POL 213 – Spring 2023

Quantitative Analysis in Political Science II

Lecture 4

Model Diagnostics and Assessment

Lauren Peritz

U.C. Davis

lperitz@ucdavis.edu

April 17, 2023

Table of Contents

OLS review

What are we trying to do with OLS?

1. Model some variable $\boldsymbol{y} = [y_1, y_2, ..., y_n]'$ as a linear function of some variable(s)

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{1,1} & ... & x_{1,k} \\ \vdots & & & \vdots \\ 1 & x_{n,1} & ... & x_{n,k} \end{bmatrix}$$

2. That is, we wish to regress $\boldsymbol{y}$ onto $\boldsymbol{X}$, estimating the population parameters $\boldsymbol{\beta} = [\beta_0, ..., \beta_k]'$ in the function: $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{e}$.

We'd like these estimates to have a couple of properties:

1. Minimize both bias (residuals) and sampling variance.
2. Allow for statistical inference, modeling uncertainty in estimates.

When the data do not behave perfectly in accordance with our assumptions, we need to modify our model, moderate our conclusions, or both.

Outliers, Leverage, and Influence

In simple regression, an **outlier** is an observation whose response-variable value is conditionally unusual given the explanatory variable (discrepancy).

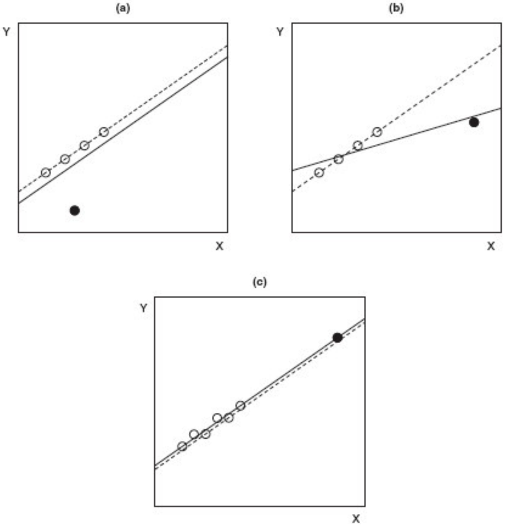
Some outliers exert strong **leverage** on the regression coefficients, swaying them substantially.

Thus we can use the following heuristic:

$$\texttt{Influence = Leverage} \times \texttt{Discrepancy}$$

Outliers, Leverage, and Influence

Leverage and influence in a simple regression



Source: Fox (2015) figure 11-2.

Outliers, Leverage, and Influence

In each graph of the preceding figure, the solid line gives the least-squares regression for all the data, while the broken line gives the least-squares regression with the unusual data point (the black circle) omitted.

(a) An outlier near the mean of X has low leverage and little influence on the regression coefficients.

(b) An outlier far from the mean of X has high leverage and substantial influence on the regression coefficients.

(c) A high-leverage observation in line with the rest of the data does not influence the regression coefficients. The two regression lines are separated slightly for visual effect.

Assessing Leverage (bivariate)

Measure leverage with “hat value” h_{ij} which capture the contribution of observation Y_i to the fitted value $\hat{Y}_j$

$$\hat{Y}_j = h_{1j}Y_1 + h_{12}Y_2 + ... + h_{jj}Y_j + ... + h_{nj}Y_n = \sum_{i=1}^{n} h_{ij}Y_i$$

In simple (bivariate) regression, the hat-value measures distance of each point from the mean of X:

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^{n}(X_j - \bar{X})^2}$$

Assessing Leverage (multivariate)

In multiple regression, $h_{ii} \equiv h_i$ measures the distance from the centroid of the X's, taking into account the correlational and variational structure of the X's. These are the diagonal entries of the hat matrix, the matrix that transforms $\boldsymbol{y}$ into $\hat{\boldsymbol{y}}$:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \mathbf{H}\boldsymbol{y}$$

Note that the *hat-matrix* is symmetric ($\boldsymbol{H} = \boldsymbol{H}'$) and idempotent ($\boldsymbol{H}^2 = \boldsymbol{H}$) so we can focus on the diagonal entries.

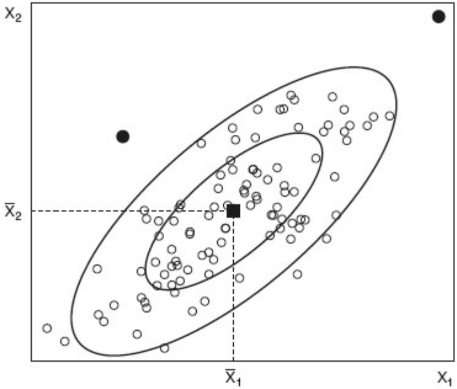
$$h_i \equiv \mathbf{h_i}'\mathbf{h_i} = \sum_{j=1}^{n} h_{ij}^2$$

Because $\boldsymbol{H}$ is a projection matrix, projecting $\boldsymbol{y}$ orthogonally onto the $(k+1$ dimension) subspace spanned by columns of $\boldsymbol{X}$, it follows that $\sum h_i = k+1$.

