Linear Model Diagnostics (and What to Do if You Flunk One)

EH6127 – Quantitative Methods

Steven V. Miller

Department of Economic History and International Relations

Stockholm University

Goal(s) for Today

1. Summarize the important assumptions of OLS.
2. Familiarize you with basic diagnostic tests you should run on every OLS model.
3. Introduce you to what to do if you flunk a diagnostic test.

Which Assumptions?

OLS has *lots* of assumptions.

- Every definition is an assumption, which compounds matters.

The problem (for us as teachers): *which* assumptions are most important?

- Maximalist position: any violation is sufficient to reject the model.
- However: violation of certain assumptions have certain *implications*.

A LINE Mnemonic

Think of this `LINE` mnemonic for OLS' assumptions:

- `L`: the outcome *y* is a *l*inear function of the right-hand side variables.
- `I`: the errors (residuals) are *i*ndependent from each other (i.e. no autocorrelation).
- `N`: the distribution of errors is *n*ormal
- `E`: the variance of the errors is *e*qual/constant (i.e. no heteroskedasticity).

This is roughly in order of importance.

- I'd actually put `E` before `N`, but `LIEN` conjures image of debt collection.
- That said, ask me about this one again if I catch you doing the "linear probability model."

This also says nothing about the validity of the data or representativeness of the sample.

- Those are assumptions too!

First, a Caveat

There are typically “textbook” assumptions I want to briefly mention:

1. Multicollinearity
2. Specification (“all relevant variables”)

My caveat: these are important, but the problem is trivial (or your book is misleading you).

Multicollinearity

Multicollinearity is when two right-hand side variables are so highly correlated that OLS cannot reliably return their partial effects.

- *Perfect* collinearity: the model cannot be identified.
- High collinearity: your standard errors start to explode.

Diagnostics: correlation matrix, variance inflation factors

- Correlation matrix: absolute value above .8 indicates a problem.
- Variance inflation factor: a value above 5 typically indicates a problem.

Solution(s):

1. Kick one of the offending variables out.
2. Principal components analysis (i.e. create a latent measure)

Specification Issues

Model specification issues are often presented as “including all relevant variables” that predict *y*. Caveat:

- There’s no formal test for this (in the way you’re thinking about it).

Specification issues are only critical for adjusting for confounding. Assume three scenarios:

1. `X` and `Z` both explain variation in `Y`, but `X` and `Z` have no overlap (correlation).
2. `X` and `Z` both explain variation in `Y`, and `X` and `Z` have overlap (correlation).
3. `X` (but not `Z`) explains variation in `Y`, and `X` and `Z` have overlap (correlation).

Specification Issue Scenarios

First scenario: omitting Z has no bearing on "true" effect of X on Y.

- Omitting Z will just decrease R^2, which is no real problem for causal identification.
- *Ask yourself what the goal of the model is*.

Second scenario: omitting Z biases relationship between X and Y in direction of the overlap.

- Including Z fixes this, barring a massive collinearity problem.

Third scenario: this is an instrumental variable problem (an advanced topic).

- Even these have their own peculiarities (i.e. in the real world, they are a huge leap of faith).

Linearity

OLS assumes an outcome *y* is a linear function of some predictors.

- This also implies the model is *additive*.
- Without this assumption, the model is no longer "linear."

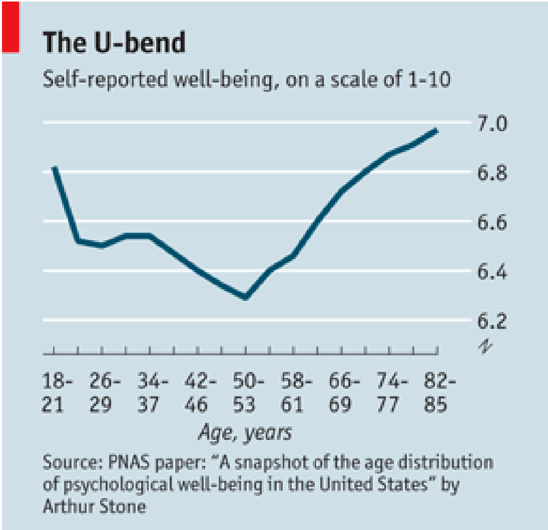
Diagnostics: mostly visual (esp. fitted-residual plot). But also:

- Utts' (1982) "Rainbow" test
- Harvey-Collier test
- *ed. Neither of these are very good; just look at the data/model*.

