Week 7: Linear model assumptions and diagnostics

Marcelo Coca Perraillon

University of Colorado
Anschutz Medical Campus

Health Services Research Methods I
HSMP 7607
2019

Updated notes are here: https://clas.ucdenver.edu/marcelo-perraillon/teaching/health-services-research-methods-i-hsmp-7607

Outline

- Review of assumptions
- Gauss-Markov theorem
- The linear model is BLUE
- Using residuals to diagnose non-normality and non-linearity

Big picture

- We have made several assumptions when estimating the linear model using OLS and MLE
- Today, we will cover more formally some assumptions to show that (paraphrasing) **the linear model is the bomb**
- (If you are into skiing and white hair is not yet a concern: the linear model is rad, epic; insert many !!!!!!)
- Note that I'm saying that linear regression is the bomb, not OLS (we saw that MLE is pretty much the same)
- Once we understand the role of each of the assumptions, we can start talking about diagnosing violations; what is usually called **regression diagnostics**
- We are also going to cover ways of "fixing" some problems

Assumptions 1 to 4

- **1)** Linear in parameters: The parameters of the population model are linear with respect to the outcome. So we can write the model as $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$
- **2)** Random sampling: We assume that there was random sampling and we obtained data for Y and X_1 to X_p (note: not saying that it came from a randomized experiment)
- **3)** No perfect **collinearity**: In the sample and population, none of the covariates is constant and there are no exact linear relationships among the independent variables (Chapter 9)
- **4) Zero conditional mean**: The error has an expected value of zero given any values of the independent variables: $E[\epsilon|x_1, ..., x_p] = 0$; alternatively, no confounders were left behind (in the error, that is)

Number 4, again

- Assumption number 4, zero conditional mean, is about the **population**, not the model in the sample
- We saw from the first order conditions that the residuals always add up to zero and that the covariance, and thus correlation, between the residuals and the explanatory variables is zero
- The distinction between sample and population is key for understanding causal inference the economist way. Without the zero conditional mean, we can't say that our model estimated using a sample represents a conditional expectation
- We saw that the "modern" way of understanding causal inference separates the design of the study and the particular way of estimating a relationship

Big picture

- Remember that we covered some basic properties of estimators
- The first one is that $E[\hat{\beta}] = \beta$, that is, **unbiasness**
- The second one was **efficiency**. Among estimators, we want the one that has smaller standard error; or the one that is more *precise*
- The last property is asymptotic: the larger the sample size the closer, in probability, our estimator should get to the true but unknown population parameter (**consistency**)
- With the four assumptions of the previous slide, we can show that the parameters estimated from OLS (or MLE) are **unbiased**. So: $E[\hat{\beta}_j] = \beta_j$ (note that the zero conditional mean is the key one)

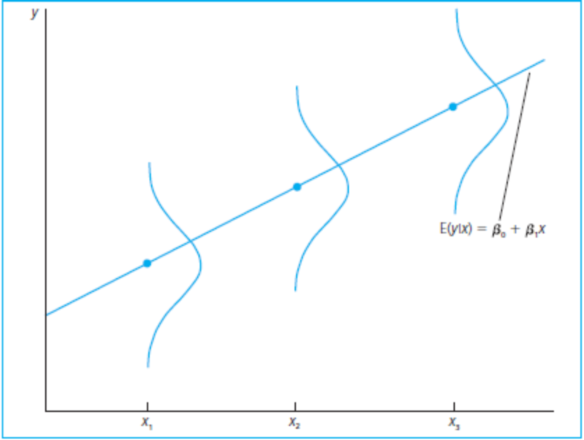
Assumption 5

- **5)** Homoskedasticity: The error ϵ has the same variance given any value of the explanatory variable. That is, $var(\epsilon|x_1, ..., x_p) = 0$
- We have seen this one before but we didn't pay much attention to it
- For OLS, we made no assumptions about the variance. We did need that assumption to figure out the distribution of $\hat{\beta}_j$
- We said that $\epsilon \sim N(0, \sigma^2)$. Note that it is σ^2 and not σ_i^2. We are saying that observations are idd
- When we covered MLE, we assumed that the data came from a normal distribution whose mean was explained by covariates but we said that the observations had a normal distribution with σ^2 (again, not σ_i^2)
- Also, we didn't make σ^2 a function of covariates (we could, it's actually not that hard to estimate using MLE)

Graphically

- From Wooldridge:

FIGURE 2.1 E(*y*|*x*) as a linear function of *x*.