As with the bivariate case, h_i, the leverage of the ith observation, is directly related to the distance of the this observation from the center of the explanatory variable scatter.

Assessing Leverage

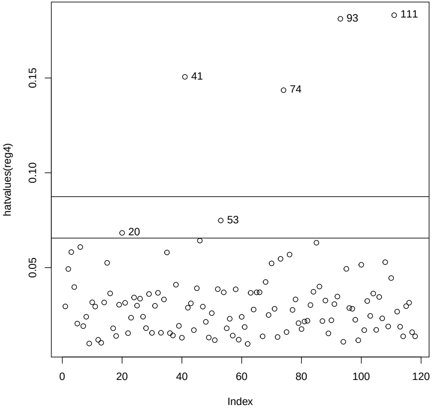
Elliptical contours of constant leverage (constant hat values) for $k = 2$ explanatory variables.



Source: Fox (2015) figure 11-4.

Example

```
library(car)

UN <- read.table(
"http://socserv.mcmaster.ca/jfox/Books/Applied-Regression-2E/
datasets/UnitedNations.txt",
header=TRUE)

reg4 <- lm(tfr ~ GDPperCapita + illiteracyFemale + contraception, UN)

plot(hatvalues(reg4))  # index plot of hat-values
abline(h=c(3, 4)*4/183)  # three and four-times average hat-value
identify(hatvalues(reg4)) # identify high leverage points

print(UN[c(20,  41,  53,  74,  93, 111),])
```



```
> print(UN[c(20,  41,  53,  74,  93, 1
             region  tfr
Benin         Africa 5.83
Cook.Islands Oceania 3.50
Ecuador      America 3.10
Guam         Oceania 3.04
Japan           Asia 1.48
Macau           Asia 1.60
```

Detecting Outliers

To identify an outlying observation, we need an index of the unusualness of Y given the X's. Discrepant observations usually have large residuals. Even if the errors ε_i have equal variances, the residuals do not:

$$V(E_i) = \sigma_\varepsilon^2(1 - h_i)$$

High leverage observations tend to have small residuals (they pull regression surface toward them).

We can form a **standardized residual** by calculating

$$E_i' \equiv \frac{E_i}{S_E\sqrt{1 - h_i}}$$

where $S_E = \sqrt{\sum E_i^2/(n-k-1)}$. This measure is inconvenient because the numerator and denominator are not independent preventing E_i' from following a t-distribution.

Detecting Outliers

So to detect outliers, we usually use a **studentized residual**. Its num. and den. following a t-distribution with $n - k - 2$ degrees of freedom and it has an estimate of the standard error of the regression based on deleting the ith observation and using the remaining $n - 1$ data points.

$$E_i^* \equiv \frac{E_i}{S_{E(-i)}\sqrt{1 - h_i}}$$

Outliers in the data might reveal (1) measurement error (2) unexpected patterns that cause us to rethink the data generating process.

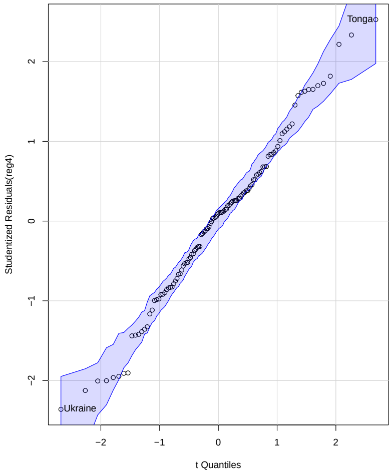
Detecting Outliers

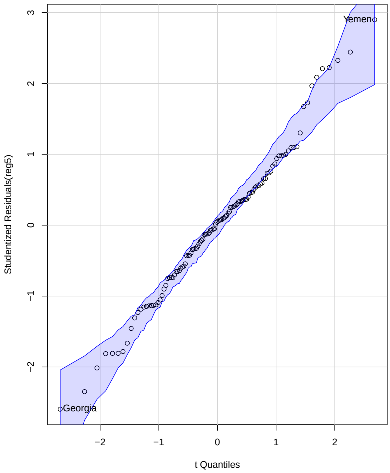
There are a few ways to do this. Here's an example with what's referred to as a Bonferroni test for the largest absolute residual (fox p 248):

```
reg4 <- lm(tfr ~ GDPperCapita + illiteracyFemale + contraception, UN)
reg5 <- lm(tfr ~ GDPperCapita + illiteracyFemale + contraception +
as.factor(region), UN)

> outlierTest(reg4)  # Bonferroni t-test
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
      rstudent unadjusted p-value Bonferroni p
Tonga  2.52752           0.012869           NA

> qqPlot(reg4, simulate=TRUE, line="none")  # QQ plot for stud. res.
  Tonga Ukraine
    186     193
```

If the model is correct and there are no true outliers, then each studentized residual follows a t-distribution with $n - k - 2$ degrees of freedom.

