

d10

Matthew Draper

March 3, 2019

to change: output: pdf_document: keep_tex: yes

```
setwd("C:/Users/mwdra/Google Drive/UCSD/Courses/204B Quantitative Methods/Homework 8")  
library(stargazer)
```

```
## Warning: package 'stargazer' was built under R version 3.5.2
```

```
##  
## Please cite as:
```

```
## Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
## R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
require(knitr)
```

```
## Loading required package: knitr
```

```
## Warning: package 'knitr' was built under R version 3.5.2
```

```
require(markdown)
```

```
## Loading required package: markdown
```

```
## Warning: package 'markdown' was built under R version 3.5.2
```

```
require(dotwhisker)
```

```
## Loading required package: dotwhisker
```

```
## Warning: package 'dotwhisker' was built under R version 3.5.2
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.5.2
```

```
set.seed(030419)
```

Question 1.

We will begin by reviewing last week's regression model.

```
## Reading in data, defining variables of interest:
Tabellini<-read.csv("Tabellini.csv", header=T)
cpi <- Tabellini$CPI
elf <- Tabellini$ELF
pop <- Tabellini$POP
luk <- Tabellini$LEGOR_UK
edu <- Tabellini$EDU
legor_fr <- Tabellini$LEGOR_FR
y <- Tabellini$Y

## H0: Population, education levels, ethnolinguistic fractionalization, GDP and French l
egal origins do not coorrelate with CPI.

reg10 <- lm(cpi ~ edu + legor_fr + y + elf + pop + (elf:pop))
summary(reg10)
```

```
##
## Call:
## lm(formula = cpi ~ edu + legor_fr + y + elf + pop + (elf:pop))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.86176 -0.60012 -0.05559  0.72937  2.28870
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.768e+00  7.955e-01  11.022 < 2e-16 ***
## edu          -2.726e-02  8.232e-03  -3.311 0.001423 **
## legor_fr     8.706e-01  2.255e-01   3.860 0.000236 ***
## y           -2.075e-04  1.676e-05 -12.379 < 2e-16 ***
## elf          4.130e-01  5.339e-01   0.774 0.441594
## pop          1.310e-02  3.433e-03   3.816 0.000274 ***
## elf:pop     -1.748e-02  5.245e-03  -3.333 0.001329 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9865 on 76 degrees of freedom
## (3 observations deleted due to missingness)
## Multiple R-squared:  0.8456, Adjusted R-squared:  0.8334
## F-statistic: 69.36 on 6 and 76 DF,  p-value: < 2.2e-16
```

```
stargazer::stargazer(reg10, type = "html",
  title = "Reg10")
```

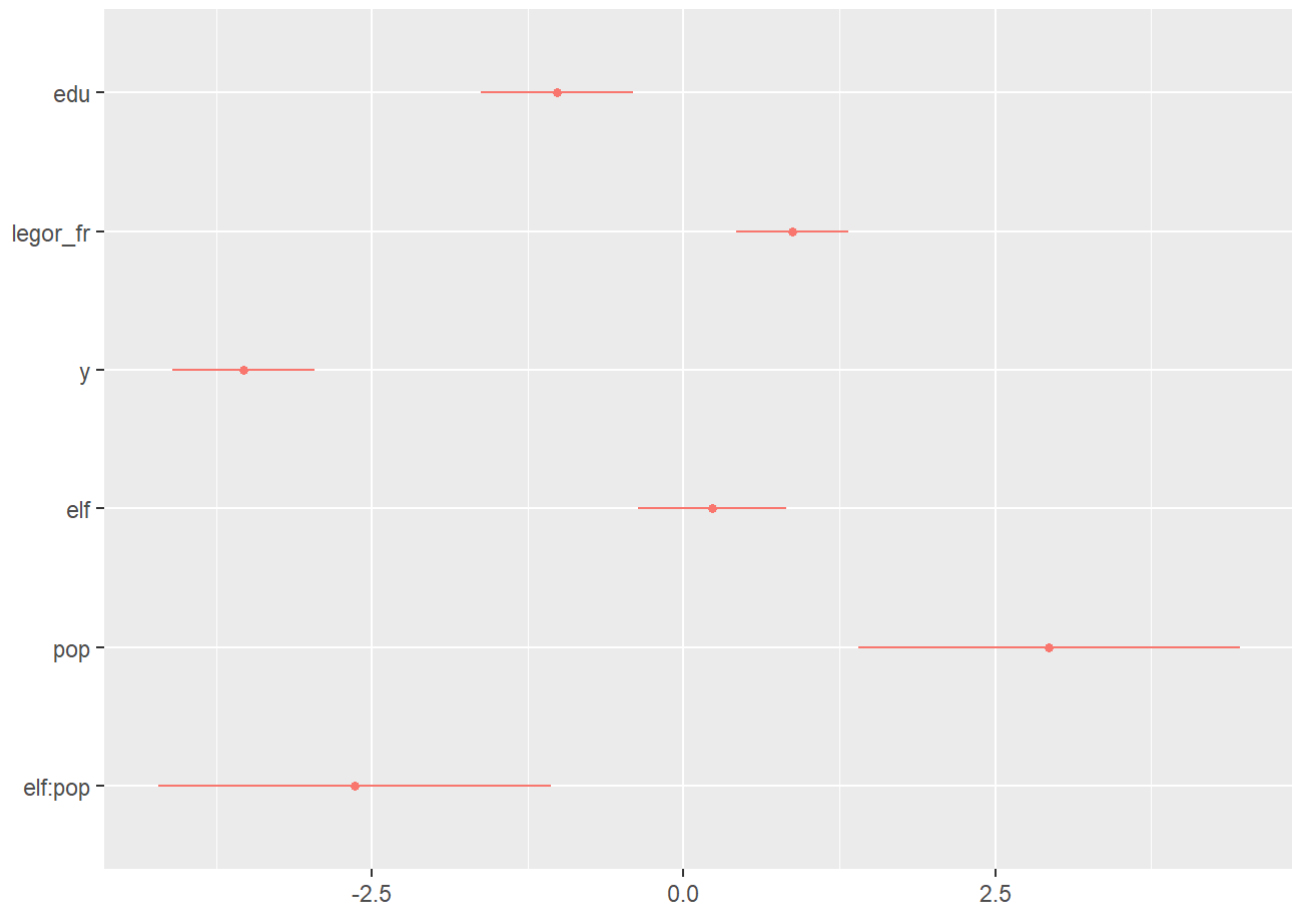
Reg10

<i>Dependent variable:</i>	
cpi	
edu	-0.027*** (0.008)
legor_fr	0.871*** (0.226)
y	-0.0002*** (0.00002)
elf	0.413 (0.534)
pop	0.013*** (0.003)
elf:pop	-0.017*** (0.005)
Constant	8.768*** (0.795)
Observations	83
R ²	0.846
Adjusted R ²	0.833
Residual Std. Error	0.987 (df = 76)
F Statistic	69.365*** (df = 6; 76)

Note: $p < 0.1$; **$p < 0.05$** ; $p < 0.01$

Let's examine the regression coefficients graphically:

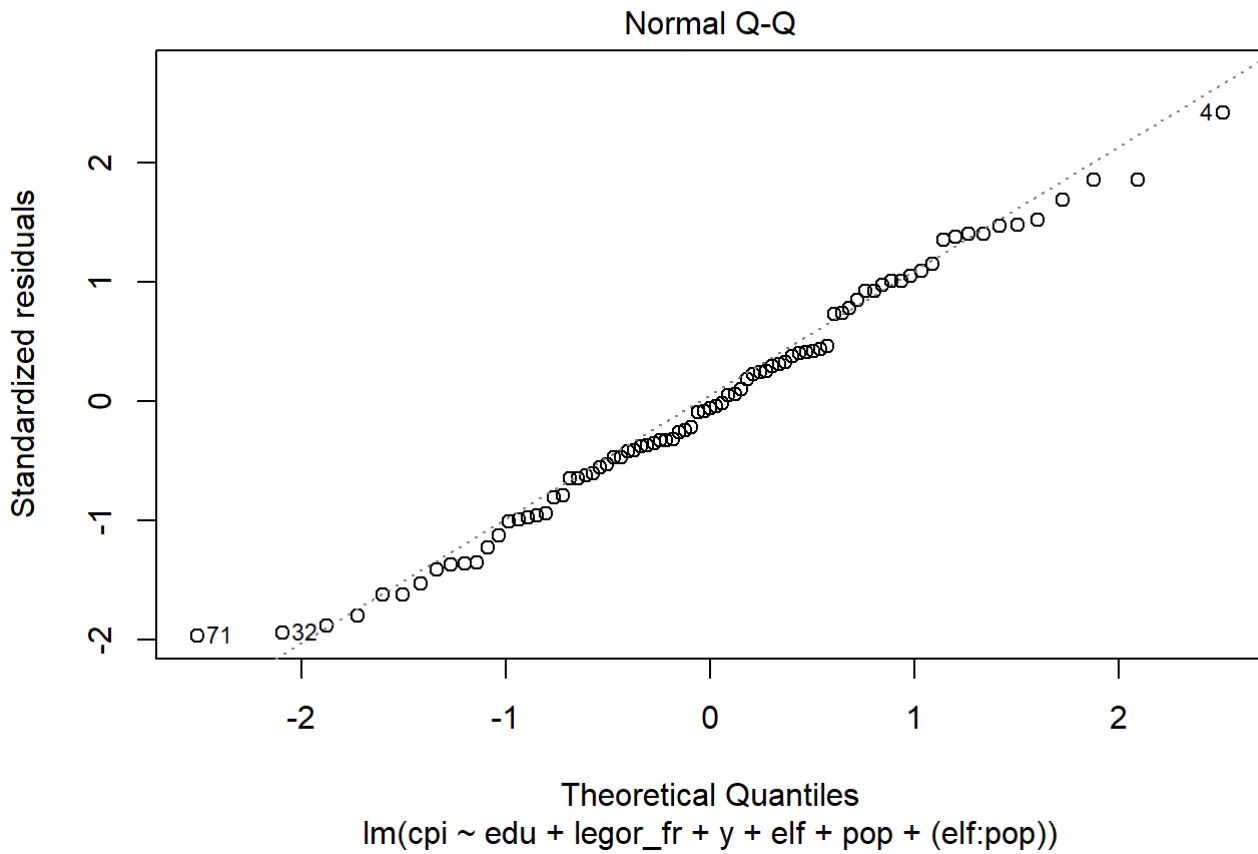
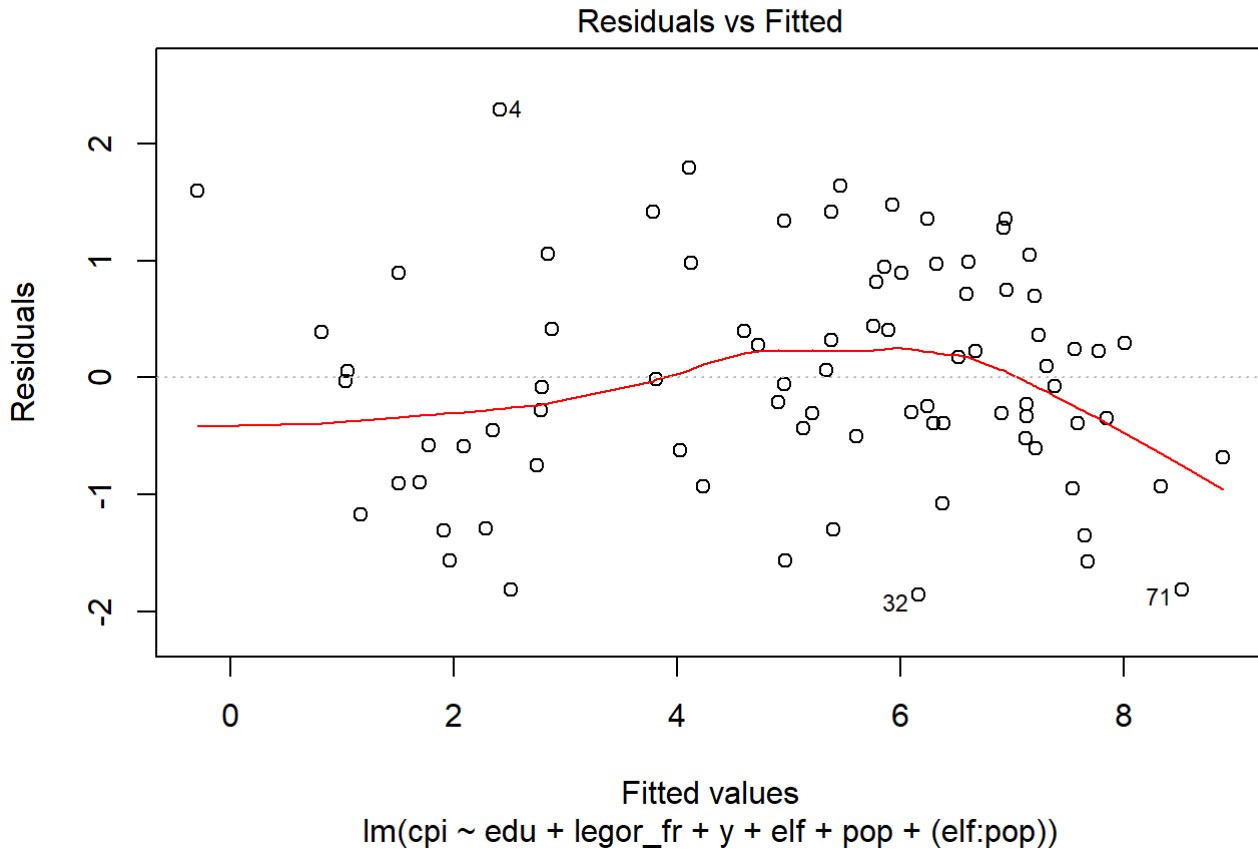
```
dwplot(reg10)
```