Heteroskedasticity is often a concern in applied analysis

- It's often the case that the variance increases with the values of a covariate, if only because samples sizes are smaller
- Think about income (Y) and education (X) for example. At the high end of education, those with PhDs or MDs or MDs and PhDs tend to have higher incomes (surgeon?) or relatively low incomes (teacher?)
- For those that have fewer years of education, income is more homogeneous. Also, only 2% of the population has a PhD or another advanced degree
- We will see ways to deal with **heteroskedasticity** (Chapter 7)
- **But remember**: even with heteroskedasticity, we still have *unbiased estimators*; in the linear model, **heteroskedasticy is a problem for inference because the standard errors are going to be wrong**
- How wrong? It depends, but it's not hard to solve this problem (with larger samples)

Gauss-Markov theorem

- Under assumptions 1-5, the Gauss-Markov theorem says that the estimates from the linear model (obtained via OLS or MLE) are BLUE
- BLUE: **B**est **L**inear **U**nbiased **E**stimator
- The unbiased part is easy; the "best" here means that the parameters $\hat{\beta}_j$ have the smallest variances among the class of all linear unbiased estimators
- In other words; With OLS, we get unbiasness, efficiency, and consistency

The linear model is BLUE

- This is purely for **didactic** purposes so you remember
- Most Valentine's days some economist friend has to post a variation of the following:

  Roses are red; OLS is BLUE; I'll run regressions anytime with you

  Roses are red; OLS is BLUE; I'm 95% confident that I really love you!
- **Please don't do it**; but you get the fascination with BLUE
- Also, don't make the same mistake economists tend to make all the time: **it's not OLS**, it's the linear model; OLS is just a method of estimation (to be fair, MLE has the wrong variance but it's easy to fix)
- It's the model $Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_p X_{pi} + \epsilon_i$, where ϵ_i are iid and $\epsilon_i \sim N(0, \sigma^2)$. This is equivalent as saying that $Y \sim N(\beta_0 + \beta_1 X_{1i} + ... + \beta_p X_{pi}, \sigma^2)$

Simulating a normal model

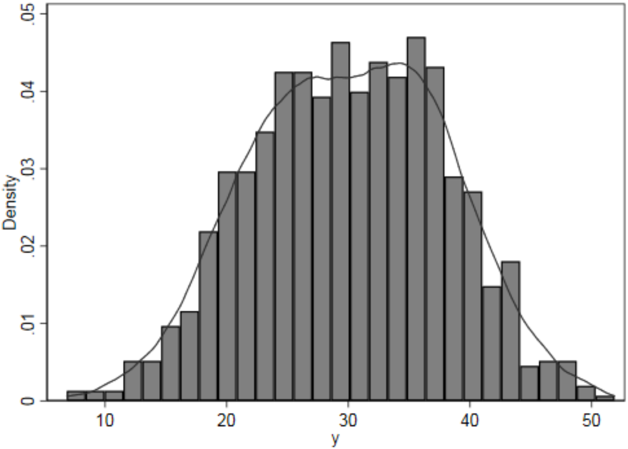
- You can easily simulate a linear model. **Don't underestimate simulations**. They help you understand the abstract concepts
- (Btw, this is the simulation I should have done when covering MLE adding a covariate to model the mean as a function of one covariate X)

```
set obs 1000
* I just need an X, doesn't need to be random
gen x =  runiform()*10
gen y = 20 + 2*x + rnormal(0,5)
* Note the 5 there, that's the standard deviation of y, so variance is 5^2

sum y x
 Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
           y |      1,000    30.06353    7.923546   6.852018   51.87202
           x |      1,000    4.891377    2.938474   .0002378   9.997806
```

$y \sim N(\hat{\beta}_0 + \hat{\beta}_1 x, \sigma^2)$

- So Y also distributes normal, with conditional variance of 5^2



We can estimate our simulated model

- We should recover the parameters using the reg command

```
reg y x
      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(1, 998)       =   1519.18
       Model |  37852.9849         1  37852.9849   Prob > F        =    0.0000
    Residual |  24866.8191       998  24.9166524   R-squared       =    0.6035
-------------+----------------------------------   Adj R-squared   =    0.6031
       Total |  62719.8041       999  62.7825866   Root MSE        =    4.9917

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   2.094814   .0537453    38.98   0.000     1.989347     2.20028
       _cons |   19.81701   .3066381    64.63   0.000     19.21528    20.41874
------------------------------------------------------------------------------
```