Detecting Outliers





Measuring Influence

Observations that combine high leverage with large studentized residual exert substantial *influence* on regression coefficients. Cook's D statistic provides a summary index of influence.

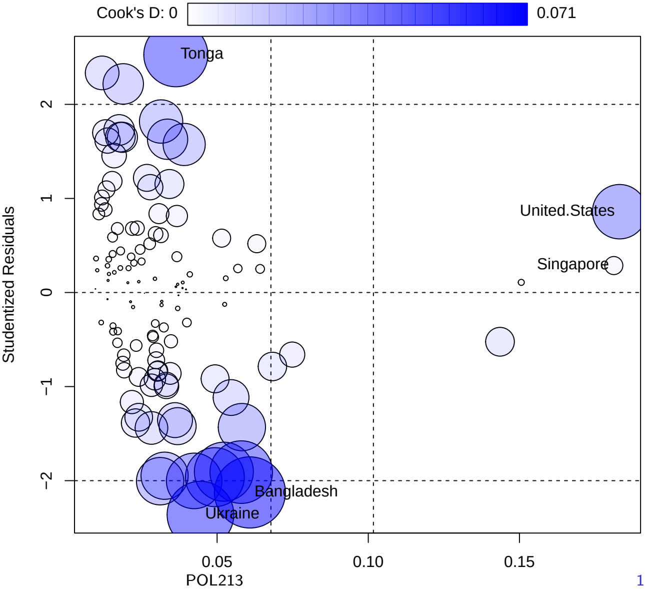
$$D_i = \frac{E_i'^2}{k+1} \times \frac{h_i}{1-h_i}$$

where the first term, the standardized residual, is a measure of discrepancy and the second is a measure of leverage.

There are other statistics (DFBETA) that can accomplish a similar task.

Measuring Influence

```
>influencePlot(reg4)
```



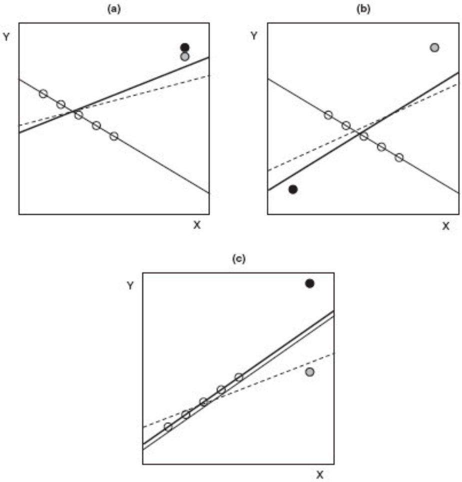
Joint Influence

Sometime, subsets of observations can be *jointly influential* or can offset each other's influence. While one could systematically delete influential points and rerun the models, the large number of possible subsets renders this approach impractical. Instead, we use *added variable plots*, also called *partial regression plots*.

Does the presence of jointly influential points tell us something about our data generating process or modeling approach?

Joint Influence

In each graph, the heavier solid line gives the least-squares regression for all of the data, the broken line gives the regression with the black circle deleted, and the lighter solid line gives the regression with both the black circle and the gray circle deleted. (a) Jointly influential observations located close to one another: Deletion of both observations has a much greater impact than deletion of only one. (b) Jointly influential observations located on opposite sides of the data. (c) Observations that offset one another: The regression with both observations deleted is the same as for the whole data set (the two lines are separated slightly for visual effect).



Source: Fox fig. 11-8.

