Regression III
Outliers, Heteroskedasticity, Inference

Dave Armstrong

University of Western Ontario
Department of Political Science
Department of Statistics and Actuarial Science (by courtesy)

e: dave.armstrong@uwo.ca
w: www.quantoid.net/teachicpsr/regression3/

July 20, 2017

Types of Unusual Observations (1)

- An observation that is unconditionally unusual in either its Y or X value is called a univariate outlier, but it is not necessarily a regression outlier
- A regression outlier is an observation that has an unusual value of the outcome variable Y, conditional on its value of the explanatory variable X
    - In other words, for a regression outlier, neither the X nor the Y value is necessarily unusual on its own
- Regression outliers often have large residuals but do not necessarily affect the regression slope coefficient
- Also sometimes referred to as vertical outliers

Types of Unusual Observations (2)

- An observation that has an unusual X value - i.e., it is far from the mean of X - has leverage on the regression line
    - The further the outlier sits from the mean of X (either in a positive or negative direction), the more leverage it has
- High leverage does not necessarily mean that it influences the regression coefficients
    - It is possible to have a high leverage and yet follow straight in line with the pattern of the rest of the data. Such cases are sometimes called “good” leverage points because they help the precision of the estimates. Remember, $V(B) = \sigma^2_\varepsilon(\mathbf{X'X})^{-\mathbf{1}}$, so outliers could increase the variance of X.

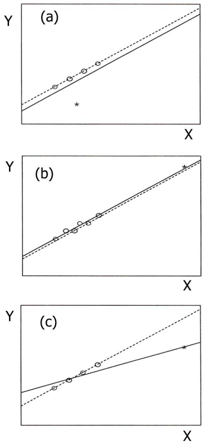
Types of Unusual Observations (3)

- An observation with high leverage that is also a regression outlier will strongly influence the regression line
    - In other words, it must have an unusual X-value with an unusual Y-value given its X-value
- In such cases both the intercept and slope are affected, as the line chases the observation

Discrepancy × Leverage = Influence

Types of Unusual Observations (4)

- Figure (a): Outlier without influence. Although its Y value is unusual given its X value, it has little influence on the regression line because it is in the middle of the X-range
- Figure (b) High leverage because it has a high value of X. However, because its value of Y puts it in line with the general pattern of the data it has no influence
- Figure (c): Combination of discrepancy (unusual Y value) and leverage (unusual X value) results in strong influence. When this case is deleted both the slope and intercept change dramatically.



Adapted from Figure 11.1 (Fox, 1997)

Assessing Leverage: Hat Values (2)

- If h_{ij} is large, the i^{th} observation has a substantial impact on the j^{th} fitted value
- Since $\boldsymbol{H}$ is symmetric and idempotent, the diagonal entries represent both the i^{th} row and the i^{th} column:

$$\begin{aligned} h_i &= \boldsymbol{h_i'h}_i \\ &= \sum_{i=1}^{n} h_{ij}^2 \end{aligned}$$

- This means that $h_i = h_{ii}$
- As a result, the hat value h_i measures the *potential leverage of Y_i on all the fitted values*

Properties of Hat Values

- The average hat value is: $\bar{h} = \frac{k+1}{n}$
- The hat values are bound between $\frac{1}{n}$ and 1
- In simple regression hat values measure distance from the mean of X:

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

- In multiple regression, h_i measures the distance from the centroid point of all of the X's (point of means)
- Commonly used Cut-off:
  - Hat values exceeding about twice the average hat-value should be considered noteworthy
  - With large sample sizes, however, this cut-off is unlikely to identify any observations regardless of whether they deserve attention

Assessing Leverage: Hat Values (2)

- If h_{ij} is large, the i^{th} observation has a substantial impact on the j^{th} fitted value
- Since $\boldsymbol{H}$ is symmetric and idempotent, the diagonal entries represent both the i^{th} row and the i^{th} column:

$$\begin{aligned} h_i &= \boldsymbol{h_i'h}_i \\ &= \sum_{i=1}^{n} h_{ij}^2 \end{aligned}$$

- This means that $h_i = h_{ii}$
- As a result, the hat value h_i measures the *potential leverage of Y_i on all the fitted values*

Properties of Hat Values

- The average hat value is: $\bar{h} = \frac{k+1}{n}$
- The hat values are bound between $\frac{1}{n}$ and 1
- In simple regression hat values measure distance from the mean of X:

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

- In multiple regression, h_i measures the distance from the centroid point of all of the X's (point of means)
- Commonly used Cut-off:
  - Hat values exceeding about twice the average hat-value should be considered noteworthy
  - With large sample sizes, however, this cut-off is unlikely to identify any observations regardless of whether they deserve attention

Hat Values in Multiple Regression

- The diagram to the right shows elliptical contours of hat values for two explanatory variables
- As the contours suggest, hat values in multiple regression take into consideration the correlational and variational structure of the X's
- As a result, outliers in multi-dimensional X-space are high leverage observations - i.e., the outcome variable values are irrelevant in calculating h_i

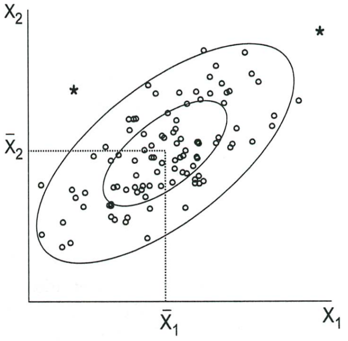


Figure 11.3 from Fox (1997)

Studentized Residuals (1)

- If we refit the model deleting the i^{th} observation we obtain an estimate of the standard deviation of the residuals $S_{E(-1)}$ (standard error of the regression) that is based on the $n - 1$ observations
- We then calculate the studentized residuals E_i^*'s, which have an independent numerator and denominator:

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

  Studentized residuals follow a t-distribution with $n - k - 2$ degrees of freedom
- We might employ this method when we have several cases that might be outliers
- Observations that have a studentized residual outside the ± 2 range are considered statistically significant at the 95% level

Studentized Residuals (2)

- An alternative, but equivalent, method of calculating studentized residuals is the so-called *'mean-shift' outlier model*:

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \gamma D_i + \varepsilon_i$$

  where D is a dummy regressor coded 1 for observation i and 0 otherwise
- We test the null hypothesis that the outlier i does not differ from the rest of the observations, $H_0 : \gamma = 0$, by calculating the t-test:

$$t_0 = \frac{\tilde{\gamma}}{\widehat{SE}(\tilde{\gamma})}$$

- The test statistic is the studentized residual E_i^* and is distributed as t_{n-k-2}
- This method is most suitable when, after looking at the data, we have determined that a particular case might be an outlier

Influential Observations: Cook’s D (1)

- Cook’s D measures the ‘distance’ between B_j and $B_{j(-i)}$ by calculating an F-test for the hypothesis that $b_j = B_{j(-i)}$, for $j = 0, 1, \ldots, k$. An F-test is calculated for each observation as follows:
$$D_i = \frac{E_i'^2}{k+1} \times \frac{h_i}{1-h_i}$$
where h_i is the hat value for each observation and E_i' is the standardized residual
- The first fraction measures discrepancy; the second fraction measures leverage
- There is no significance test for D_i (i.e., the F-test here measures only distance) but a commonly used cut-off is:
$$D_i > \frac{4}{n-k-1}$$
- The cut-off is useful, but there is no substitution for examining relative discrepancies in plots of Cook’s D versus cases, or of E_i' against h_i

Unusual Cases: Solutions

- Unusual observations may reflect miscoding, in which case the observations can be rectified or deleted entirely
- Outliers are sometimes of substantive interest:
  - If only a few cases, we may decide to deal separately with them
  - Several outliers may reflect model misspecification - i.e., an important explanatory variable that accounts for the subset of the data that are outliers has been neglected
- Unless there are strong reasons to remove outliers we may decide to keep them in the analysis and use alternative models to OLS, for example robust regression, which down weight outlying data.
  - Often these models give similar results to an OLS model that omits the influential cases, because they assign very low weight to highly influential cases

Defining “Robust”

- Statistical inferences are based both on observations and on prior assumptions about the underlying distributions and relationships between variables
  - Although the assumptions are never *exactly* true, some statistical models are more sensitive to small deviations from these assumptions than others
- Following Huber (1981) *robustness* signifies insensitivity to deviations from the assumptions the model imposes
  - A model is robust then, if it is (1) reasonably efficient and unbiased, (2) small deviations from model assumptions will not substantially impair the performance of the model and (3) somewhat larger deviations will not invalidate the model completely
- Robust regression is concerned with distributional robustness and outlier resistance
  - Although conceptually distinct, these are for practical purposes synonymous

Breakdown Point (1)

- Assume a sample, $\boldsymbol{Z}$, with n observations, and let $\boldsymbol{T}$ be a regression estimator.
- Applying $\boldsymbol{T}$ to $\boldsymbol{Z}$ gives us the vector of regression coefficients:
$$\boldsymbol{T}(\boldsymbol{Z}) = \hat{\boldsymbol{\beta}}$$
- Imagine all possible “corrupted” samples $\boldsymbol{Z}'$ that replace any observations m of the dataset with arbitrary values (i.e., influential cases)
- The maximum bias that could arise from these substitutions is:
$$\text{effect}(m; \boldsymbol{T}, \boldsymbol{Z}) = \sup_{\boldsymbol{Z}'} \| \boldsymbol{T}(\boldsymbol{Z}') - \boldsymbol{T}(\boldsymbol{Z}) \|$$
where the supremum is over all possible $\boldsymbol{Z}'$

Breakdown Point (2)

- if the effect$(m; \boldsymbol{T}, \boldsymbol{Z})$ is infinite, the m outliers have an arbitrarily large effect on $\boldsymbol{T}$
- The breakdown point for an estimator $\boldsymbol{T}$ for a finite sample $\boldsymbol{Z}$ is:

$$BDP(T, Z) = min\left\{\frac{m}{n}; \text{effect}(m; \boldsymbol{T}, \boldsymbol{Z})\text{is infinite}\right\}$$

- In other words, the breakdown point is the smallest fraction of "bad" data (outliers or data grouped in the extreme tail of the distribution) the estimator can tolerate without taking on values arbitrarily far from $\boldsymbol{T}(\boldsymbol{Z})$
  - For OLS regression, one unusual case is enough to influence the coefficient estimates. Its breakdown point then is:

$$BDP = \frac{1}{n}$$

- As n gets larger, $\frac{1}{n}$ tends toward 0, meaning that the breakdown point for OLS is 0%

Influence Function (or Influence Curve)

- While the breakdown point measures global robustness, the influence function (IF) measures local robustness
- More specifically, the IF measures the impact of a single observation Y that contaminates the theoretically assumed distribution F on an estimator T

$$\text{IF}(Y, F, T) = \lim_{\lambda \to 0} \frac{T\{(1 - \lambda)F + \lambda\delta_Y\} - T\{F\}}{\lambda}$$

  where δ_Y is the probability distribution that puts its mass at the point Y (i.e., δ_Y=1 at Y and 0 otherwise), and λ is the proportion of contamination at Y
- Simply put, the IF indicates the bias caused by adding outliers at the point Y, standardized by the proportion of contamination
- The IF can be calculated from the first derivative of the estimator