Solutions:

- New model?
- Logarithmic transformations
  - e.g. if $y = abc$, then $log(y) = log(a) + log(b) + log(c)$
- Interactions/square terms?

An Example: Age and Happiness



The U-bend

Self-reported well-being, on a scale of 1-10

Source: PNAS paper: “A snapshot of the age distribution of psychological well-being in the United States” by Arthur Stone

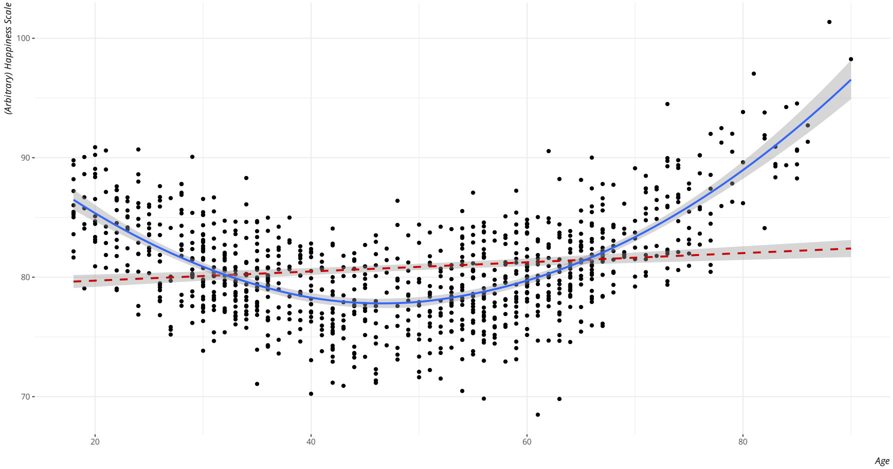
An Illustration of a Curvilinear Effect

`fakeHappiness` in `{stevedata}` has an illustration of this phenomenon.

- These (fake) data follow a basic form for age and happiness: `happy ~ .95*age + .01*(age^2) + controls + e`

A Curvilinear Relationship: What OLS Wants to Say (Red) vs. What it Actually Is (Blue)

This textbook curvilinear relationship can be fixed with a simple square term in the model.



Data: ?fakeHappiness in {stevedata}. Data are simulated to illustrate a curvilinear relationship.

The Fitted-Residual Plot

The fitted-residual plot will point this out (and other things).

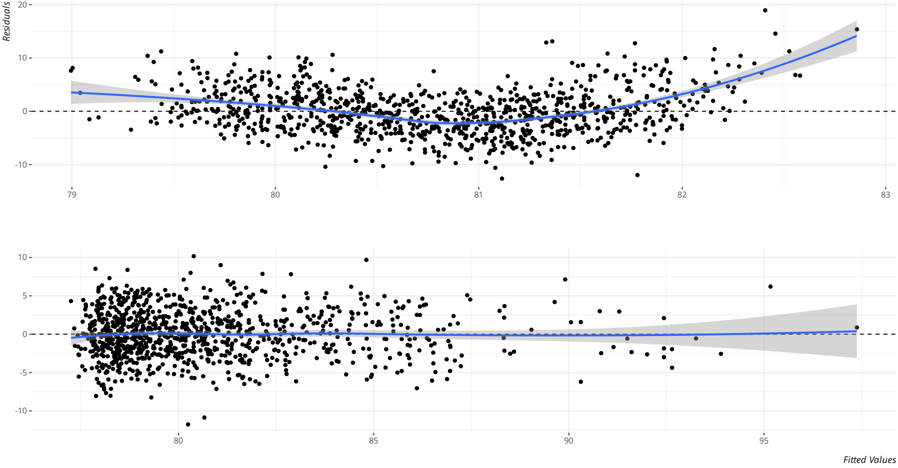
- Fit the model, extract the fitted values and residuals
- Plot the fitted values on the *x*-axis, and residuals on the *y*-axis.