Added Variable Plots

The approach to examining influential groups of data in the AV plot is as follows:

- Regress Y on all the X's with the exception of X_1 and obtain the residuals ($Y_i^{(1)}$)
- Likewise, regress X_1 on all other Xs and obtain the residuals ($X_i^{(1)}$)

Then the residuals $Y_i^{(1)}$ and $X_i^{(1)}$ have the following properties:

(i.) the slope from the LS regression of $Y_i^{(1)}$ on $X_i^{(1)}$ is the least squares slope B_1 for the full multiple regression.[1]

(ii.) the residuals from the simple regression of $Y_i^{(1)}$ on $X_i^{(1)}$ is the same as those from the full regression

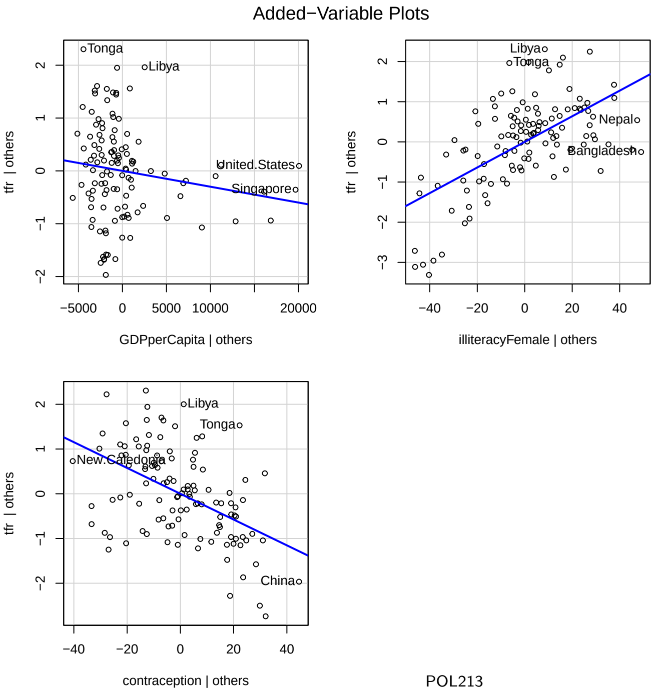
(iii.) the standard error of B_i is the same as the multiple regression standard error

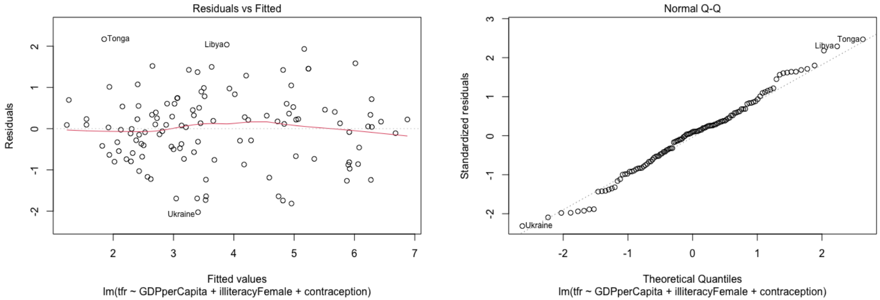
As a result, we can plot $Y_i^{(1)}$ against $X_i^{(1)}$ to examine leverage and influence of the observations of B_1 and visualize easily. The AV plots are useful for detecting the joint influence of groups of observations on the regression coefficients.

[1] This concept comes into play with instrumental variables

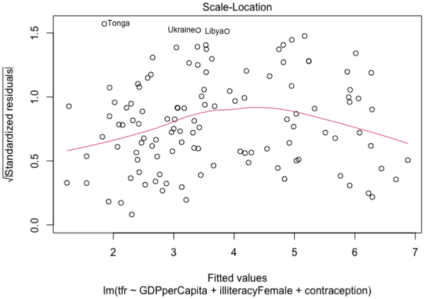
Diagnostics process

```
# added-variable plots
avPlots(reg4, ask=FALSE)

# 4 plots: Residual vs Fitted,
# Normal Probability Plot, Scale-Location,
# Residual vs Leverage
plot(reg4)
```


Added−Variable Plots
tfr | others
GDPperCapita | others
−5000
0
5000
10000
20000
−2
−1
0
1
2
Tonga
Libya
United.States
Singapore
tfr | others
illiteracyFemale | others
−40
−20
0
20
40
−3
−2
−1
0
1
2
Libya
Tonga
Nepal
Bangladesh
tfr | others
contraception | others
−40
−20
0
20
40
−2
−1
0
1
2
Libya
Tonga
New.Caledonia
China


Residuals vs Fitted
Residuals
2
1
0
-1
-2
Tonga
Libya
Ukraine
2
3
4
5
6
7
Fitted values
lm(tfr ~ GDPperCapita + illiteracyFemale + contraception)
Normal Q-Q
Standardized residuals
3
2
1
0
-1
-2
Tonga
Libya
Ukraine
-2
-1
0
1
2
Theoretical Quantiles
lm(tfr ~ GDPperCapita + illiteracyFemale + contraception)