- Note the difference between Root MSE and standard deviation of Y from summarize command (observed vs unexplained by model)

Back to diagnostics

- We seldom can be certain that we can satisfy all five assumptions but some are more trivial than other [e.g. 1) and 2)]
- **The zero conditional mean cannot be verified with data**; you need knowledge about the subject and how the data was collected (experimental? Observational?). For causality, we need to understand the **assignment mechanism**
- We can check some of the other assumption and make corrections if necessary
- Besides causal inference (zero conditional mean), the most common to be concerned about are **heteroskedasticity** and **collinearity** (not perfect collinearity, though, that one is easy – you won't get estimates or you will get a warning from Stata)
- What other assumptions did we make or **did NOT make**?

Other assumptions or lack thereof

- We made no assumptions about the distribution of the explanatory variables X
- However, an implicit assumption is that the variables X were **measured without error**
- There are models that are used when some explanatory variables are measured with error (for example, random coefficient models)
- The main consequence of measurement error is that it makes the estimates less precise (higher standard errors) and the Wald tests are "biased towards the null" (when the null is $H_0 = \beta_j = 0$)
- Remember, less precise estimates mean larger standard errors and thus larger confidence intervals, so we will tend to not reject the null
- There are several types of measurement error (systematic, random, classical, etc)

Regression diagnostics

- We will use regression diagnostics to check for violations of some assumptions or other important problems
- In particular:
  1. Deviations from the normality assumption
  2. Observations that have leverage, influence, or are outliers
  3. Multicollinearity
  4. Heteroskedasticity
- For now, 1) and 2)

Using the residuals

- We will use the residuals to check some assumptions
- Remember, we define the residuals as $\hat{\epsilon}_i = y_i - \hat{y}_i$ for the n observations $i = 1, ..., n$
- We calculate residuals by predicting y and subtracting from the observed y: $(y_i - \hat{y}_i) = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p)$
- There is another way to think about predicted values. We could write them as a function of observed values:
- $\hat{y}_i = p_{i1} y_1 + p_{i2} y_2 + \cdots + p_{in} y_n$, for $i = 1, 2, ..., n$
- More compact: $\hat{y}_i = \sum_{j=1}^{n} p_{ij} y_j$
- In other words, for each observation i, the predicted outcome $\hat{y}$ can be written as the weighted sum of all observed values, weighted by p_{ij}
- Intuitively, p_{ij} has to depend on the **value of the predictor variables**

Leverage

- Look at the formula again: $\hat{y}_i = p_{i1}y_1 + p_{i2}y_2 + \cdots + p_{in}y_n$, for $i = 1, 2, ..., n$
- We are saying that each predicted value for each observation can be written as a **weighted sum of all the other outcomes values** y in the dataset
- Think about regression towards the mean and how the prediction for one observation depends on the value of all other observations
- So each observation has it's own weight. In the simple case of one predictor:

  $p_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$
- We call p_{ii} the leverage value of observation i
- Note that observations with a value x_i that is **far away from the average** $\bar{x}$ will have more leverage
- You will be happy to know that p_{ii} are the diagonal elements of your **beloved projection matrix P**

Leverage example

- Simulate a regression model with one predictor
- The post-estimation command **predict** has an option “leverage” to calculate the leverage for each observation

```
set obs 100
gen x =  runiform()*10
gen y = 20 + 2*x + rnormal(0,5)
sum x
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
           x |        100    5.150371     2.77518   .2311043   9.997318

* Calculate leverage
predict ylev, leverage
* Sort in descending order (so highest leverage on top)
gsort -ylev
```