By definition, the linear line will be flat at 0.

- You want to overlay a LOESS smoother to check for patterns.

The Top Plot is Bad; The Bottom Plot is Good

The top plot tries to impose one age effect. The bottom outright models the square term. The fitted-residual plot on top points to a problem.



Data: ?fakeHappiness in {stevedata}. Models include relevant controls too.

(Error) Independence (or: No Autocorrelation)

Another biggie: OLS assumes the data are randomly drawn from an underlying population.

- The inclusion of one observation should have no bearing on the inclusion of another observation.
- The residual value for one observation cannot depend on the residual for other data points.
- If they do, OLS loses its inferential value.

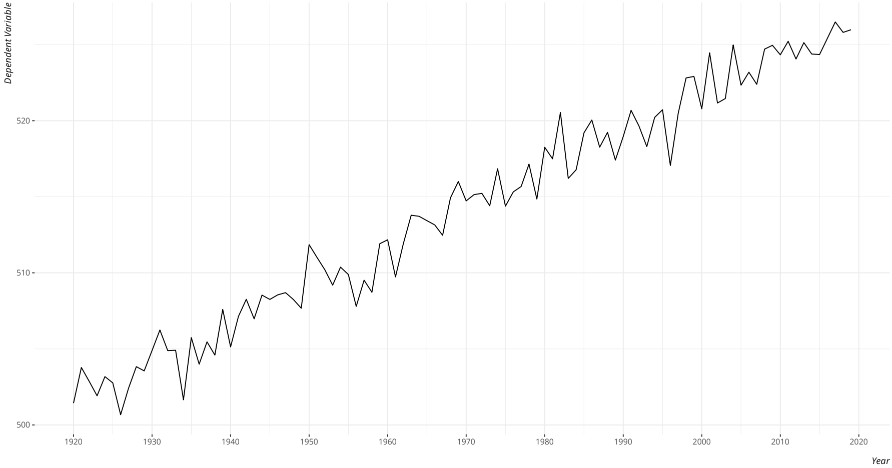
Diagnostics: honestly? Just know what you're doing. There are three common situations for this.

1. Time series (i.e. y is informed by past values of y)
2. "Multilevel"/hierarchical models (i.e. individuals are clustered within higher units)
3. Omitted variable bias (a bit harder to diagnose)

Solutions: various, depending on the non-independence.

Consider This (Fake) Time Series with a Clear Linear Trend

This time series has systematic determinants (x1 and x2), but also a clear time-dependent trajectory.



*Data: ?fakeTSD in {stevedata}. y = .25*year + .25*x1 + x2 + e, where x2 is binary and x1 is normal(5,2).*

Time Series Diagnostics

The diagnostic test for a problem here starts with estimating your linear model and doing one of two tests.

- Durbin-Watson test
- Breusch-Godfrey test
- *ed. just do Breusch-Godfrey. Durbin-Watson is more restrictive, but has some informational value.*

For both: if the *p*-value is below some threshold (e.g. .05), you have a problem.

- Solutions in this case may include first-differences, lag effects, time-trends.
- Also kind of depends on the point of the modeling process.

```r
# library(stevedata); library(lmtest)
M1 <- lm(y ~ x1 + x2, fakeTSD)
dwtest(M1)
#> 
#>  Durbin-Watson test
#> 
#> data:  M1
#> DW = 0.05506, p-value < 2.2e-16
#> alternative hypothesis: true autocorrelation is greater than 0
bgtest(M1)
#> 
#>  Breusch-Godfrey test for serial correlation of order up to 1
#> 
#> data:  M1
#> LM test = 92.359, df = 1, p-value < 2.2e-16
```

Normality of Errors

OLS assumes the distribution of residuals is normal with a mean of 0 and some variance derived from the model. Caveats:

- This is *not* an assumption the DV is normal, but it does kind of imply it.
- Says nothing about the IVs.
- All diagnostic “tests” for this are kinda bad and are sensitive to any discreteness in the DV.
- The implication of this violation is about the errors and not the regression line itself.