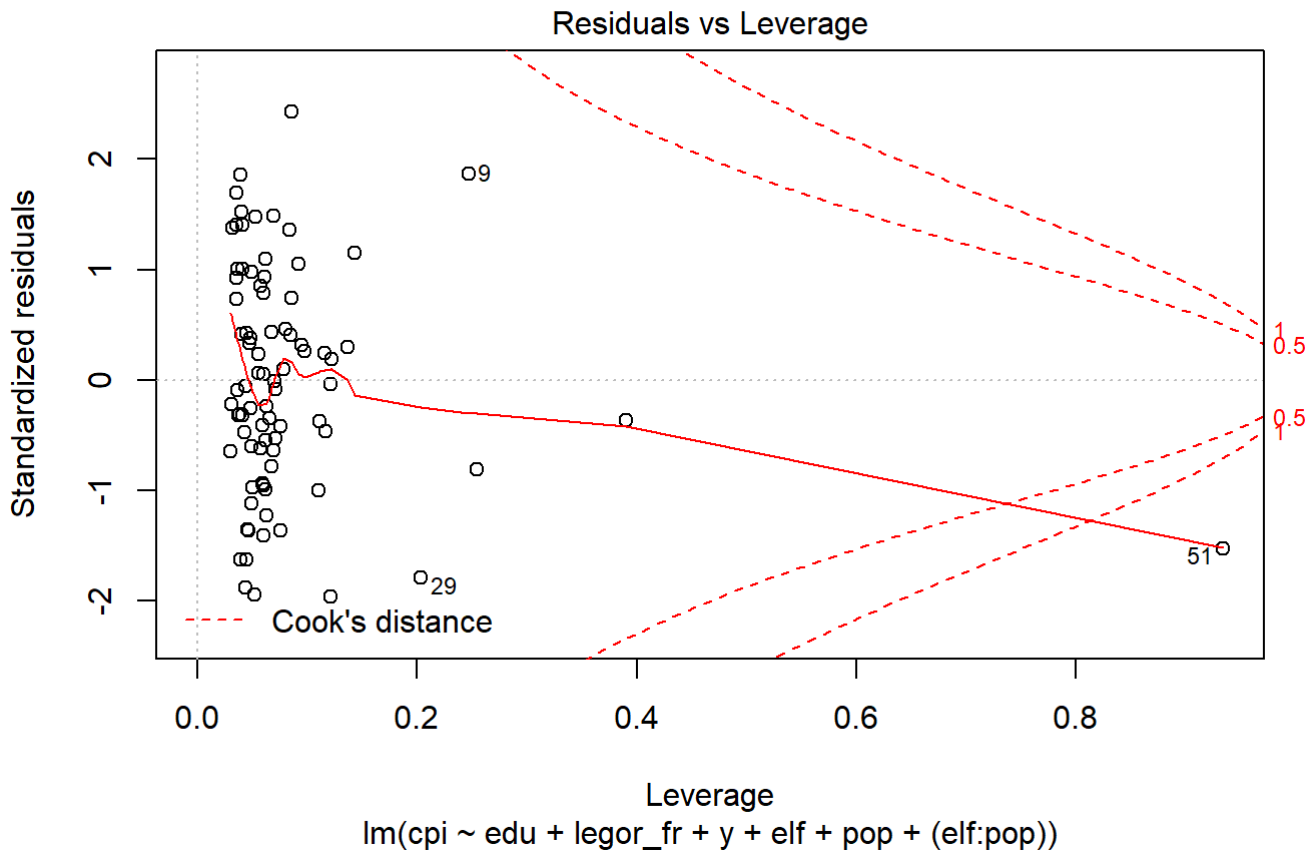
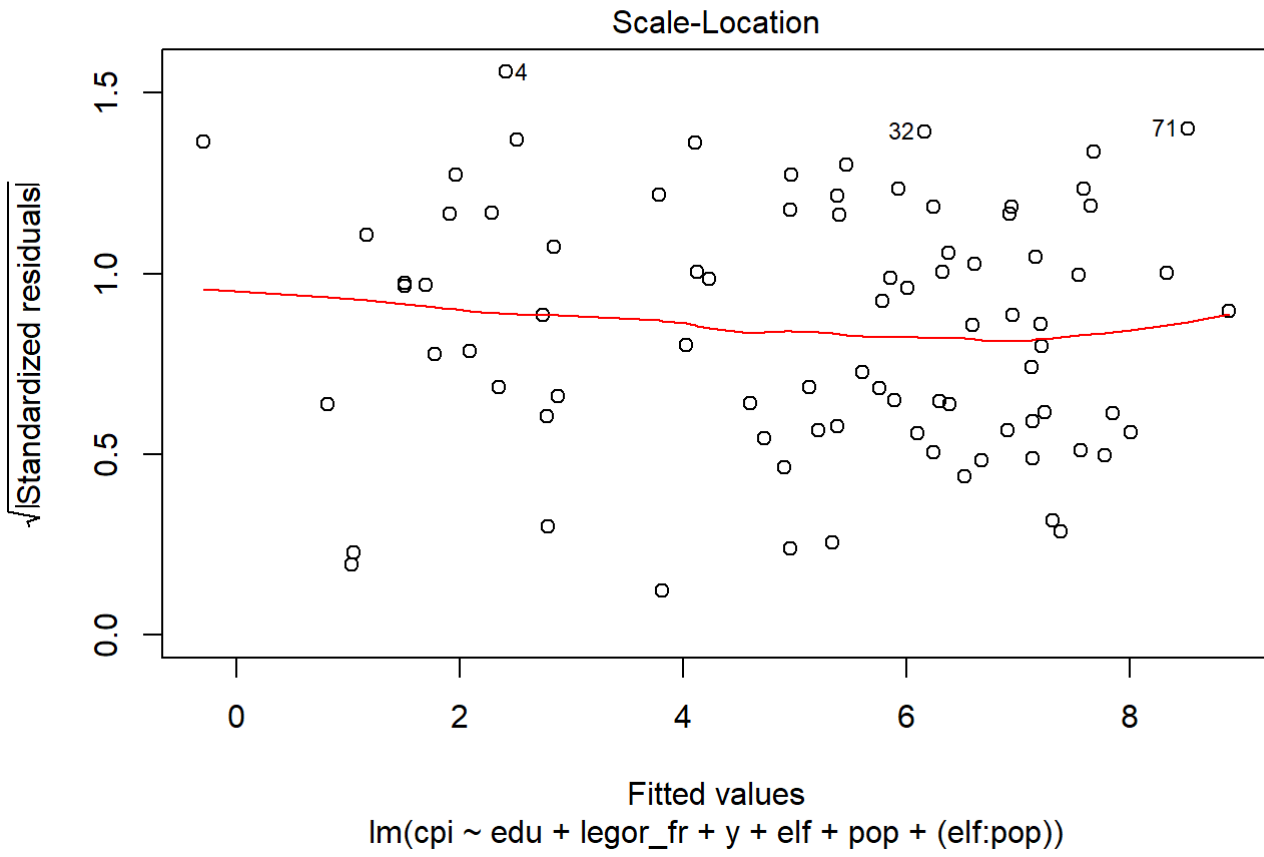


We will now apply the diagnostic tools we learned last week: residual plots, leverages, Studentized residuals, DFBETAs, Cook's distance, and multicollinearity. We want to know whether any of the core regression assumptions (linearity, normality and fixity) are violated.

First we will examine the residual plots:

```
plot(reg10)
```



In the residuals vs. fitted plot, we observe weak linearity at the left and center of the distribution, with a hint of curvature at the right side. In addition, we observe homoskedasticity, as error terms appear to be uncorrelated with x .