M-Estimation for Regression

- OLS minimizes the sum of squares function

$$\text{min} \sum_{i=1}^{n} (E_i)^2$$

- Following from M-estimation of location, a robust regression M-estimate minimizes the sum of a less rapidly increasing function ρ of the residuals

$$\text{min} \sum_{i=1}^{n} \rho(E_i)$$

- Since the solution is not scale invariant. the residuals must be standardized by a robust estimate of their scale, σ_ε, which is estimated simultaneously. Usually, the median absolute deviation is used:

$$\text{min} \sum_{i=1}^{n} \rho\left(\frac{E_i}{\hat{\sigma}_E}\right)$$

M-estimation for Regression (2)

- Taking the derivative and solving produces the shape of the influence function:

$$\sum_{i=1}^{n} \Psi\left(\frac{E_i}{\hat{\sigma}_E}\right) x_i, \text{where } \psi = \rho'$$

- We then substitute Ψ with an appropriate weight function

$$\sum_{i=1}^{n} w_i \left(\frac{E_i}{\hat{\sigma}_E}\right) x_i$$

- Typically the Huber or bisquare weight is employed. In other words, the solution assigns a different weight to each case depending on the size of its residual and thus minimizes the weighted sum of squares

$$\sum_{i=1}^{n} w_i E_i^2$$

M-Estimation and Regression (3)

- Since the weights cannot be estimated before fitting the model and estimates cannot be found without the weights, an iterative procedure is needed to find estimates
- Initial estimates of $\boldsymbol{b}$ are selected using weighted least squares
- The residuals from this model are used to calculate an estimate of the scale of the residuals $\sigma_e^{(0)}$ and the weights $w_i^{(0)}$
- The model is then refit with several iterations minimizing the weighted sum of squares to obtain new estimates of $\boldsymbol{b}$

$$\boldsymbol{b}^{(l)} = (\boldsymbol{X'WX})^{-1}\boldsymbol{X'Wy}$$

- where l is the iteration counter; in the i^{th} row of the model matrix are x_i' and $\boldsymbol{W} \equiv diag\{w_i^{l-1}\}$
- This process continues until the model converges ($\boldsymbol{b}^{(l)} \simeq \boldsymbol{b}^{(l-1)}$)

Asymptotic Standard Errors

- For all M-estimators (including the MM-estimator), asymptotic standard errors are given by the square root of the diagonal entries of the estimated asymptotic covariance matrix $(\boldsymbol{X'WX})^{-1}\sigma_E^2$ from the final IWLS fit
- The ASE for a particular coefficient, is then given by:

$$SE\hat{\beta} = \sqrt{\frac{\sum[\boldsymbol{W}(E_i)]^2}{[\sum \boldsymbol{W}'(E_i)/n]^2}(\boldsymbol{X'X})^{-1}}$$

- The ASEs are relaibale if the sample size n is sufficiently large relative to the number of parameters estimated
  - Other evidence suggests that their reliability also decreases as the proportion of influential cases increases
- As a result, if $n < 50$ bootstrapping should be considered

Combining Resistance and Efficiency: MM-estimation (1)

- MM estimation is perhaps the most commonly employed method today
- "MM" in the name refers to the fact that more than one M-estimation procedure is used to calculate the final estimates
- Combine a high breakdown point (50%), bounded influence function and high efficiency under normal errors (≈ 95%)

Steps to MM-estimation (1)

1. Initial estimates of the coefficients $\boldsymbol{B}^{(1)}$ and corresponding residuals $e_i^{(1)}$ are taken from a highly resistant robust regression (i.e., a regression with a breakdown point of 50%)
   - Although the estimator must be consistent, it is not necessary that it is efficient. As a result, S-estimation with Huber or Bisquare weights is typically employed here
2. The Residuals $E_i^{(1)}$ from the S-estimation stage 1 are used to compute an M-estimation of the scale of the residuals
3. Finally, the initial estimates of the residuals $e_i^{(1)}$ from stage 1 and of the residual scale σ_E from stage 2 are used to compute a single-step M-estimate

$$\sum_{i=1}^{n} w_i \left(\frac{E_i^{(1)}}{\hat{\sigma}_E}\right) x_i$$

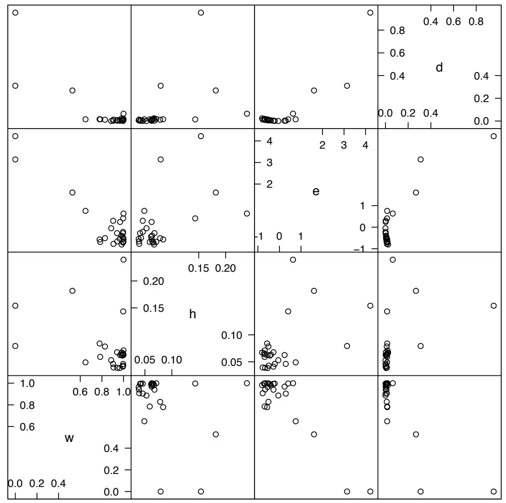
   where the w_i are typically Huber or bisquare weights. In other words, the M-estimation procedure at this stage needs only a single iteration of weighted least squares

MM-estimation in **R**

```
weakliem <- read.table(
    "http://quantoid.net/files/reg3/weakliem2.txt",
    sep=",", header=T, row.names=1)
library(MASS)
mod5 <- rlm(secpay ~ gini, data=weakliem,
        method="MM")
summary(mod5)

##
## Call: rlm(formula = secpay ~ gini, data = weakliem, method = "
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.7533 -2.3654  0.4203  4.6373 57.8463
##
## Coefficients:
##             Value   Std. Error t value
## (Intercept) -4.5408  4.8936    -0.9279
## gini         0.4561  0.1276     3.5759
##
## Residual standard error: 6.068 on 24 degrees of freedom
```