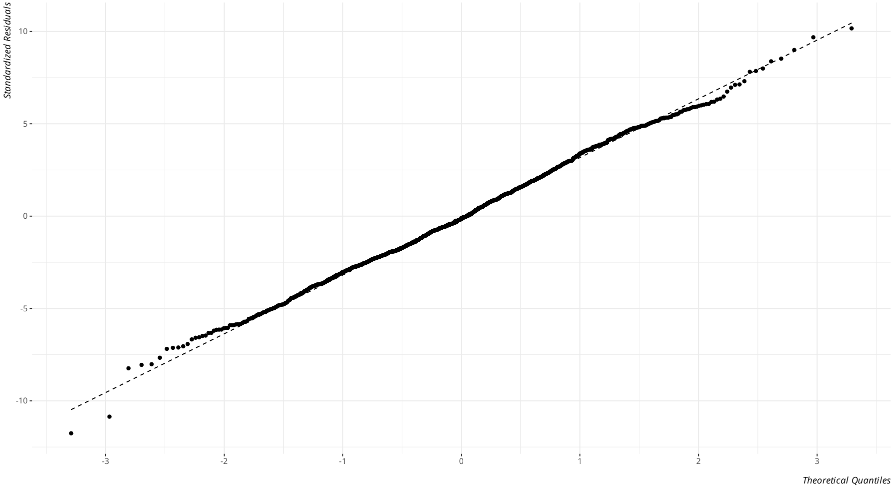
Diagnostics: visual (Q-Q plot), and assorted normality tests (which are all bad)

Solutions: model the non-normality through a GLM

- i.e. you’re very likely trying to impose OLS on a DV with a finite set of values.

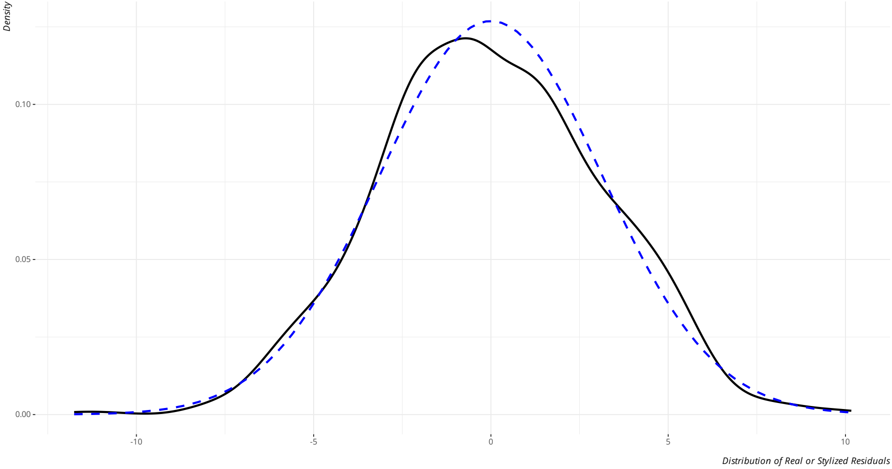
The Q-Q Plot of Our Square-Term Age and Happiness Model

Ideally, plotting the quantiles of residuals against the standardized residuals puts everything on a straight line



Another Way: Plot the Distribution of Residuals Against a True Normal Distribution Matching Its Description

The real thing is always going to look kind of lumpy, but you can see just how bad any issue might be.



Black solid line is the actual ditribution of residuals. Blue-dashed line is the stylized residuals matching its description.

Equal/Constant Error Variance (i.e. No Heteroskedasticity)

OLS assumes the dispersion of the error terms does not depend on the fitted values (homoskedasticity).

- If they do, the line is fine but the standard errors are wrong (in ways you don't know).
- This has major implications for null hypothesis significance testing.

Diagnostics: fitted-residual plot, Breusch-Pagan test

Solutions: more "robustness tests" comparing OLS to some other approach.