The q-q plot verifies the normality of our distribution, with notable outliers at 4, 71 and 32 (Belgium, Senegal and Costa Rica).

The scale-location plot shows an even spread of residuals along the range of predictors, with a nearly-horizontal fitted line, indicating homoskedasticity.

Finally, the residuals vs. leverage plot will assist us in identifying influential cases. Observations beyond the Cook's distance lines are relatively more influential. In our case, observation #51 (India) may be driving our results.

To investigate, we will drop observation #51 from the model and compare the results:

Excellent

```
Tabellini<-Tabellini[-c(51),]
reg10 <- lm(cpi ~ edu + legor_fr + y + elf + pop + (elf:pop))
summary(reg10)
```

```
##
## Call:
## lm(formula = cpi ~ edu + legor_fr + y + elf + pop + (elf:pop))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.86176 -0.60012 -0.05559  0.72937  2.28870
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.768e+00  7.955e-01  11.022 < 2e-16 ***
## edu          -2.726e-02  8.232e-03  -3.311 0.001423 **
## legor_fr     8.706e-01  2.255e-01   3.860 0.000236 ***
## y           -2.075e-04  1.676e-05 -12.379 < 2e-16 ***
## elf          4.130e-01  5.339e-01   0.774 0.441594
## pop          1.310e-02  3.433e-03   3.816 0.000274 ***
## elf:pop     -1.748e-02  5.245e-03  -3.333 0.001329 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9865 on 76 degrees of freedom
## (3 observations deleted due to missingness)
## Multiple R-squared:  0.8456, Adjusted R-squared:  0.8334
## F-statistic: 69.36 on 6 and 76 DF,  p-value: < 2.2e-16
```

The results are substantially the same without observation #51.

Good

Now we will examine residual plots for each variable:

```
require(car)
```

```
## Loading required package: car
```


:DUQLQJ SDFNDJH FDU ZDV EXLOW XQGHU 5 YHUVLRQ

/RDGLQJ UHTXLUHG SDFNDJH FDU'DWD

:DUQLQJ SDFNDJH FDU'DWD ZDV EXLOW XQGHU 5 YHUVLRQ

UHVLGXDO3ORWV UHJ

Good

7HVW VWDW 3U !_7HVW VWDW_

HGX
OHJRUBIU
\
HOI
SRS
7XNH\ WHVW

6LJQLI FRGHV

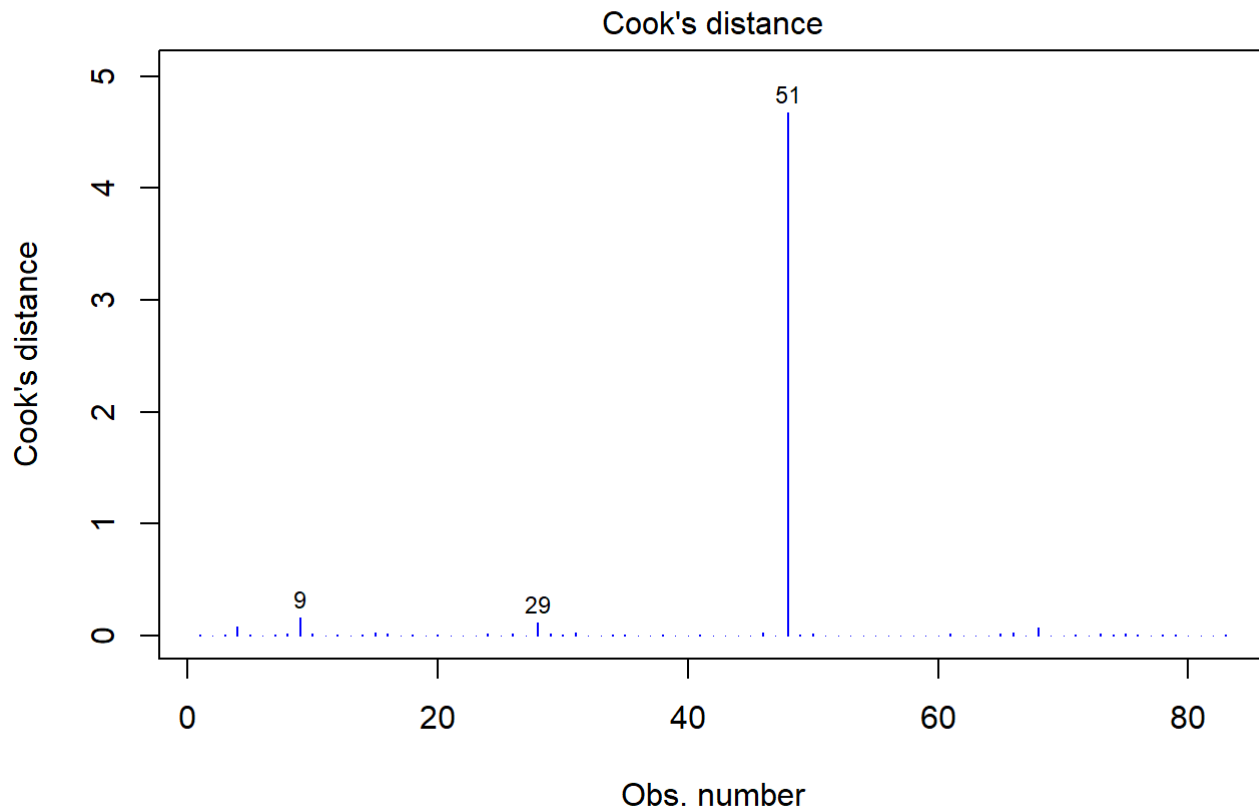
ONUMRNK\TT` HUKJH[LMNPHS] HPH \$Z ^L UL] LY[CLS Z Z IL J\ Y] BULHYLS[RUZ QZ RJM Y] HPH \$Z ; CZ
JHZ ZKV I [VU[CL \$LHP` HZ ZT VPRU

* VU[RA RN\ YR] LZ [PHRU ^L ^ \$SSVRH[\$] LYNL I` NYMOPN[CL CH] \$LZ QRO \$] LYNL WRJZ V [\$YZ RJ
_!

SORW KDWYDOXHV UHJK ~~FR~~EOXH\ODE+DW 9DOX~~HP~~DLQ+DW YDOXHV
S Q
DEOLQH K S Q OW\ 7KUHVKROG IRU VXVSHFWV

5L_[^L^~~SL~~_HT RL * WRZKZ HWL!

SORW UHJ ZKL~~FR~~EOXH



Here,

Obs. number
lm(cpi ~ edu + legor_fr + y + elf + pop + (elf:pop))

observation #48 (Bangladesh) is the only significant outlier.

Finally, we will examine the DFBetas:

```
par(mfrow=c(2,3))
for (j in 1:7){
  plot(abs((dfbetas(reg10))[,j]), col=4, type='h', ylab='DFBETAS')
  abline(h = 2/sqrt(n), lty=2) # Threshold for suspects
}
```

> OFF

RKPK HSI Z D HPVZ JYZ ZCL JVWLYU [OLZ OS VUHS] HPI SZ YS [PLS M^ VM] CLT KV HUK RHHKPPU
ÀOH 8VHUV FVHOOND*RLD H32/JLYH% +RPHZRUNV +: 6XEPLVVLQRV 'UDSHU +: KWPO

different observations have high DFBetas for different variables (rather than the same ones).