Comparison of Diagnostics



Scatter Plot Matrix

Non-constant Error Variance

- Also called *Heteroskedasticity*
- An important assumption of the least-squares regression model is that the variance of the errors around the regression surface is everywhere the same: $V(E) = V(Y|x_1, \ldots, x_k) = \sigma^2$.
- Non-constant error variance does not cause biased estimates, but it does pose problems for efficiency and the usual formulas for standard errors are inaccurate
    - OLS estimates are inefficient because they give equal weight to all observations regardless of the fact that those with large residuals contain less information about the regression
- Two types of non-constant error variance are relatively common:
    - Error variance increases as the expectation of Y increases;
    - There is a systematic relationship between the errors and one of the X's

Assessing Non-constant Error Variance

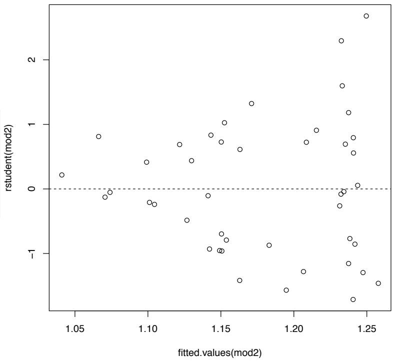
- Direct examination of the data is usually not helpful in assessing non-constant error variance, especially if there are many predictors. Instead, we look to the residuals to uncover the distribution of the errors.
    - It is also not helpful to plot Y against the residuals E, because there is a built-in correlation between Y and E:

$$Y = \hat{Y} + E$$

- The least squares fit ensures that the correlation between $\hat{Y}$ and E is 0, so a plot of these (residual plot) can help us uncover non-constant error variance.
    - The pattern of changing spread is often more easily seen using studentized residuals E_i^{*2} against $\hat{Y}$
- If the values of Y are all positive, we can use a Spread-level plot
    - plot log(| E_i^* |) (called the log spread) against log $\hat{Y}$ (called the log level)
    - The slope b of the regression line fit to this plot suggests the variance-stabilizing transformation $Y^{(p)}$, with $p = 1 - b$

Example Model

```
Weakliem <- read.table(
"http://quantoid.net/files/reg3/weakliem.txt")
W <- Weakliem[-c(21,22,24, 25,49), ]
mod2 <- lm(secpay ~ log(gdp), data=W)
plot(fitted.values(mod2), rstudent(mod2))
abline(h=0, lty=2)
```



Testing for Non-Constant Error variance (1)

- Assume that a discrete X (or combination of X's) partitions the data into m groups.
- Let Y_{ij} denote that ith of n_j outcome-variable scores in group j
- Within-group sample variances are then calculated as follows:

$$S_j^2 = \frac{\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y}_j)^2}{n_j - 1}$$

- We could then compare these within-group sample variances to see if they differ
- If the distribution of the errors is non-normal, however, tests that examine S_j^2 directly are not valid because the mean is not a good summary of the data

Testing for Non-Constant Error variance (2): Score Test

- A score test for the null hypothesis that all of the error variances σ^2 are the same provides a better alternative

1. We start by calculating the standardized squared residuals

$$U_i = \frac{E_i^2}{\hat{\sigma}^2} = \frac{E_i^2}{\frac{\sum E_i^2}{n}}$$

2. Regress the U_i on all of the explanatory variable X's, finding the fitted values:

$$U_i = \eta_0 + \eta_1 X_{i1} + \cdots + \eta_p X_{ip} + \omega_i$$

3. The score test, which s distributed as χ^2 with p degrees of freedom is:

$$S_0^2 = \frac{\sum(\hat{U}_i - \bar{U})^2}{2}$$

R-script testing for non-constant error variance

- The `ncvTest` function in the `car` library provides a simple way to carry out the score test
- The result below shows that the non-constant error variance is statistically significant

```
library(car)

## Warning:  package 'car' was built under R version 3.4.1

ncvTest(mod2, data=W)

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 6.025183    Df = 1     p = 0.01410317

ncvTest(mod2, var.formula=~log(gdp), data=W)

## Non-constant Variance Score Test
## Variance formula: ~ log(gdp)
## Chisquare = 6.025183    Df = 1     p = 0.01410317
```

Generalized Least Squares (1)

- Sometimes, we do not know the relationship between $\boldsymbol{x}_i$ and $var(u_i|\boldsymbol{x}_i)$.
- In this case, we can use a Feasible GLS model.
- FGLS estimates the weight from the data. That weight is then used in a WLS fashion.

GLS: Steps

1. Regress y on $\boldsymbol{x}_i$ and obtain residuals $\hat{u}_i$.
2. Create $log(\hat{u}_i^2)$ by squaring and then taking the natural log of the OLS residuals from step 1.
3. Run a regression of $log(\hat{u}_i^2)$ on $\boldsymbol{x}_i$ and obtain the fitted values $\hat{g}_i$.
4. Generate $\hat{h}_i = exp(\hat{g}_i)$.
5. Estimate the WLS of y on $\boldsymbol{x}_i$ with weights of $\frac{1}{\hat{h}_i}$.