- e.g. IV/DV transformations, weighted least squares, on-the-fly heteroskedasticity corrections, bootstrapping

A Tell-Tale Case of Heteroskedasticity

Assume this simple data-generating process to illustrate heteroskedasticity.

```r
set.seed(8675309)
tibble(
  x = rnorm(1000, 5, 2),
  e = rnorm(1000, 0, 1),
  y = 5 + .25*x + e
) -> C # for constant

M3 <- lm(y ~ x, C)

tibble(
  x = rnorm(1000, 5, 2),
  e = rnorm(1000, 0, 1 + x), # uh-oh...
  y = 5 + .25*x + e
) -> H # for heteroskedasticity

M4 <- lm(y ~ x, H, na.action=na.exclude)
# ^ I'm thinking ahead to those of you wanting to copy-paste code for homework.
# If you have missing data you want to ignore, put that argument in the lm() function.
```

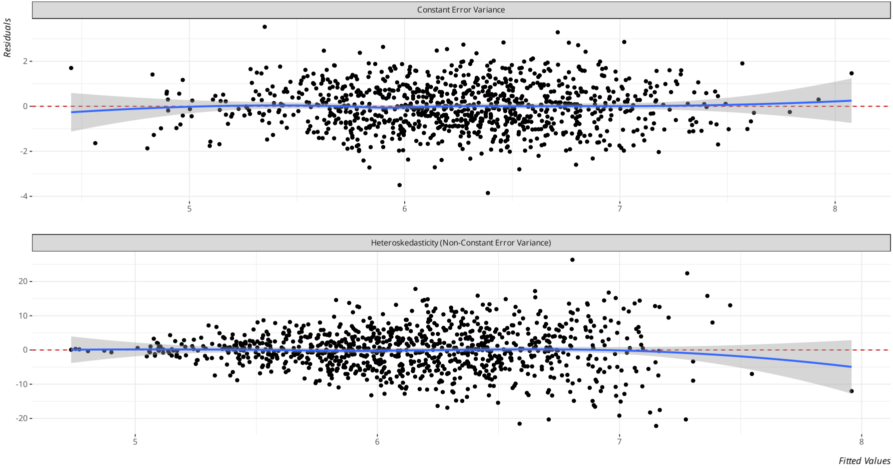
Table 1: Comparing Two Models With and Without Constant Error Variance

	C	**H**
Independent Variable (x)	0.264***	0.244*
	(0.016)	(0.096)
Intercept	4.952***	4.920***
	(0.085)	(0.515)
Num.Obs.	1000	1000
R2 Adj.	0.216	0.005

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The Top Plot is Good; The Bottom Plot is Bad

Non-constant error variances (heteroskedasticity) results in a tell-tale 'cone of shame', making error estimates automatically suspect.



Data: fake data from this lecture. See previous slides. Notice the line is largely unaffected.

The Breusch-Pagan Test

The Breusch-Pagan test (in `{lmtest}`) will tell you this as well.

- *p*-value below certain threshold = heteroskedasticity

```
bptest(M3)
#>
#>  studentized Breusch-Pagan test
#>
#> data:  M3
#> BP = 1.1241, df = 1, p-value = 0.289
bptest(M4)
#>
#>  studentized Breusch-Pagan test
#>
#> data:  M4
#> BP = 132.1, df = 1, p-value < 2.2e-16
```

Approach 1: Weighted Least Squares (WLS)

This procedure is a mouthful to explain, but:

1. Run the offending model (which we already did).
2. Grab the residuals and fitted values
3. Regress the absolute value of the residuals on the fitted values of the original model.
4. Extract those fitted values.
5. Square them, and...
6. Divide 1 over those values.
7. Finally, apply those as weights in the linear model once more for a re-estimation.