Question 2a.

The question asks us what happens to the estimate of b_1 when x_2 is missing, in two cases: x_1 and x_2 have positive covariance, and x_1 and x_2 have negative covariance. We will use the omitted variable simulation from lecture:

```
omitsim<-function(n=100,numsims=1000,rho1=2){
  par(mfrow=c(2,2))
  B1<-runif(n)*4
  B2<-rnorm(n,mean=100,sd=10)+(B1*rho1) ## The variable rho1 is our measure of correlation.
  print(cor(B1,B2))
  res<-array(NA,dim=c(numsims,2))
  for(i in 1:numsims){
    Y<-10+2*B2 + 2*B1+rnorm(n)
    res[i,1]<-lm(Y~B1+B2)$coef[2]
    res[i,2]<-lm(Y~B1)$coef[2]}
  print(lm(Y~B1+B2))
  print(lm(Y~B1))
  hist(res[,1],main="Correct Model",xlab="Est of B(B1)")
  hist(res[,2],main="Model W/O B2",xlab="Est of B(B1)")
  plot(res[,1],res[,2],xlim=c(min(res),max(res)),ylim=c(min(res),max(res)))
  abline(0,1)
}
```

We will now run the omitsim function. B_1 and B_2 are positively correlated because of the relationship in line 19. Our correlation measure is a positive number, so the simulation will produce a positive covariance, with residuals clustering in the top left corner. When x_2 is missing, our overestimate of B_1 will increase. As positive covariance increases, the amount by which we overestimate B_1 increases when we drop x_2 .

```
omitsim(n=100,numsims=1000,rho1=2)
```

```
## [1] 0.1003581
##
## Call:
## lm(formula = Y ~ B1 + B2)
##
## Coefficients:
## (Intercept)          B1          B2
##      9.968      1.987      2.000
##
##
## Call:
## lm(formula = Y ~ B1)
##
## Coefficients:
## (Intercept)          B1
##      214.552      3.716
```

OM L

RUYLHZ L[QL T HNU\ KL VMOV [QL NHI L[^ LLU[QL T VKLS ^ S RUYLHZ IL

RPLWVLP QQXPVLPV UKR

```

> @

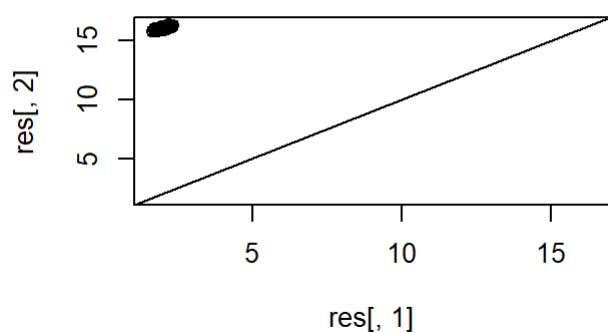
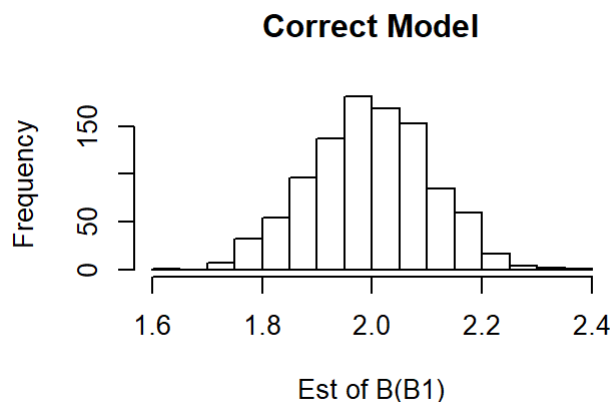
&DOO
OP IRUPXOD < a % %

&RHIILFLHQWV
,QWHUFHSW % %

&DOO
OP IRUPXOD < a %

&RHIILFLHQWV
,QWHUFHSW %

```

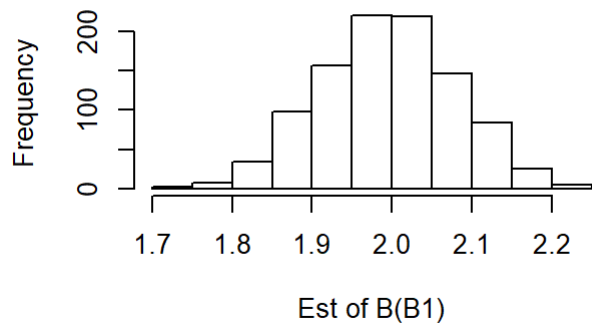
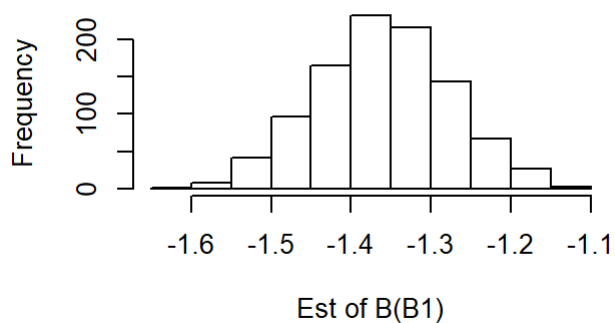


When

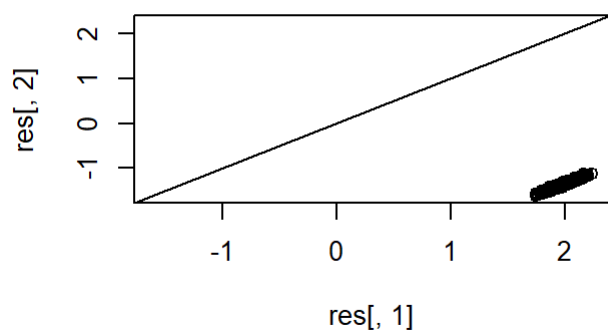
we change the rho1 value to a negative number, this creates a negative covariance between B1 and B2. Residuals will cluster in the bottom right corner. When x2 is missing, our overestimate of B1 will decrease. As negative covariance increases, the amount by which we overestimate B1 increases when we drop x2.

```
omitsim(n=100,numsims=1000,rho1=-2)
```

```
## [1] -0.1890616
##
## Call:
## lm(formula = Y ~ B1 + B2)
##
## Coefficients:
## (Intercept)      B1      B2
##      8.301      2.025      2.016
##
## Call:
## lm(formula = Y ~ B1)
##
## Coefficients:
## (Intercept)      B1
##      204.646      -1.357
```

Correct Model**Model W/O B2**

If we



increase the magnitude of rho1, the gap between the models will increase:

```
omitsim(n=100,numsims=1000,rho1=-7)
```