FGLS Example: Inequality Data

```
W2 <- Weakliem[-c(25,49), ]
mod1.ols <- lm(secpay ~ gini*democrat, data=W2)
aux.mod1 <- update(mod1.ols, log(resid(mod1.ols)^2) ~ .)
h <- exp(predict(aux.mod1))
mod.fgls <- update(mod1.ols, weight=1/h)
with(summary(mod.fgls), printCoefmat(coefficients))


##                  Estimate Std. Error t value  Pr(>|t|)
## (Intercept)     0.9619416  0.0476415 20.1913 < 2.2e-16 ***
## gini            0.0044149  0.0013294  3.3210  0.001836 **
## democrat        0.4662928  0.0806427  5.7822 7.577e-07 ***
## gini:democrat -0.0102999  0.0022211 -4.6374 3.288e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Heteroskedastic Regression in R

Charles Franklin wrote a library of ML techniques in R including heteroskedastic regression. We can do the following:

```
source("http://www.quantoid.net/files/reg3/mlhetreg.r")
X <- model.matrix(mod1.ols)[,-1]
Z <- X
het.mod <- MLhetreg(W2$secpay, X, Z)
summary(het.mod)


##
## Heteroskedastic Linear Regression
##
##   Estimated Parameters
##                  Estimate Std. Error z-value  Pr(>|z|)
## Constant        0.9844411  0.0491585 20.0259 < 2.2e-16 ***
## gini            0.0037411  0.0015136  2.4717 0.0134456 *
## democrat        0.4474368  0.0753627  5.9371 2.901e-09 ***
## gini:democrat -0.0097254  0.0021095 -4.6103 4.020e-06 ***
## ZConstant     -9.9427564  1.4019902 -7.0919 1.323e-12 ***
## gini            0.1038281  0.0359335  2.8894 0.0038592 **
## democrat        7.1919603  1.8088570  3.9760 7.009e-05 ***
## gini:democrat -0.1847861  0.0495987 -3.7256 0.0001948 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-Likelihood:  109.9483
##
## Wald Test for Heteroskedasticity
##   Wald statistic:  16.16162 with  3  degrees of freedom
##                    p= 0.001050662
```

Table: Comparing Models

	OLS	FGLS	HetReg
(Intercept)	0.941	0.962	0.984
	(0.060)	(0.048)	(0.049)
gini	0.005	0.004	0.004
	(0.002)	(0.001)	(0.002)
democrat	0.486	0.466	0.447
	(0.088)	(0.081)	(0.075)
gini:democrat	-0.011	-0.010	-0.010
	(0.002)	(0.002)	(0.002)

Robust Standard Errors (1)

- Robust standard errors can be calculated to compensate for an *unknown* pattern of non-constant error variance
- Robust standard errors require fewer assumptions about the model than WLS (which is better if there is increasing error variance in the level of Y)
  - Robust standard errors do not change the OLS coefficient estimates or solve the inefficiency problem, but do give more accurate p-values.
- There are several methods for calculating heteroskedasticity consistent standard errors (e.g., known variously as White, Eicker or Huber standard errors) but most are variants on the method originally proposed by White (1980).

