36-401 Class Notes and Exercise
Regression Criticism I: Checking the Assumptions

The regression model is

$$y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{k-1} X_{i,k-1} + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2) \qquad (*)$$

We check assumptions by thinking about the experimental design, viewing plots, and sometimes by performing special statistical tests.

Checking the assumptions – Qualitative

- **No measurement error in X's?** Are the covariates (X's) measured with no (little) error? Is there a clear, precise, accurate method for measuring each covariate? Could we *"theoretically" obtain many random samples, each with the same X but with different Y's due to different ϵ's.?*

- **Linearity?** Is it reasonable to think that the Y's are a linear function of the X's? Is it reasonable to think that, in the region of the covariate space where we have observed the X's (a.k.a. the "design space"), the relationship between each covariate and the outcome may be monotone, and further, nearly linear?

- **Additivity?** We need to add interaction terms if the effects of each separate covariate is not identical at all levels of the other covariates.

- **Normality of errors?** From the problem description, is it reasonable to say $Y \sim N(\mu, \sigma^2)$? Use a histogram with overlaid normal curve and/or a Quantile-Normal plot to detect deviations from normality. Boxplots and symmetry plots will indicate skew or lack of skew, but lack of skew alone does not prove normality. Non-normal symmetric distributions include those with heavy or light tails. On a box-plot short length whiskers suggest lighter that Normal tails, and too many outliers suggests heavier than normal tails. Formal tests like Kolmogorov-Smirnov exist.

- **Independent errors?** Does knowing whether a certain unit has a positive residual tell us something about the residual of another specific unit (e.g. when we have 20 observations on 5 subjects)? Is there a clear relationship in a plot of e_i vs. e_{i-1}? Formal tests for serial correlation like Durbin-Watson exist. Adding appropriate covariates may restore independence.

- **Equal variances?** Is it clear that at different parts of the covariate space the variability of Y will vary more than at others on an absolute scale? Check the fitted vs. residual plot for different spread in different regions. Don't be fooled by the effects of different numbers of cases in different regions.

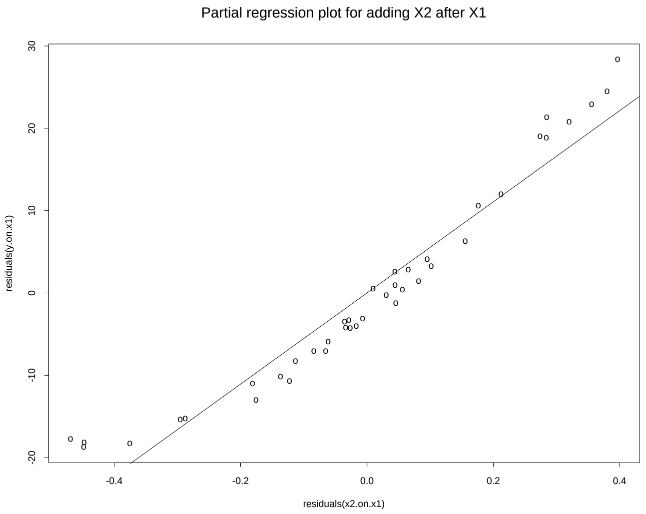
Checking the assumptions – Quantitative

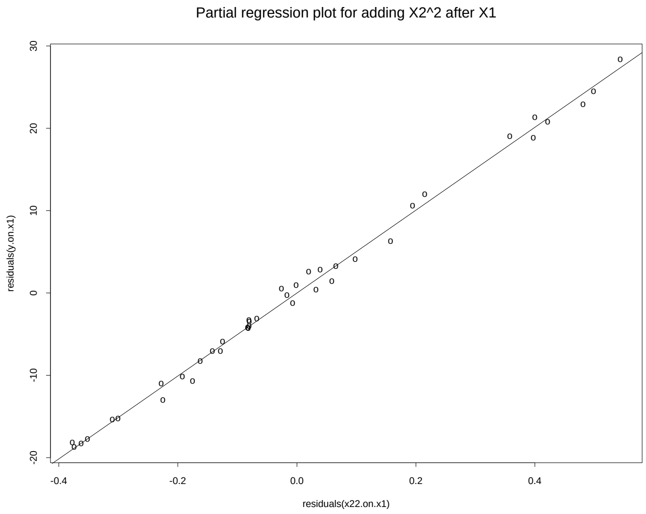
Least Squares coefficient estimates are linear combinations of the Y's. Under all the assumptions of $(*)$, the estimates are unbiased, the standard error estimates are unbiased, and the t-tests, F-tests, and confidence intervals are correct, even for small samples. In addition, the coefficient are the most efficient (lowest variance) of all unbiased estimators. *What are the effects of violating the assumptions in* $(*)$ *on the ordinary least squares (OLS) estimates?*