Unusual and Influential Data



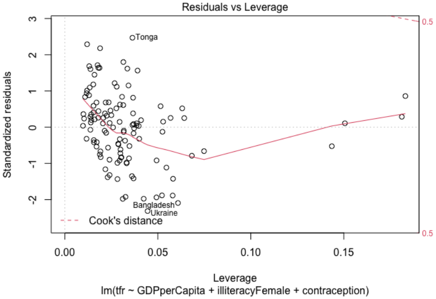


Table of Contents

Unusual and Influential Data

Diagnostics for Errors and Linearity

Summary

The Gauss-Markov theorem

Recall, we want to (1) minimize both bias (residuals) and sampling variance. And we want to (2) allow for statistical inference, modeling uncertainty in estimates. OLS is good at accomplishing these things provided key assumptions of the Gauss Markov theorem are met:

- The first two (linearity and nonstochastic regressors) relate to bias.
- The next two (homoskedasticity and independence of errors) relate to variance.
- The above four assumptions are sufficient to prove OLS is BLUE (the Gauss-Markov theorem).
- We can add one more assumption (normality of errors) to complete the strong set and make statistical inference using the normal distribution.

OLS is unbiased if. . .

1. The functional form is correct, or the disturbance has a conditional mean of zero: $E(u_i|X_{1i}, X_{2i}, \ldots, X_{ki}) = 0$, $\forall i$.
2. The regressors are fixed, exogenous, or independent of the disturbance term: $cov(X_{1i}, u_i) = cov(X_{2i}, u_i) = \cdots = cov(X_{ki}, u_i) = 0$.

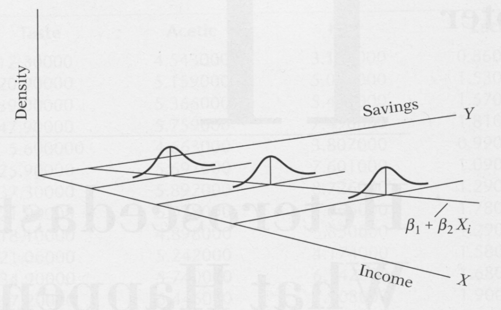
- Violating either of these two Gauss-Markov assumptions causes bias.

Possible Violations: Specification Errors

1. Omission of a relevant variable/wrong functional form
2. Errors of measurement
3. Incorrect specification of the error term (note: nonlinear regression)
4. Reciprocal causation (note: instrumental variables)

Homoskedasticity and independent errors

These don't deal with bias, but violations do means the estimates are no longer efficient (i.e., lowest variance of sampling distributions among all estimators in class)



And finally. . .

- The assumption of normal disturbances completes the strong set: OLS is now the best unbiased estimator among all (not just linear) estimators (i.e., also the MLE estimate).
- The sampling variances of the estimators are now **normal**, allowing for easy inference.
- Not possible if we're dealing with discrete dependent variables.
- **BUT**, no such restriction on the independent/predictor variables.

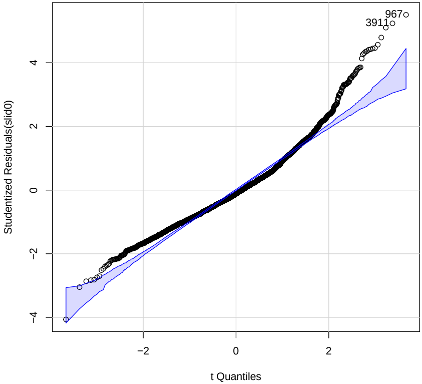
Assessing non-normality

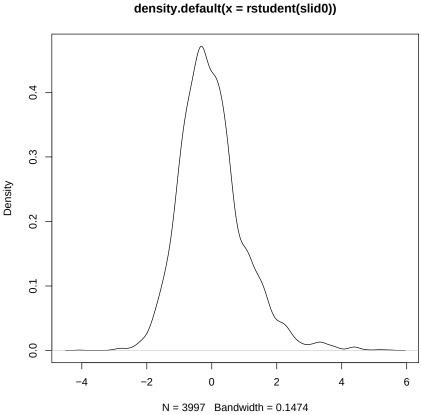
- Non-normal errors: OLS no longer necessarily MLE.
- Can assess by looking at (studentized) residuals: quantile-comparison plots and density of studentized residuals
- Examples of residual problems we can detect: highly skewed error distribution, multimodal error distribution

Non-normal errors

```
SLID <- read.table(
"http://socserv.socsci.mcmaster.ca/
jfox/Books/Applied-Regression-2E/
datasets/SLID-Ontario.txt",
header=TRUE)

slid0 <- lm(compositeHourlyWages ~ sex + age + yearsEducation,
data=SLID)
summary(slid0)

qqPlot(slid0, simulate=TRUE, line="none", labels=FALSE)

plot(density(rstudent(slid0)))
```