Robust Standard Errors (2): White's Standard Errors

- The covariance matrix of the OLS estimator is:

$$\begin{aligned} V(\boldsymbol{b}) &= (\boldsymbol{X'X})^{-1}\boldsymbol{X'\Sigma X}(\boldsymbol{X'X})^{-1} \\ &= (\boldsymbol{X'X})^{-1}\boldsymbol{X'}V(\boldsymbol{y})\boldsymbol{X}(\boldsymbol{X'X})^{-1} \end{aligned}$$

- Where $V(\boldsymbol{y}) = \sigma_\varepsilon^2 \boldsymbol{I}_n$ if the assumptions of normality and homoskedasticity are satisfied. The variance simplifies to:

$$V(\boldsymbol{b}) = \sigma_\varepsilon^2(\boldsymbol{X'X})^{-1}$$

- In the presence of non-constant error variance, however, $V(\boldsymbol{y})$ contains nonzero covariance and unequal variances
  - In these cases, White suggests a consistent estimator of the variance that constrains $\boldsymbol{\Sigma}$ to a diagonal matrix containing only squared residuals

Robust Standard Errors (3): White's Standard Errors

- The *heteroskedasticity consistent covariance matrix* (HCCM) estimator is then:

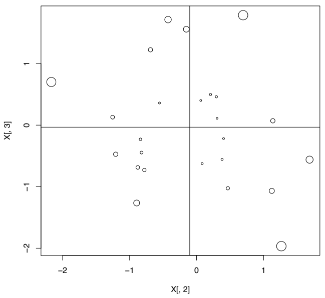
$$V(\boldsymbol{b}) = (\boldsymbol{X'X})^{-1}\boldsymbol{X'}\hat{\boldsymbol{\Phi}}\boldsymbol{X}(\boldsymbol{X'X})^{-1}$$

  where $\hat{\boldsymbol{\Phi}} = e_i^2\boldsymbol{I}_n$ and the e_i are the OLS residuals
- This is what is known as HC0 - White's (1980) original recipe.

Hat Values

Other HCCMs use the “hat value” which are the diagonal elements of $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$

- These give a sense of how far each observation is from the mean of the X’s.
- Below is a figure that shows two hypothetical X variables and the plotting symbols are proportional in size to the hat value

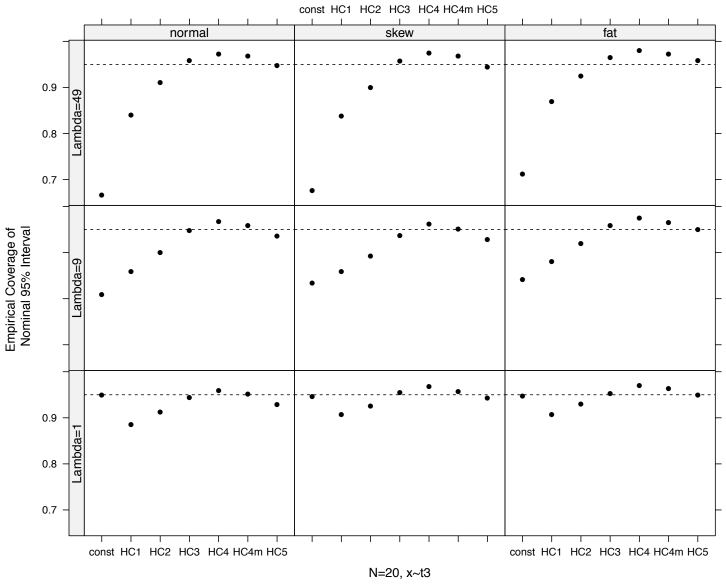


Other HCCM’s

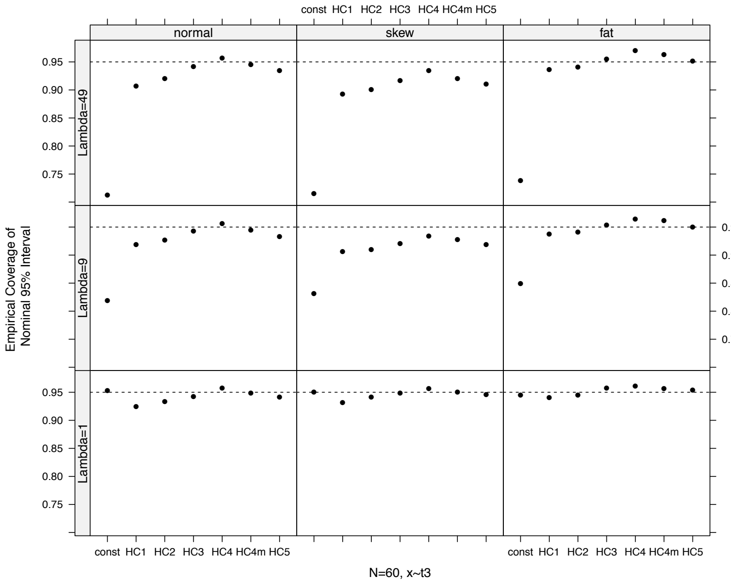
MacKinnon and White (1985) considered three alternatives: HC1, HC2 and HC3, each of which offers a different method for finding $\boldsymbol{\Phi}$.

- HC1: $\frac{N}{N-K}\times$ HC0.
- HC2: $\hat{\boldsymbol{\Phi}}$ = diag$\left[\frac{e_i^2}{1-h_{ii}}\right]$ where $h_{ii}=\boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i'$
- HC3: $\hat{\boldsymbol{\Phi}}$ = diag$\left[\frac{e_i^2}{(1-h_{ii})^2}\right]$
- HC4: $\hat{\boldsymbol{\Phi}}$diag$\left[\frac{e_i^2}{(1-h_{ii})^{\delta_i}}\right]$, where $\delta_i = min\left\{4, \frac{Nh_{ii}}{p}\right\}$
- HC4m: $\hat{\boldsymbol{\Phi}}$ = diag$\left[\frac{e_i^2}{(1-h_{ii})^{\delta_i}}\right]$, where $\delta_i = min\left\{\gamma_1, \frac{nh_{ii}}{p}\right\} + min\left\{\gamma_2, \frac{nh_{ii}}{p}\right\}$, $\gamma_1 = 1$ and $\gamma_2 = 1.5$.