- **Linearity:** 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.

```
> X1 _ runif(40)
> X2 _ runif(40)
> Y _ 10 + 100*X1 + 50*X2^2 + rnorm(40)
> plot(residuals(x2.on.x1),residuals(y.on.x1),pch="o")
> abline(lm(residuals(y.on.x1) ~ residuals(x2.on.x1)))
> title("Partial regression plot for adding X2 after X1")
> x22.on.x1 _ lm(X2^2 ~ X1)
> plot(residuals(x22.on.x1),residuals(y.on.x1),pch="o")
> abline(lm(residuals(y.on.x1) ~ residuals(x22.on.x1)))
> title("Partial regression plot for adding X2^2 after X1")
```





- **Measurement error in X's:** 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.)

  Assume the usual simple regression model:

  $$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \; \epsilon_i \sim N(0, \sigma^2), \text{ i.i.d}$$

  But we cannot see the X's. Instead we see a noisy version, $\tilde{X}$.

  $$\tilde{X} = X + u, \; u \sim N(0, \tau^2), \text{ i.i.d}$$

  Then the relationship between the usual estimate b_1 and the true value, β_1 is:

  $$\mathbf{b} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{Y} \;\; \Rightarrow \;\; E[b_1] = \frac{\sigma^2}{\sigma^2 + \tau^2}\beta_1$$

- **Normality of errors:** 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

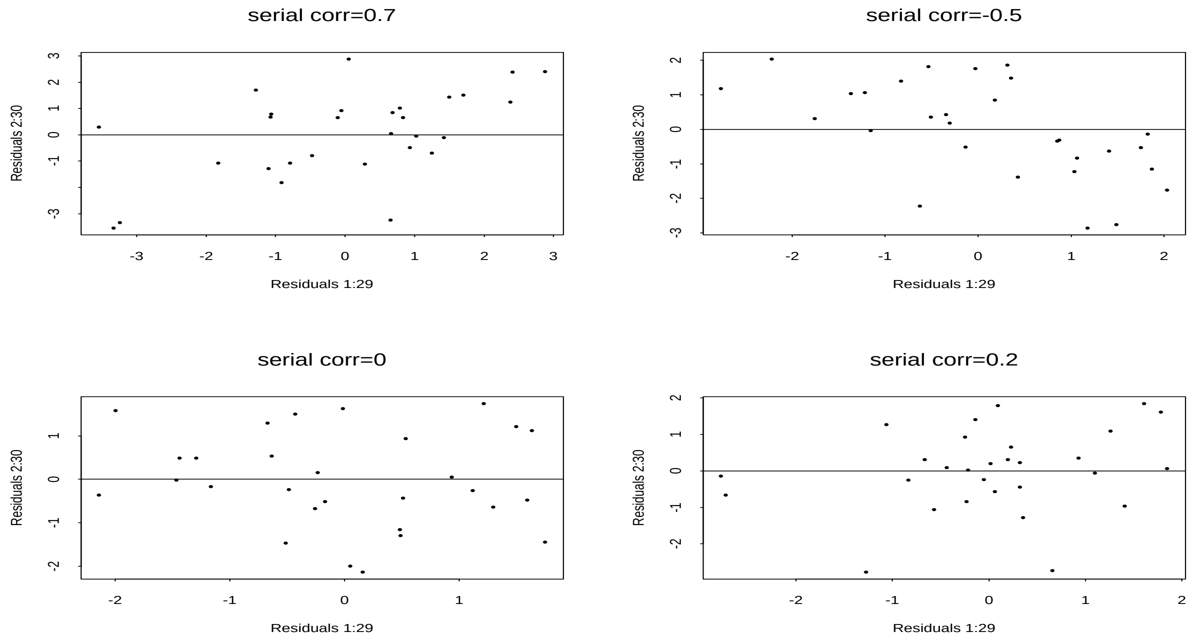
  $$V(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{Var(Y)}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

  which equals $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ if Var(Y) $= \sigma^2\mathbf{I}$. So, if $E(\epsilon_i) = 0$, Var$(\epsilon_i) = \sigma^2$, 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.

- **Correlated errors:** 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

A special case of correlated errors: Serially Correlated Errors.