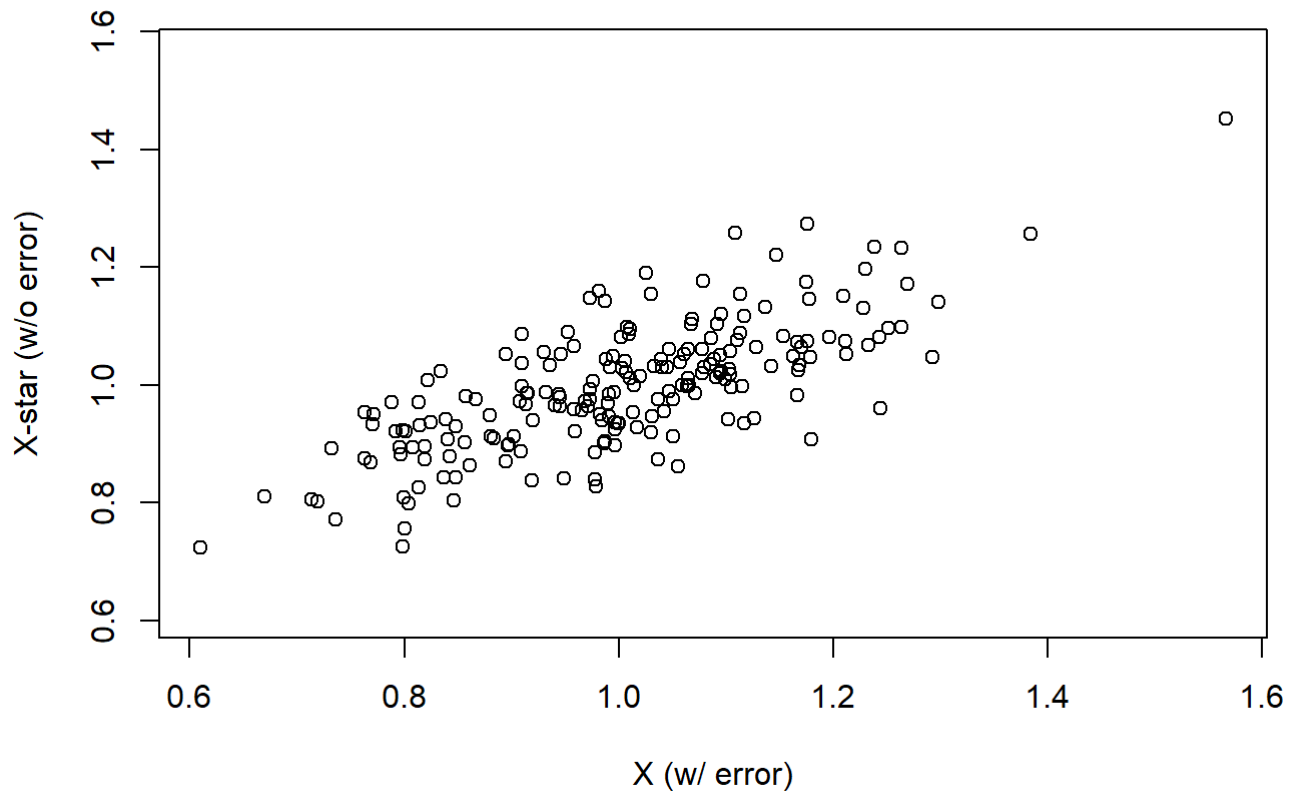
```
## [1] -0.6169707
##
## Call:
## lm(formula = Y ~ B1 + B2)
##
## Coefficients:
## (Intercept)          B1          B2
##      11.430       1.905       1.986
##
##
## Call:
## lm(formula = Y ~ B1)
##
## Coefficients:
## (Intercept)          B1
##      212.89       -12.16
```


8\ LZ PU I

UNLULYS^ CLU_ FT LHZ \LK^ POYUKVT LYWY[CL LYWZ^ S[LUK[V AH][LU[CL Z SWL VMCL YNLZ ZP SL
; OZ EI LJH Z L[CL LYWY^ S[LUK[V KEZ WZ L[CL KH H YZ \SPN RUIH I RNLY SZ ZWLJEZ LJS K (Z YUKVT LYWY
RUYLHZ IZ ^ L [CL YMYL L_VLJ[[VZ IL I L[HCH HWWWUOaLY

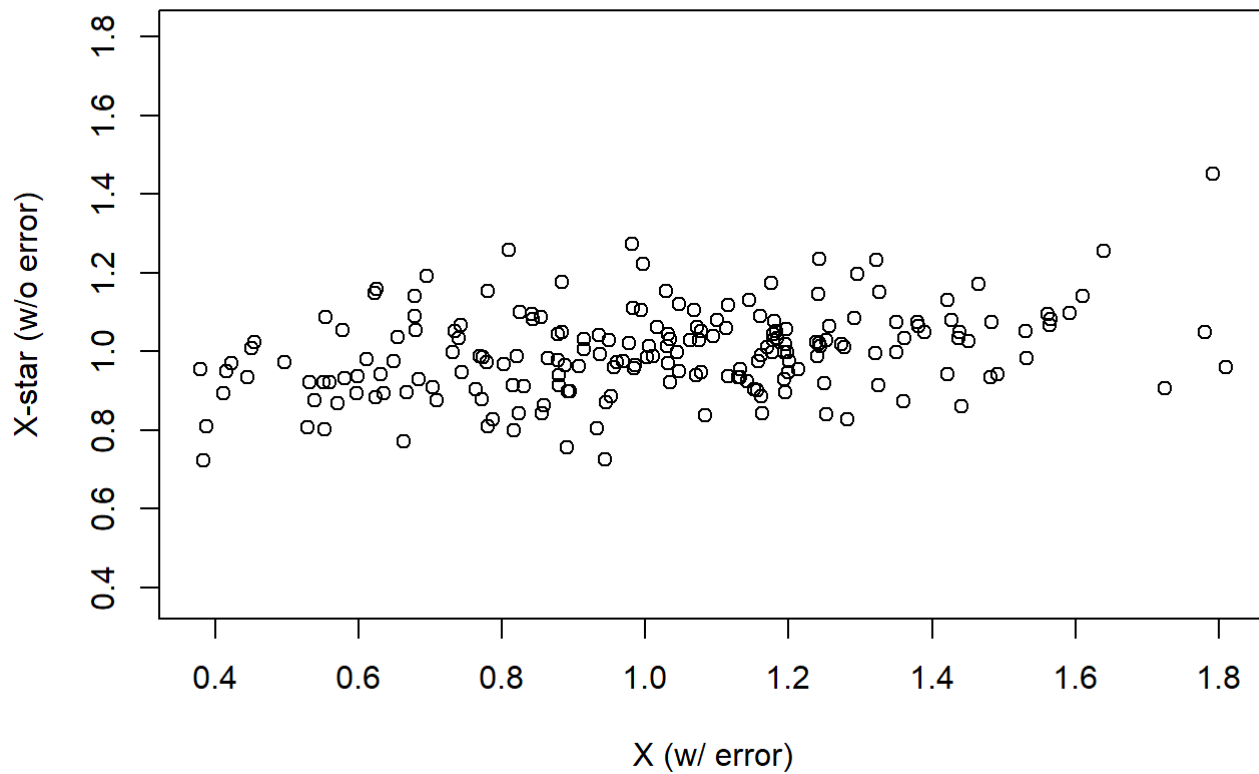
- PZ [^ L L HT RL [CL KH H^ POHUYT HS KEZ TP \ [LK LYWY[LYT ^ POHZ FUKHK KL] P[PUVM > L ^ S[NHVO_
^ POLYWY VU[CL _ H_Z HNRZ [_ Z FY [X L] HSL VU[CL ` H_Z > L L_VLJ[[VZ IL KEZ WZ VUHSUN[CL _ H_Z

```
VHW VHHG
UHJVLPPIXQFWLQQ QXPVLPV DOSKDEHWDVLJPDJDPVHUHJ^
; UQRUP Q PHDQ
UHV DUU1D\GLP F QXPVLPV
IRUL LQ QXPVLPV ^
<VWDU DOSKD EHWD ; UQRUP Q VG VHUHJ
< <VWDU UQRUP Q VG
UHV>L@ OP <VWDUa; @RHI>
UHV>L@ OP <a; FR@I>
,
SORW UH@>UHV@ [ODE Z HUURU\ODE; VWDU Z R HUURU\@LP F PLQ UHV
UHV \OLP F PLQ UHV PD[ UHV
,
UHJVLP
```