- HC5: $\hat{\boldsymbol{\Phi}}$ = diag$\left[\frac{e_i^2}{(1-h_{ii})^{\delta_i}}\right]$, where $\delta_i = min\left\{\frac{nh_{ii}}{p}, max\left\{4, \frac{nkh_{max}}{p}\right\}\right\}$ with $k = 0.7$

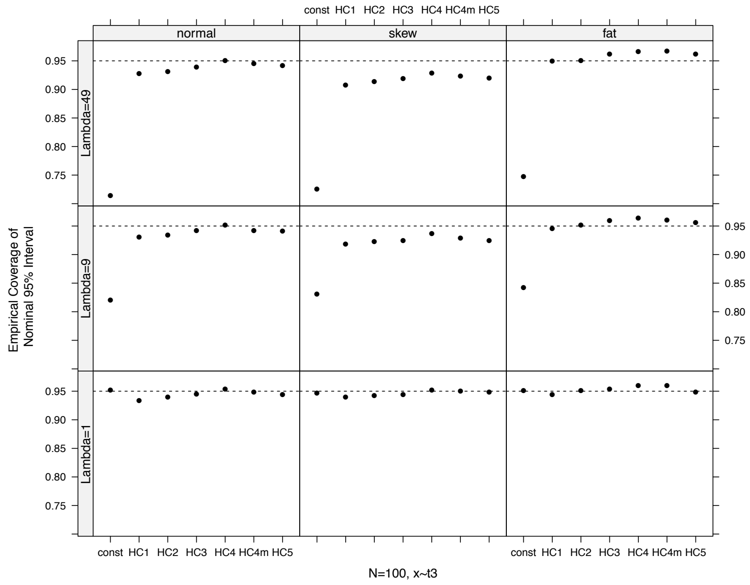
Coverage Percentages for HCCMs (N=20)



Coverage Percentages for HCCMs (N=60)



Coverage Percentages for HCCMs (N=100)



Coverage Percentages for Bootstrap-t Intervals

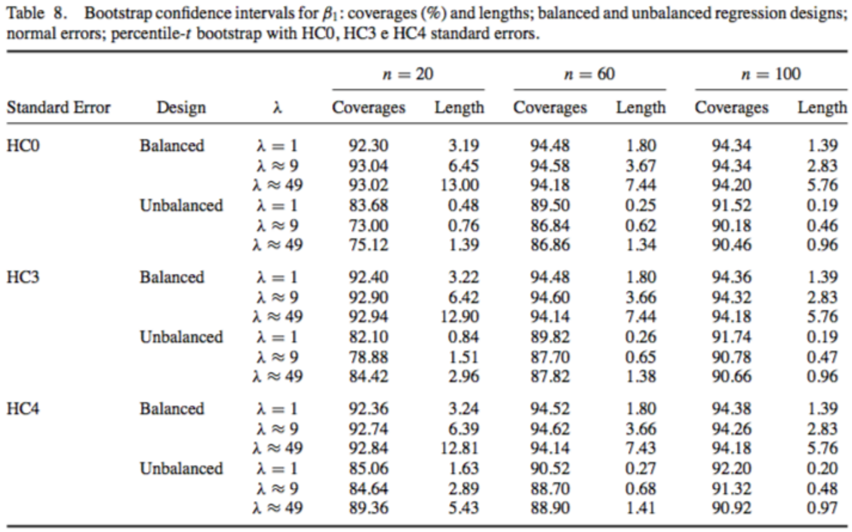


Table 8. Bootstrap confidence intervals for β_1: coverages (%) and lengths; balanced and unbalanced regression designs; normal errors; percentile-t bootstrap with HC0, HC3 e HC4 standard errors.

			$n = 20$		$n = 60$		$n = 100$	
Standard Error	Design	λ	Coverages	Length	Coverages	Length	Coverages	Length
HC0	Balanced	$\lambda = 1$	92.30	3.19	94.48	1.80	94.34	1.39
		$\lambda \approx 9$	93.04	6.45	94.58	3.67	94.34	2.83
		$\lambda \approx 49$	93.02	13.00	94.18	7.44	94.20	5.76
	Unbalanced	$\lambda = 1$	83.68	0.48	89.50	0.25	91.52	0.19
		$\lambda \approx 9$	73.00	0.76	86.84	0.62	90.18	0.46
		$\lambda \approx 49$	75.12	1.39	86.86	1.34	90.46	0.96
HC3	Balanced	$\lambda = 1$	92.40	3.22	94.48	1.80	94.36	1.39
		$\lambda \approx 9$	92.90	6.42	94.60	3.66	94.32	2.83
		$\lambda \approx 49$	92.94	12.90	94.14	7.44	94.18	5.76
	Unbalanced	$\lambda = 1$	82.10	0.84	89.82	0.26	91.74	0.19
		$\lambda \approx 9$	78.88	1.51	87.70	0.65	90.78	0.47
		$\lambda \approx 49$	84.42	2.96	87.82	1.38	90.66	0.96
HC4	Balanced	$\lambda = 1$	92.36	3.24	94.52	1.80	94.38	1.39
		$\lambda \approx 9$	92.74	6.39	94.62	3.66	94.26	2.83
		$\lambda \approx 49$	92.84	12.81	94.14	7.43	94.18	5.76
	Unbalanced	$\lambda = 1$	85.06	1.63	90.52	0.27	92.20	0.20
		$\lambda \approx 9$	84.64	2.89	88.70	0.68	91.32	0.48
		$\lambda \approx 49$	89.36	5.43	88.90	1.41	90.92	0.97

Comparison of Robust and Classical Standard Errors

King and Roberts (2014) suggest that when robust and classical standard errors diverge, it is not an indication that Robust SEs should be used, but that the model is mis-specified.

- A formal test can tell whether the mis-specification is bad enough.

GIM Test

The classical variance-covariance matrix of parameters in MLE is the negative inverse of the Hessian:

- $V_c(\hat{\beta}) = -P^{-1}$

The robust variance-covariance matrix is given by:

- $V_r(\hat{\beta}) = P^{-1}MP^{-1}$ where M is the square of the gradient.
- $V_r(\hat{\beta}) = V_c(\hat{\beta})$ when $M = -P$

A test of model mis-specification comes by evaluating $E(M + P) = 0$ and obtaining sampling variance estimates of the test statistic with a parametric bootstrap.

GIM Test in R