Diagnostics for Errors and Linearity



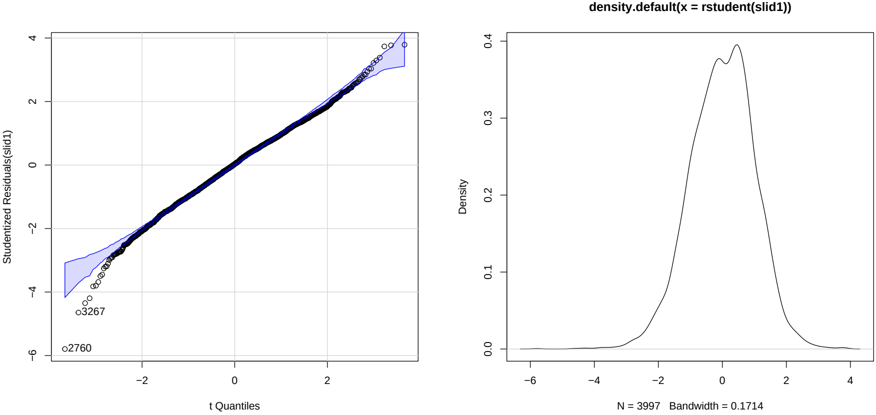


And after log transforming the DV

To handle skewness in the DV, we can log transform it with a suitable base. Here, base 2 seems to work well.

```
#Visualize distribution
hist(SLID$compositeHourlyWages)
hist(log2(SLID$compositeHourlyWages))

slid1 <- lm(log2(compositeHourlyWages) ~
sex + age + yearsEducation, data=SLID)
summary(slid1)

qq.plot(slid1, simulate=TRUE, line="none", labels=FALSE)
plot(density(rstudent(slid1)))
boxplot(rstudent(slid1))
```

Diagnostics for Errors and Linearity



Assessing heteroskedasticity

Key assumption of regression model is that variation of the response variable around the regression surface is everywhere the same:

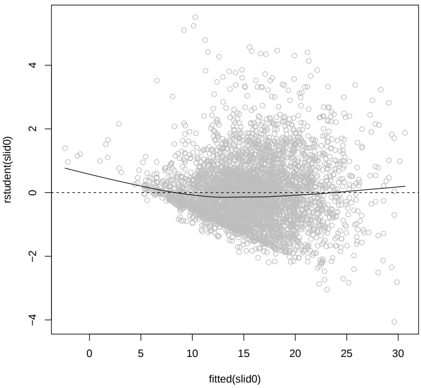
$$V(\varepsilon) = V(Y|x_1, x_2, ..., x_k) = \sigma_\varepsilon^2$$

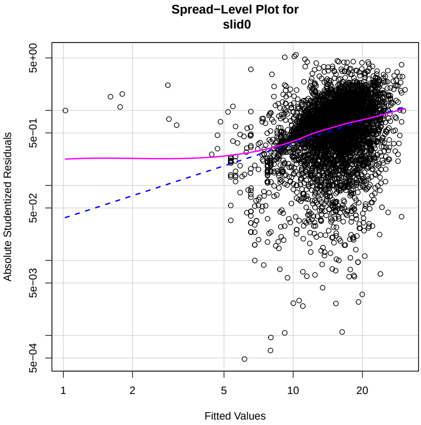
- Non-constant error variance: bad standard errors. (Huber-White/robust SEs)
- Arises from incorrect specification (omitting an important effect on Y).
- Fitted vs. residual plots:

Residual plot

```
plot(fitted(slid0), rstudent(slid0), col="gray")
abline(h=0, lty=2)
lines(lowess(fitted(slid0), rstudent(slid0)))
spread.level.plot(slid0)
```