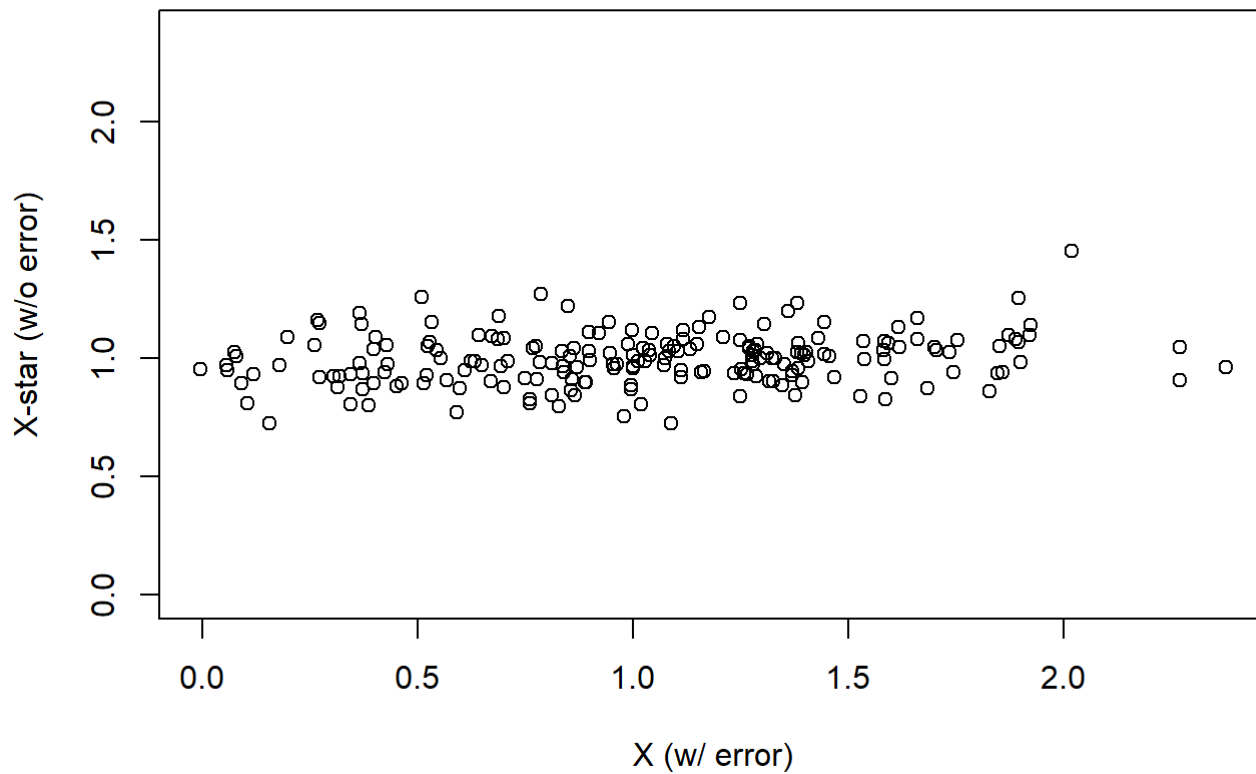
As we increase standard deviation, the distribution flattens:

```
set.seed(030419)
regsim0<-function(n=100,numsims=200,alpha=1,beta=1,sigmagamma=1,sereg=1){
  X<-rnorm(n,mean=1)
  res<-array(NA,dim=c(numsims,2))
  for(i in 1:numsims){
    Ystar<-alpha+beta*X+rnorm(n,sd=sereg)
    Y<-Ystar+(3*rnorm(n,sd=1))
    res[i,1]<-lm(Ystar~X)$coef[2]
    res[i,2]<-lm(Y~X)$coef[2]
  }
  plot(res[,2],res[,1],xlab="X (w/ error)",ylab="X-star (w/o error)",xlim=c(min(res),max
(res)),ylim=c(min(res),max(res)))
}
regsim0()
```



Eventually, the error term dominates the result:

```
set.seed(030419)
regsim0<-function(n=100,numsims=200,alpha=1,beta=1,sigmagamma=1,sereg=1){
  X<-rnorm(n,mean=1)
  res<-array(NA,dim=c(numsims,2))
  for(i in 1:numsims){
    Ystar<-alpha+beta*X+rnorm(n,sd=sereg)
    Y<-Ystar+(5*rnorm(n,sd=1))
    res[i,1]<-lm(Ystar~X)$coef[2]
    res[i,2]<-lm(Y~X)$coef[2]
  }
  plot(res[,2],res[,1],xlab="X (w/ error)",ylab="X-star (w/o error)",xlim=c(min(res),max
(res)),ylim=c(min(res),max(res)))
}
regsim0()
```



```

set.seed(030419)
regsim0<-function(n=100,numsims=200,alpha=1,beta=1,sigmagamma=1,sereg=1){
  X<-rnorm(n,mean=1)
  res<-array(NA,dim=c(numsims,2))
  for(i in 1:numsims){
    Ystar<-alpha+beta*X+rnorm(n,sd=sereg)
    Y<-Ystar+(7*rnorm(n,sd=1))
    res[i,1]<-lm(Ystar~X)$coef[2]
    res[i,2]<-lm(Y~X)$coef[2]
  }
  plot(res[,2],res[,1],xlab="X (w/ error)",ylab="X-star (w/o error)",xlim=c(min(res),max
(res)),ylim=c(min(res),max(res)))
}
regsim0()

```

> L^SUV^ JYLHLH]HPI \$ A JVYLS[LK^PO[CL [XL_ !

Z[ZILK YNZP # MW[RUUS U TZPZ\$ HVOH\$ I[LH\$ ZINTHNHTTH\$ ZLN\$ b?# YUVT UTLHU\$ A# ? YUVT UTLHU\$? CH# S Ae? WVKR[RUZ# WVKR[? CH UL^KH\$? CH ZLÄ\$;9<, R[LYH\$,JVUÄKLWL' \$]L\$ YLZ# HYH 5(KR\$J U T Z F Z MYPRU !U T Z F Z b @Z]H# HVOH I L[H? CH @ @Z]H YUVT UZ K\$ YLZ B # S @Z]H? CJVLNB DYLB # S @? CJVLNB Dd WS[YLZ BDYLB BD_S\$ \$,? ^ LYWY' ` S\$ \$,? Z]HY^ VLYWY' _S\$ \$J T R U Y L Z T H_ Y L Z ` S\$ \$J T R U Y L Z T H_ Y L Z d YNZP ,'

FF ([[LTW !