Woof, Okay...

```r
# We already did step 1
tibble(resid = resid(M4),
       fitted = fitted(M4)) -> fitred # 2

M5 <- lm(abs(resid) ~ fitted,
         data = fitred, na.action=na.exclude) # 3
# ^ again, be mindful about missing data you may have.
H$wts <- 1/(fitted(M5)^2) # 4, 5, 6

M6 <- lm(y ~ x, data=H, weights = wts) # 7
# psst, {stevemisc} can do this:
# wls(M4)
```

Table 2: Comparing Two Models With and Without Constant Error Variance (with Robustness Tests)

	C	**H**	**H (WLS)**
Independent Variable (x)	0.264***	0.244*	0.225***
	(0.016)	(0.096)	(0.046)
Intercept	4.952***	4.920***	5.011***
	(0.085)	(0.515)	(0.133)
Num.Obs.	1000	1000	1000
R2 Adj.	0.216	0.005	0.023

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Approach 2: Some Kind of On-the-Fly Standard Error Correction

These are multiple, but `feols()` in `{fixest}` can do this easily.

```r
library(fixest)

M7 <- feols(y ~ x,
            data=H, se = "hetero")
```

Table 3: Comparing Two Models With and Without Constant Error Variance (with Robustness Tests)

	C	**H**	**H (WLS)**	**H (OtF-SE)**
Independent Variable (x)	0.264***	0.244*	0.225***	0.244*
	(0.016)	(0.096)	(0.046)	(0.108)
Intercept	4.952***	4.920***	5.011***	4.920***
	(0.085)	(0.515)	(0.133)	(0.459)
Num.Obs.	1000	1000	1000	1000
R2 Adj.	0.216	0.005	0.023	0.005

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Approach 3: Bootstrap! *LFG*

The simple bootstrap is a useful approach to assess what heteroskedasticity might be doing here.

- Sample, with replacement, *M* times from the data to create *M* replicates of the original data.
- Re-run the model on *each* of those replicates.
- Summarize the model parameters with the mean of the estimates and standard deviation of the estimates.

```
set.seed(8675309)
# Get the data from the model
model.frame(M4) %>%
  # draw 1000 bootstrap resamples
  modelr::bootstrap(n = 1000) %>%
  # estimate the model 1000 times
  mutate(results = map(strap, ~ update(M4, data = .))) %>%
  # extract results using `broom::tidy`
  mutate(results = map(results, tidy)) %>%
  # unnest and summarize
  unnest(results) %>%
  group_by(term) %>%
  summarize(std.error = sd(estimate),
            estimate = mean(estimate))
#> # A tibble: 2 x 3
#>   term        std.error estimate
#>   <chr>           <dbl>    <dbl>
#> 1 (Intercept)     0.468    4.94
#> 2 x               0.109    0.242
```

Table 4: Comparing Two Models With and Without Constant Error Variance (with Robustness Tests)

	C	H	H (WLS)	H (OtF-SE)	H (Bootstrap)
Independent Variable (x)	0.264***	0.244*	0.225***	0.244*	0.242*
	(0.016)	(0.096)	(0.046)	(0.108)	(0.109)
Intercept	4.952***	4.920***	5.011***	4.920***	4.940***
	(0.085)	(0.515)	(0.133)	(0.459)	(0.468)
Num.Obs.	1000	1000	1000	1000	1000
R2 Adj.	0.216	0.005	0.023	0.005	0.005

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Conclusion

OLS has assumptions, and you should really know them.

- Assumptions have varying levels of importance.
- Violation of certain assumptions have varying implications.

Important takeaways:

- Get to know the fitted-residual plot, and what it can tell you.
- Know what types of data are assuredly going to have autocorrelation.
    - Formal diagnostics have some informational value, but you can know ahead of time if you have a problem.
- The normality assumption is about the *errors* and not the *DV*.
    - That said, think long and hard about what you're doing if you want to impose OLS on a Likert item or dummy variable.
- Unequal error variances have no solution, per se.
    - Throw more rocks at your model and see if the results meaningfully change.

Table of Contents

Introduction

The LINE Mnemonic for OLS Assumptions

- A Caveat About Other Assumptions
- Linearity
- Independence (i.e. No Autocorrelation)
- Normality (of the Errors)
- Equal Error Variance (i.e. No Heteroskedasticity)

Conclusion