Diagnostics for Errors and Linearity

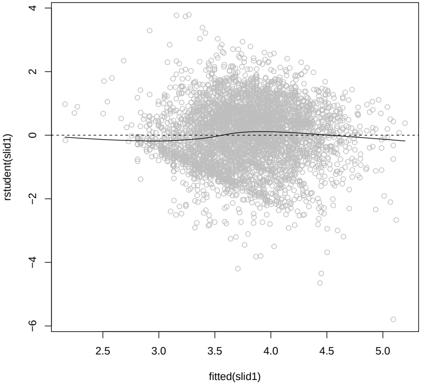


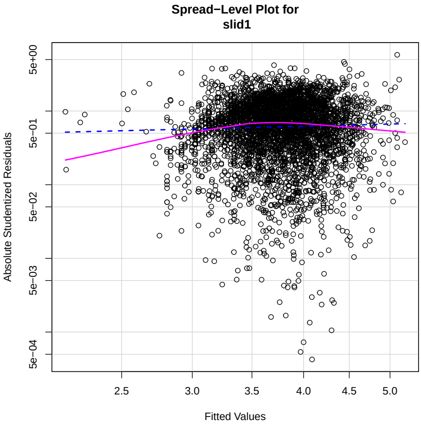


And after log transforming the DV

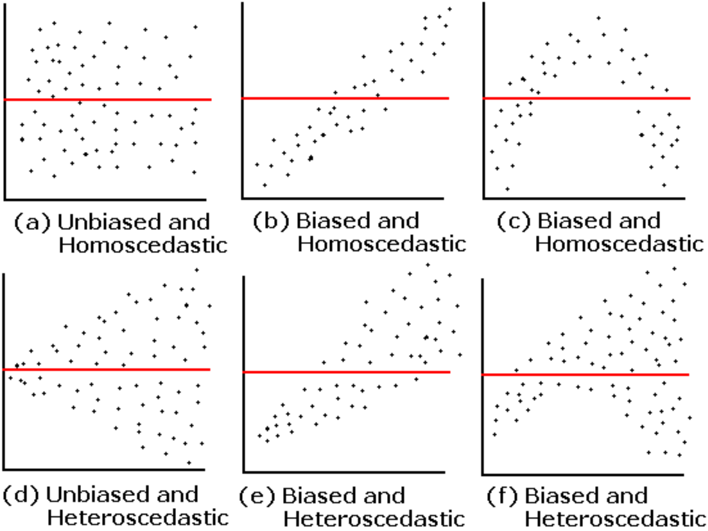
```
plot(fitted(slid1), rstudent(slid1), col="gray")
abline(h=0, lty=2)
lines(lowess(fitted(slid1), rstudent(slid1)))
spreadLevelPlot(slid1)
```

Diagnostics for Errors and Linearity





Residual plots - Examples



(a) Unbiased and Homoscedastic

(b) Biased and Homoscedastic

(c) Biased and Homoscedastic

(d) Unbiased and Heteroscedastic

(e) Biased and Heteroscedastic

(f) Biased and Heteroscedastic

Correcting OLS Standard Errors for Nonconstant Variance

- We often wish to not only diagnose but to correct for OLS standard errors with nonconstant variance.
- One common approach is to use a Weighted Least Squares estimator in which weights are assigned to observations. Greater weight is assigned to observations with smaller variance. (A practical example is here.)
- Another common approach is to modify the s.e. with White's correction (sometimes called a sandwich estimator).
- An advantage of White's approach for comping with heteroskedasticity is that knowledge of the pattern of nonconstant error variance (e.g. increased variance with the level of Y or with an X is not required).

Correcting OLS Standard Errors for Nonconstant Variance

The covariance matrix of the OLS estimator is:

$$V(\boldsymbol{b}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'V(\boldsymbol{y})X(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

Under the standard assumption of constant error variance, we have $V(\boldsymbol{y}) = \sigma_\varepsilon^2 \boldsymbol{I}_n$ and the above equation simplifies to the usual formula, $V(\boldsymbol{b}) = \sigma_\varepsilon^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

However, if the errors are heteroskedastic but independent, then $\Sigma \equiv V(\boldsymbol{y}) = \mathrm{diag}\{\sigma_1^2, \sigma_2^2, ..., \sigma_n^2\}$ and

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

Because $E(\varepsilon_i) = 0$, the variance of the ith error is $\sigma_i^2 = E(\varepsilon_i^2)$ and then we can estimate that middle term with $\hat{\Sigma} = \mathrm{diag}\{E_1^2, E_2^2, ..., E_n^2\}$ where E_i is the OLS residual for observation i.

White (1980) showed that a sandwich estimator using this approximation provides a consistent estimator of $V(\boldsymbol{b})$. The "meat" of the sandwich can also be adjusted by the hat-values of observation i for a modified estimate.

Robust and Clustered Standard Errors

A thorough and excellent discussion of corrections for non-constant variance of errors is available at the following URL. Strongly recommended for methods-track students.

https://projects.iq.harvard.edu/files/gov2001/files/sesection_print.pdf

A related problem is if your errors are dependent on one another. Then the "meat" of the sandwich should be adjusted to reflect clustering or grouping, i.e. the off-diagonal terms of the Σ matrix are nonzero. In this case, we use a clustered standard error correction.

You may see the term "clustered-robust standard errors." This is when both types of corrections are used (1) for heteroskedasticity, as revealed by nonconstant error variance along diagonal terms and (2) clustering / grouping in the data as revealed dependent error terms, i.e. nonzero off-diagonal terms.

Heteroskedasticity

Common for the variance of errors to increase with the level of the response variable. Often detected in plot of residuals against fitted values (following fan shape).