```
source("http://www.quantoid.net/files/reg3/bootstrapim.normal.r")
library(maxLik) # for numericGradient function
```

```
## Loading required package:  miscTools
## Loading required package:  methods
##
## Please cite the 'maxLik' package as:
## Henningsen, Arne and Toomet, Ott (2011).  maxLik:  A package for maximum likelihood
estimation in R. Computational Statistics 26(3), 443-458.  DOI 10.1007/s00180-010-0217-1.
##
## If you have questions, suggestions, or comments regarding the 'maxLik' package, please use
a forum or 'tracker' at maxLik's R-Forge site:
## https://r-forge.r-project.org/projects/maxlik/
```

```
bs.out <- bootstrapIM.normal(formula(mod2), W, 10, 10)
```

```
bs.out
```

```
## $stat
##          [,1]
## [1,] 17.43959
##
## $pval
## [1] 0.5454545
```

Test Results

The test suggests that

- We do have heteroskedasticity according to `ncvTest()`.
- The test suggests that the difference between robust SEs and classical SEs is sufficiently big that the choice between the two is meaningful. Thus, we should think about other ways to re-specify the model.
- Might be some question about whether this advice is strictly necessary in the linear model where the variances are separable.

P-values

As Berk (2004) suggests - one of the fundamental question about statistical inference is - when p-values concern confidence intervals and statistical tests, to what does the probability refer? That is, "probability of what"?

- Assuming X is either fixed by design or considered fixed when the particular set of x values arise in the data, then random sampling results in inferences as one would expect.
- Sampling schemes other than randomness result in rather different properties with respect to inference.

Dealing with Apparent Populations

Below are some methods for dealing with non-random sample selection (particularly when we have non-randomly collected something more like a population). We will talk about each in turn.

1. Treat the data as the population.
2. Treat the data *as if* they were randomly sampled from a population
3. Redefine the population.
4. Invent a population.
5. Model-based sampling.

Treat Data Like a Population

- Description is the only game - the relationship you calculate is the relationship in the population.
- To the extent descriptions are not great (i.e., don't explain a lot of variation), the coefficients you calculate may still not provide insight into the DGP.
- Many problems do not require the frequentist thought experiment of infinite repeated sampling - treating the data as fixed is fine.
  - This might be particularly true in policy situations.

Treat the data *as if* they were randomly sampled from a population

Suggest that the data are approximately randomly sampled from a *real* population.

- Very good data and/or theory required to justify this.
- "Full disclosure" is insufficient - saying you assume your sample is a random sample without evaluation of those assumptions leaves readers not knowing which results to believe.
- Even if the population from which the sample is well-defined, sampling often does not happen in anything close to a random fashion.

Consequences:

- Regression parameters are bad estimates of population parameters (coefficients and variance explained can be attenuated).

Example

Sampling strategy: Collect water samples from a beach every Wednesday around noon over the Summer to measure levels of toxins from storm overflow. Want to infer to all days/times for that beach. Need to make (at least) the following assumptions:

- Toxin concentrations are independent of time of day and day of week.
- The 7-day time gap is sufficient to remove any "memory" (thus observations would be independent).

Data could (and should) be marshaled to provide evidence in favor of these assumptions.

Redefine the Population

Another seemingly reasonable strategy might be to redefine the population post-sampling such that the sample is a random draw from the population.

- Since the missing data occurred after the sampling procedure happened, this is also not appropriate.
- The process that made the data unavailable may be confounded with the relationship of interest.

This makes the justification for the inferential process circular. Convenience samples are, simply, not good fodder for (frequentist) inference.

> *To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of. (Fisher, 1938)*

Invent a Population: Superpopulation

The superpopulation argument is one of the mot prevalent in political science.

- Definition is often circular - the population is the set of other possible circumstances from which this sample could be a random sample.
- Superpopulations are not real and thus not well-defined, so inferences to superpopulations are tenuous at best.
- Some superpopulations *could* exist, but to make inferences to those sorts of populations, the conditions under which the superpopulation exists should be both well-defined and theoretically/empirically justifiable.

Even more problematic is the much of our “population” data suffers from non-random, perhaps non-ignorable missingness.

Superpopulation Example (Berk)

For example, monthly economic indicators from the year 2000 might be treated as “random” realizations of indicators that *could* have existed. Things that need to be specified for this to make sense:

- The data generating process (i.e., a strong and convincing theory, particularly about the random processes leading to this rather than another set of values),
- The conditioning factors pertaining to the observation - what were the particularities of the year 2000 (e.g., exchange rates with other countries, patterns of tariffs and constraints on international trade). These serve as the foundation for the superpopulation.
- Technically, even then you would need to show empirically that there are other years like the year 2000 and that the year 2000 can be meaningfully treated as a sample.

Model-Based Sampling

Propose a model by which nature produces data inferences are made to the model said to be responsible for the data generating process.

- The outcome variables are thought to be random realizations from the model (that is, the model could have generated different values of y).
- Since the outcome variables are random variables, regardless of how many of them we have (even if we have all N of them), each one was one realization of a random process.
- Even if we have the population, uncertainty remains regarding the parameters of the model that generated this (and could generate another) realization.
- The sampling scheme is irrelevant because all observations are assumed to be caused by the same natural process.

Justification of Model-based Sampling

Distributional assumption is key - why would we expect deviations from expectation to be normally distributed?

- CLT says that the sum of a bunch of independent, normal random variables approaches normality.
- The disturbance term in the model can be thought of in such a way as to justify such a distribution.
- A story is still needed to justify this - what sorts of things comprise the disturbance term? Can they really be thought to be independent?

Problems in Model-based Sampling

At best in the literature a hypothesis is done with the null hypothesis being the assumed distribution - failure to reject is taken as evidence the null is true.

- Treats failure to reject “accepting the null”.
- These tests often have low power (leading to fewer rejections of falls null hypotheses than we would like).
- Many different random processes (i.e., distributions) can produce functionally equivalent-looking data.

For inference to make sense, the method that nature uses to make data has to be well-understood and explicated.

Sampling Conclusions

- Remember, we are data analysts/social scientists, not magicians.
- Assumptions, conditions, models required for the frequentist thought experiment of one kind or another to make sense must be reasonable, theoretically justified and empirically evaluated.

What Does “Holding Constant” Mean?

Mathematically:

- A covariance adjustment - removing variance in both x and y that can be explained by z (e.g., partial out the effect of z).

Substantively:

- Independent variables must be *independently manipulable*.
- What would it mean, when predicting income, to hold occupation constant while “manipulating” education?
  - Education is theoretically manipulable - people can gain more education and interventions aimed at such can be undertaken.
  - What would it mean to hold an eventual occupation constant while intervening on education?
  - Post hoc considerations are rather more helpful, but still unsatisfying.

$$\text{Covariance Adjustment} \not\Rightarrow \text{Independent Manipulability}$$

Some Cautions about Statistical Inference

- Models have to be right and sampling procedures well-justified for inference to make sense. Failure on either count means inferences are “right” to a greater or lesser degree.
- p-values will almost always be anti-conservative. Since we rarely (never) take into account model selection uncertainty, total variability is grossly underestimated. Even formal procedures to correct for multiple testing will be insufficient here.
- Inferences should be done on a validation dataset or cross-validation should be used to prevent overfitting and capitalizing on chance through exploration.

Significant vs Not Significant

Gelman and Stern (2006) argue that the distinction between significant and not significant is less interesting than we think.

- Comparing levels of statistical significance is not appropriate.
- When comparing models, more matters than sign and significance.
  - Need to understand whether two estimates are statistically and substantively different from *each other*.

Take Away

- There are many assumptions about our modeling that we don’t (but should) test.
  - Rejecting the null is a necessary, but not sufficient condition for being “right”.
- Pay attention to potential non-linearities. Think about the functional forms of your theories and the extent to which those assumptions are truly justified.
- When you use new tools make sure you understand what they can/can’t or should/shouldn’t be used do.