X_k to the model has helped in the modeling of Y. Among other uses, this plot is a good way to see if X_k should be transformed.

Breakout Session for 10/17/00

1. For these study descriptions state the one model assumption that is most clearly violated. Briefly explain why you conclude that the assumption is violated and how the violation affects our regression results.

   In all cases the model being fit is $Y_i = \beta_0 + \mathrm{X}_i\,\beta_1 + \epsilon_i$, $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$, i.i.d.

   (a) We want to study the effects of temperature on rat mortality. The predictor is room temperature. Levels are 66, 70, 74, and 78 degrees F. The response is mortality over a 6 month period. Ten rats are randomly assigned to each of the temperature levels. Rats who survive the entire period are coded as 1. Those who die are coded 0.

   (b) A study is conducted to test the theory that the amount of time high school students spend studying (Y) is inversely related to the amount of time they spend partying (X). The investigators interviewed a random sample of students and recorded their recollections of the number of hours spent studying and the number of hours spent partying during the previous year. (The "residual vs. fit" plot looked "all clear".)

   (c) In a study of growth of potted ivy plants, the covariate is water in quarts per pot per week. All of the plants have identical amounts of sun and fertilizer and are approximately the same size at the start of the experiment. The water levels are 0, 0.5, 1, 1.5, 5 and 20. Five plants each are randomized to each of the six levels of water. The response is plant weight at the end of fifteen weeks.

   (d) A researcher wants to study the effect of number of siblings on the number of movies watched among high school seniors. Students are instructed to keep a log of the number of movies they watch. The researcher interviews 30 consecutive students as the arrive at their senior prom, and records the number of siblings and the number of movies attended in the last two months.

2. Interpret the partial regression plots shown here:

```
rslt_lm(CMPGlog~Weight+WhlBas+EngRev+GasGal, cars)
summary(rslt)

par(mfrow=c(2,2))
partreg(xlst=c("Weight","EngRev","WhlBas","GasGal"), yname="CMPGlog",
   outname="Weight", df=cars, prop=T, id=T, tcex=0.7, cex=0.8)
partreg(xlst=c("Weight","EngRev","WhlBas","GasGal"), yname="CMPGlog",
   outname="EngRev", df=cars, prop=T, id=T, tcex=0.7, cex=0.8)
partreg(xlst=c("Weight","EngRev","WhlBas","GasGal"), yname="CMPGlog",
   outname="WhlBas", df=cars, prop=T, id=T, tcex=0.7, cex=0.8)
partreg(xlst=c("Weight","EngRev","WhlBas","GasGal"), yname="CMPGlog",
   outname="GasGal", df=cars, prop=T, id=T, tcex=0.7, cex=0.8)
ps("partreg")
```