Strategies for dealing with nonconstant error variance include:

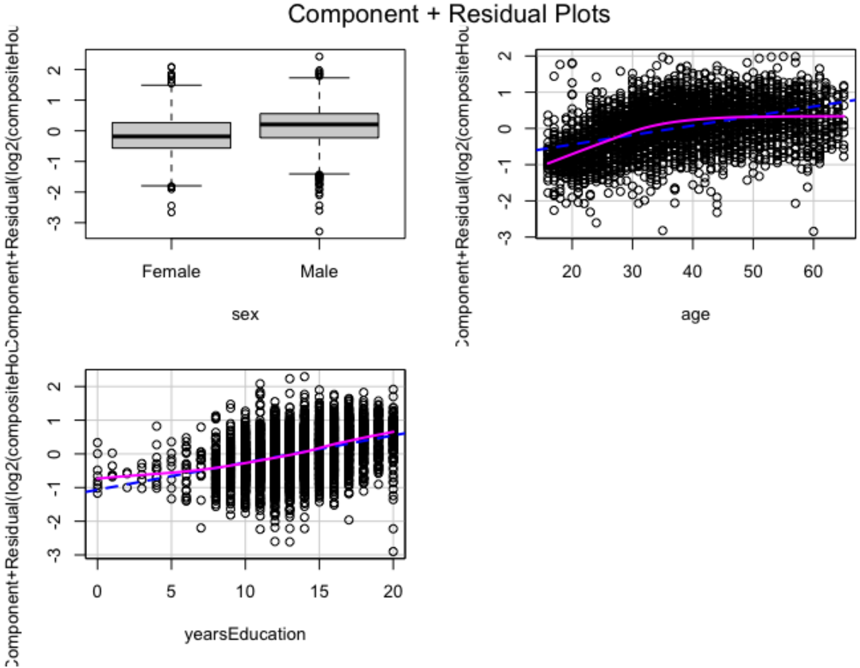
- transformations of the response variable to stabilize the variance
- substitution of the **Weighted Least Square** estimation for OLS

Assessing nonlinearity

- Nonlinearity: Violation of GM assumption that $E(\varepsilon) = 0$ everywhere.
- Often arises if we are missing interaction or polynomial terms.
- **Component-plus-residual plots** plot X_j against $\varepsilon + \hat{\beta}_j X$

Component-plus-residual plot

```
crPlots(slid1, ask=FALSE)
```


Component + Residual Plots
Component+Residual(log2(compositeHou
Female
Male
sex
20
30
40
50
60
age
0
5
10
15
20
yearsEducation
-3
-2
-1
0
1
2

```
slid2 <- lm(log2(compositeHourlyWages) ~ sex + poly(age, degree=2)
  + I(yearsEducation^2), data=SLID)
# orthogonal polynomial in age
summary(slid2)

lm(log2(compositeHourlyWages) ~ sex + poly(age, degree=2, raw=TRUE)
  + I(yearsEducation^2), data=SLID) # "raw" quadratic

lm(log2(compositeHourlyWages) ~ sex + age + I(age^2)
  + I(yearsEducation^2), data=SLID) # equivalent

crPlots(slid2, ask=FALSE)

# effects plot on original dollar scale
plot(allEffects(slid2, transformation=list(link=log2, inverse=function(x) 2^x)),
  ylab="composite hourly wages", ask=FALSE)
```

Assessing nonlinearity

Simple forms of nonlinearity can often be detected in component-plus-residual plots.

- Nonlinearity can frequently be accommodated by variable transformations (e.g. log of base *m*, box-cox)
- Sometimes altering the form of the model (e.g. adding quadratic term) can also improve model fit
- Component plus residual plots reliably reflect nonlinearity when there are not strong nonlinear relationships among the explanatory variables in a regression

Table of Contents

Unusual and Influential Data

Diagnostics for Errors and Linearity

Summary

Linear Model Diagnostics and Assessment

Goal is to ensure Gauss-Markov Assumptions are satisfied

1. Examine data for influential outliers
   - Many handy tests: measure leverage using hat values; find outliers using studentized residuals; evaluate influence using DFBETA and Cook's distance.
   - Are there influential subsets leading us to rethink our model? Can we ever drop observations?
2. Evaluate whether error terms are normally distributed.
   - While LS estimator remains valid (unbiased, consistent) with non-normal errors, it becomes *inefficient*, leading to mistaken inferences.
   - Nonconstant error variance = heteroskedasticity. Residual plots can reveal these problems.
   - Can correct for heteroskedasticity with Weighted Least Squares or White adjustment to standard errors.
3. Evaluate whether model is linear, satisfying the assumption $E(\varepsilon)$ everywhere is 0.
   - Component-plus-residual plots can be helpful in diagnosing nonlinearity in the partial relationship between Y and X.
   - Variable transformations, interaction terms, polynomials, etc, are all potential remedies.

Discussion of Final Projects.