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

Can use `DW(residuals(myfit.lm)` from `DW.q` to compute d:

Graph:	A	B	C	D
Correlation:	0.7	-0.5	0	0.2
DW statistic d:	0.89	2.92	2.23	1.38

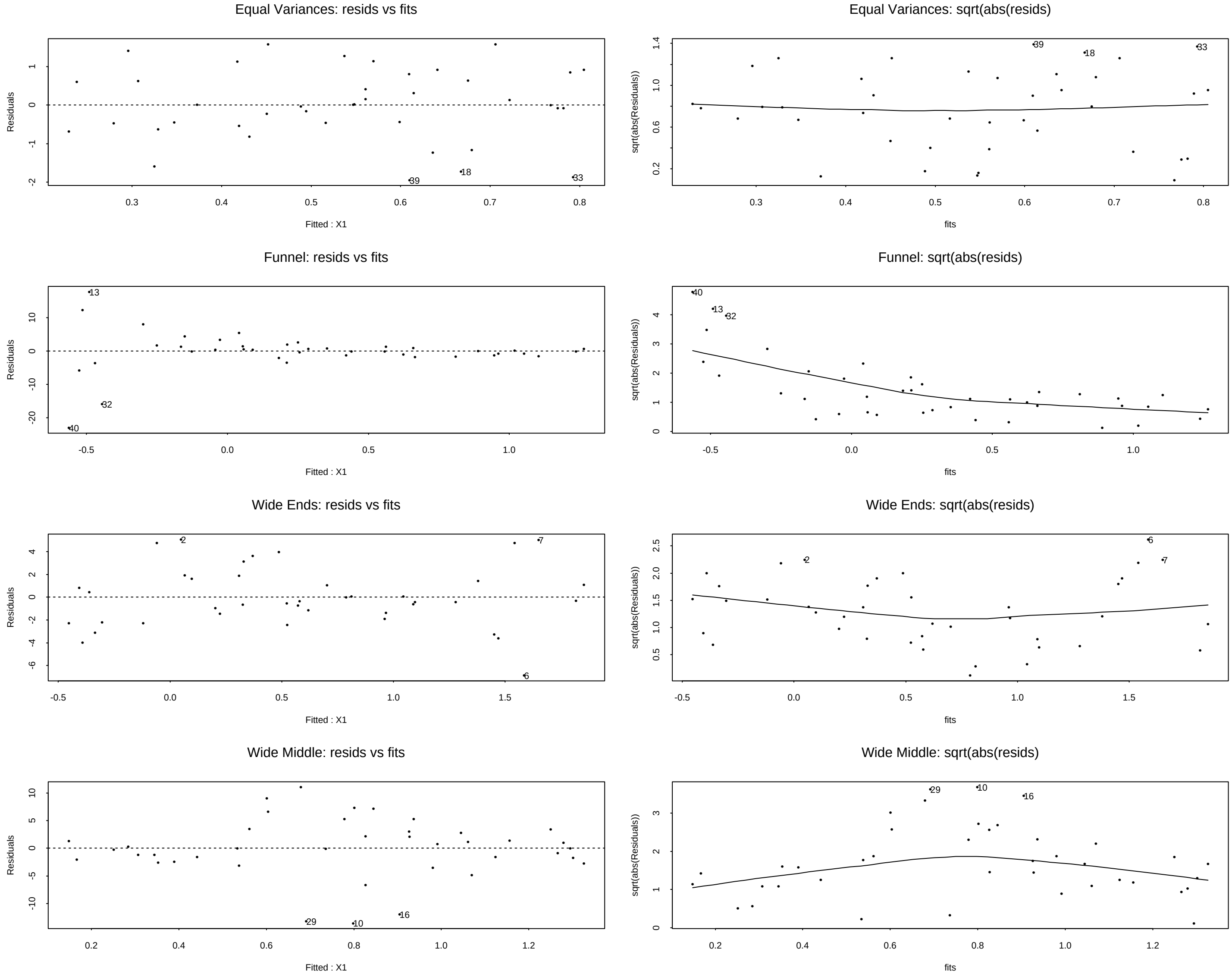
- ◇ *Test for positive correlation:* H_0 is that correlation is ≤ 0. From table in back of RwG, d_L=1.35 and d_U=1.49 (there was $k - 1 = 1$ covariate X, and $n = 30$ observations.).

  H_0 is rejected (at 0.05 level) 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.

- ◇ *Test for negative correlation:* H_0 is that correlation is ≥ 0. But now look up $d' = 4 - d$ instead of d.

  For "B" we use $d = 4 - 2.92 = 1.08$ which is significant ($< d_l = 1.35$).

- **Equal variances:** *Heteroscedasticity* (unequal error variances) can be checked by inspecting a residual plot directly. Another method 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): *How to handle* $\epsilon \sim N(\mathbf{0}, \Sigma \neq \sigma^2 \mathbf{I})$ *if all else fails.*

  More general model:

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

  New estimator:

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

Breakout Session

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 $Y_i = \beta_0 + X_i\,\beta_1 + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, 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 (for now ignore the sizes of the circles, we will talk about them next week):

```
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")
```