- We can display the top 15 observations in terms of leverage

Graphing leverage

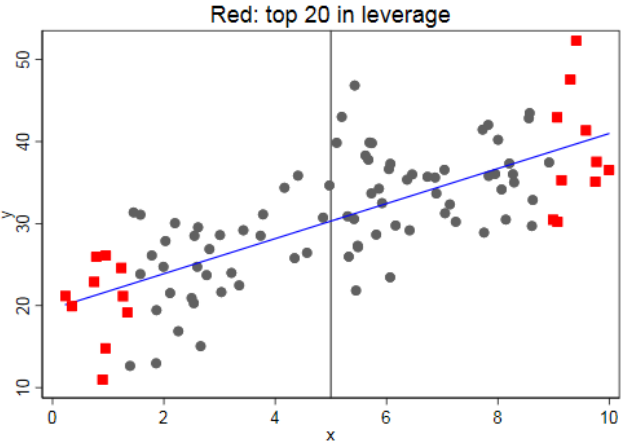
- Remember that the special variable "_n" indexes observations (so _$n == 1$ is the first observation; _$n == 100$ the last in this example because there are 100 observations)

```
scatter y x if _n > 20, xline(5) || lfit y x, color(blue) legend(off) || ///
        scatter y x if _n <= 20, color(red) title("Red: top 20 in leverage") ///
saving(leverage.gph, replace)
graph export leverage.png, replace
```

- I'm mixing three graphs. The xline(5) draws a vertical line at the mean of x, which is around 5

Graphing leverage

- As we just saw, observations with values away from the mean of X will have more leverage



What to do with leverage?

- By themselves, the leverage values are not that informative; it makes sense that observations that are away from the central value will influence predictions
- We go over leverage points because we would like to study the **residuals**, but the variance of the residuals are a function of leverage points: $var(\hat{\epsilon}) = \sigma^2(1 - p_{ii})$
- So we can standardize the residuals: $r_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1-p_{ii}}}$,

  where $\hat{\sigma} = \sqrt{\frac{SSE}{(n-p-1)}}$
- This are called the **studentized residuals** or standardized residuals, which have a mean of zero and a standard deviation of 1
- Now we have comparable residuals (same variance)

Stata

- The option rstandard of the predict command calculates the standardized residuals

```
predict res_std, rstandard
sum res_std
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     res_std |        100   -.0005914    1.003669   -2.81966   2.493449
```

- Now, checks

Check linearity and normality

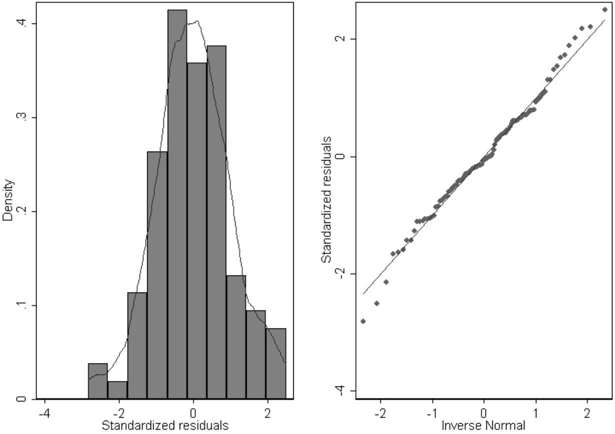
- The standardized residuals should have a normal distribution

```
hist res_std, kdensity saving(rno.gph, replace)
qnorm res_std, saving(qno.gph, replace)
graph combine rno.gph qno.gph, row(1)
graph export nor.png, replace
```

- qnorm compares the variable with a normal distribution (Q-Q plot)
- qnorm compares quantiles of the observed variable to a theoretical normal with mean and standard deviation like those of the observed variable

Normality

- We simulated the data following a normal so no surprise that the residuals follow a normal distribution



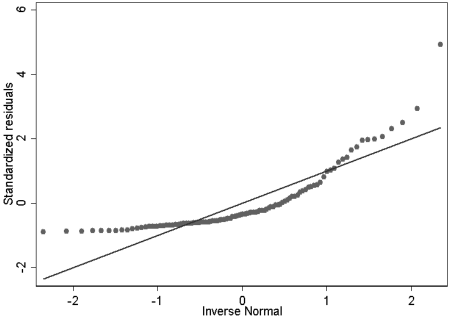
What about violations?

