W203 Statistics for Data Science
Live Session 12: Additional Notes on the 6 CLM Assumptions

6 CLM assumptions

a. Linear population model
b. Random Sampling
c. No perfect multicollinearity
d. Zero-conditional mean
e. Homoskedasticity
f. Normality of Errors

a. Linear population model

Basic setup – typically by design. Nonlinearity is usually most evident in a plot of observed versus predicted values or a plot of residuals versus predicted values

b. Random Sampling

Poorly constructed design, measurement scheme or limited range can be indicators that the data was not randomly sampled.

Methods to check for random sampling can include:
1) evaluating the data collection process
2) checking for specific patterns of correlation (e.g. grouping variables)

c. No perfect multicollinearity

R will throw an error if perfect multicollinearity exists, so no need to check this assumption separately.

Imperfect collinearity can also be a problem since it increases standard errors. The Variance Inflation Factor (VIF) explains how much the standard error of each coefficient is inflated due to collinearity with other variables. An often-heard rule of thumb is to worry about any variable with a VIF greater than 4. which implies that the R^2 exceeds 75% (i.e. VIF = $\frac{1}{1-R^2}$ = $\frac{1}{1-.75}$ = 4). However, you should really pay attention to overall research context. For example, is a coefficient of key interest nonsignificant, even though the effect size seems large, due to a high VIF?

```
library(car)

vif(model)

vif(model)>4
```

What steps can you take?

1. Consider dropping variables from the model that are correlated with variables of interest. HOWEVER, dropping variables can introduce omitted variables bias and hence, should

only be done with a clear understanding of the data and theory underlying the statistical methods used to develop the model. This is an example of a variance-bias tradeoff.

2. Sometimes, variable transformations can be used to reduce multicollinearity.
3. Increase the sample size - Small samples are particularly vulnerable to multicollinearity problems because multicollinearity reduces your effective sample size for the effects of individual predictors

d. Zero-conditional mean

Zero conditional mean assumption states that the error u has an expected value of zero, given any values of the independent variables: E(u|x1, x2,…,xk)=0.

Look at a residuals vs. fitted values plot to check this assumption. R provides a red smoothing curve that tracks the conditional mean of the residuals. Curvature in this curve indicates a violation of zero-conditional mean.

```
par(mfrow=c(2,2))
plot(model)
```

What steps can you take?

1. First, check if you have a large sample size. If sample size is sufficiently large and zero-conditional mean is not satisfied, check whether cov(x,u) = 0. Then coefficients will be biased, but consistent.
   NOTE: this condition is referred to as exogeneity in the asynchronous material and may also be called the zero covariance assumption or weak exogeneity assumption.

2. Transforming variables can sometimes fix violations of zero-conditional mean.

(For causal models) Note that omitted variables that are correlated with an X always bias coefficients and violates both the zero-conditional mean and zero covariance assumption. Since there is not direct way to check for omitted variables, you will need build a case for exogeneity using background knowledge.

e. Homoskedasticity

Once the regression model is built, set par(mfrow=c(2, 2)), then, plot the model using plot(lm.model). This produces four plots.

The condition of homoscedasticity can be seen visually on the residuals versus fitted values plot as a band of constant thickness. It can also be seen on the scale-location plot as a flat red smoothing curve.

```
par(mfrow=c(2,2))   # set 2 rows and 2 column plot layout
plot(model)
```

What steps can you take?

Generally, we simply use heteroskedasticity-robust standard errors (a.k.a. Huber-White standard errors, or "sandwich estimator").

Alternative analysis techniques, such as least absolute residuals, weighted least squares, bootstrapping, or jackknifing, are also designed to be used for heteroscedasticity problems (*Note: this is outside the scope of this class*).

f. Normality of Errors

We can visually check whether the residuals are normally distributed using the qqnorm() plot (top right plot).

```
par(mfrow=c(2,2))
plot(model)
```

The qqnorm() plot in top-right evaluates this assumption. If points lie exactly on the line, it is perfectly normal distribution. However, some deviation is to be expected, particularly near the ends, but the deviations should be small.

What steps can you take?

1. First, check if you have a large sample. If so, CLT implies that OLS coefficients have a normal sampling distribution.
2. If you have a small sample, you may have to alter your specification to achieve normality of errors. You can try
    a. Variable transformations
    b. Inclusion of different predictors
    c. Investigate outliers.

Alternatively, tools that do not require normality can be used.