- Let's simulate two situations. 1) Non-mormal errors and 2) mispecification of model

```
* True model has Chi-square errors
gen y2 = 20 + 2*x + (rnormal(0,5))^2
* We estimate a model that assumes normal errors
reg y2 x
* Calculate standardized residuals
predict y2res, rstandard
* Plot
qnorm y2res, saving(qny2.gph, replace)
graph export qny2.png, replace
```

- We should expect the residuals not to be normally distributed

Non-normal errors



- One problem with this method of detection is that the violation could be about using the wrong model (mispecification), not only, or because, the error was not normal

Model mispecification

- What if we estimate the wrong model?

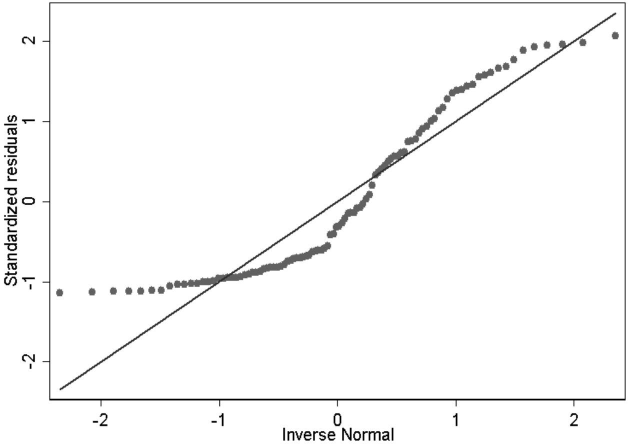
```
* The true model has normal errors (and normal outcome y) but it is quadratic on x
gen x2 = x^2
gen y3 = 20 + 2*x + 5*x2 + rnormal(0,5)

* We estimate a model that assumes a linear relationship between x an y
reg y3 x
     Source |       SS           df       MS      Number of obs   =       100
-------------+----------------------------------   F(1, 98)        =   1522.19
       Model |  2358966.01         1  2358966.01   Prob > F        =    0.0000
    Residual |  151872.536        98  1549.71976   R-squared       =    0.9395
-------------+----------------------------------   Adj R-squared   =    0.9389
       Total |  2510838.54        99  25362.0055   Root MSE        =    39.366

------------------------------------------------------------------------------
          y3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   54.43913    1.39533    39.02   0.000     51.67014    57.20811
       _cons |  -77.35757    8.62617    -8.97   0.000    -94.47592   -60.23921
------------------------------------------------------------------------------
predict y3res, rstandard
qnorm y3res, saving(qny3.gph, replace)
graph export qny3.png, replace
```

- Note the fit is good (R^2, nothing seems off with the model)

Model mispecification



- Not normal again...

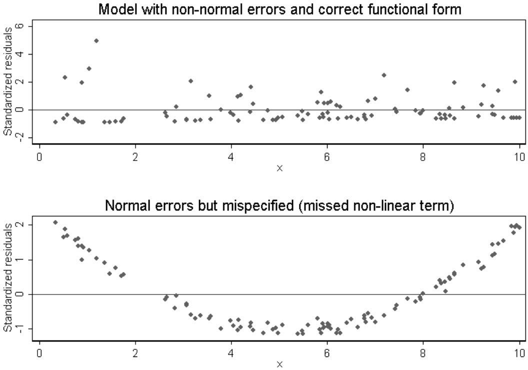
Plot residuals against *x* values

- Plotting the residuals against each of the covariates is helpful to detect **non-linearity** (mis-specification)

```
scatter y2res x, yline(0) title("Model with non-normal errors but correct functional form") ///
   saving(y2r.gph, replace)
scatter y3res x, yline(0) title("Normal errors but mispecified") ///
   saving(y3r.gph, replace)
graph combine y2r.gph y3r.gph, col(1)
graph export resplots.png, replace
```

- Remember, the cor(residual, x) is zero **but remember, too, that the correlation is about a linear relationship**

Plot residuals against x values



- Top one is ok, the bottom one shows that we didn't take into account a non-linear relationship between x and y (What's the correlation coefficient in the second one? Zero, of course)

Summary

- We can use the (standardized) residuals to diagnose **some** departures from assumptions
- We **CANNOT** use residuals for the zero conditional mean assumption or, equivalent, the ignorable treatment assignment assumption
- One key is to understand what assumptions 1 to 5 imply (say, what is the problem with heteroskedasticity?)
- Recall that some properties of linear regression are algebraic and will always be true in the **sample**; some assumptions we only needed for inference
- Next class, more diagnostics. Next week, finally dummy variables (yay!!)