YNZP # MW[RUUS U TZPZ\$ HVOH\$ I[LH\$ ZINTHNHTTH\$ ZLN\$ b?# YUVT UTLHU\$ A# ? YUVT UTLHU\$ YLZ# HYH 5(KR\$J U T Z F Z MYPRU !U T Z F Z b @ HVOH I L[H? YUVT UZ K\$ ZLN ?Z]H# ? YUVT UZ K\$ ZINTHNHTTH YLZ B # S A? CJVLNB? CH# YLZ BDYLB # S @? CH CJVLNB DYLB # S @?Z]HY JVLNB Dd WS[YLZ BDYLB BD_S\$ \$,? ^ LYWY' ` S\$ \$,? Z]HY^ V LYWY' _S\$ \$J T R U Y L Z T H_ Y L Z ` S\$ \$J T R U Y L Z T H_ Y L Z H \$L \$LZJ I L[H I L[H J T R U Y L Z T H_ Y L Z \$LZJ T R U Y L Z T H_ Y L Z J I L[H I L[H d YNZP

([[LTW ! OT Z WJSZL [CL [XAI \$ E [CH OJHU] ÄN YL V [OV^ [VWWW \$L ? CH ^ POU[CL S W W

YNZP # MW[RUUS U TZPZ\$ HVOH\$ I[LH\$ ZINTHNHTTH\$ ZLN\$ b?# YUVT UTLHU\$ A# ? YUVT UTLHU\$ YLZ# HYH 5(KR\$J U T Z F Z MYPRU !U T Z F Z b @ HVOH I L[H? YUVT UZ K\$ ZLN ?Z]H# ? YUVT UZ K\$ ZINTHNHTTH YLZ B # S A? CJVLNB? CH# YLZ BDYLB # S @? CH CJVLNB DYLB # S @?Z]HY JVLNB Dd WS[YLZ BDYLB BD_S\$ \$,? ^ LYWY' ` S\$ \$,? Z]HY^ V LYWY' _S\$ \$J T R U Y L Z T H_ Y L Z ` S\$ \$J T R U Y L Z T H_ Y L Z H \$L \$LZJ I L[H I L[H J T R U Y L Z T H_ Y L Z \$LZJ T R U Y L Z T H_ Y L Z J I L[H I L[H d YNZP

I'm not getting the results I want from these loops, but intuitively we would expect the regression on Z to track X1 to the extent of their correlation, but we will see divergence in the error terms.

Question 3.

I decided to reexamine Tabellini's 2001 corruption dataset from a different angle. I want to know whether the corruption perceptions index (CPI) score affects GDP. The proposed mechanism is that a perceptible increase in corruption-related rents would cause a corresponding deadweight loss to GDP.

H0: An increase (decrease) in CPI has no effect on GDP.

HA: An increase (decrease) in CPI causes an increase (decrease) in GDP.

```
Tabellini<-read.csv("Tabellini.csv", header=T)
y <- Tabellini$Y
cpi<-Tabellini$CPI
reg20<-lm(y ~ cpi)
```

```
stargazer::stargazer(reg20, type = "html",
  title = "lm(y ~ cpi)")
```

lm(y ~ cpi)

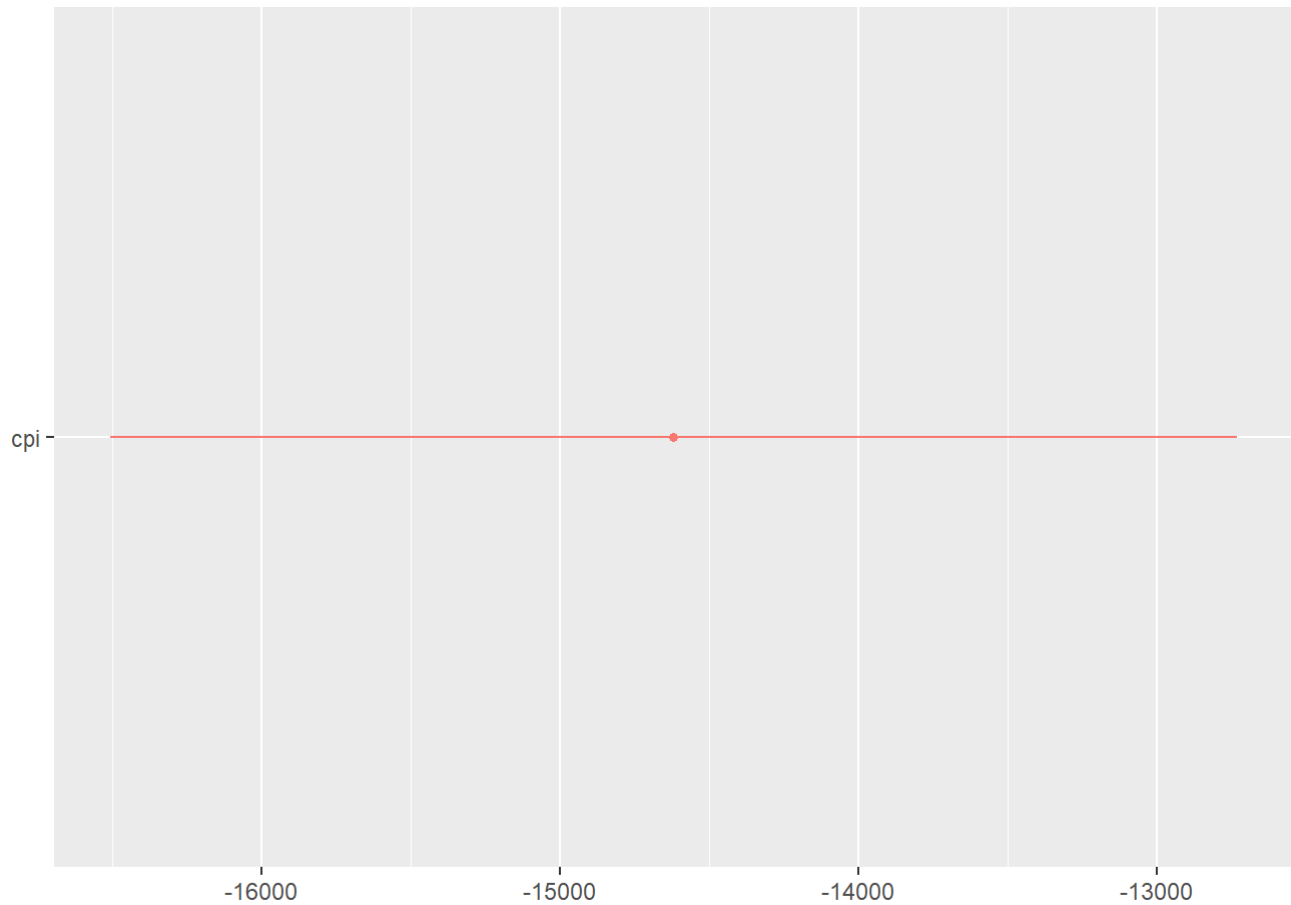
<i>Dependent variable:</i>	
	y
cpi	-3,042.323 ^{***} (197.513)
Constant	25,205.680 ^{***} (1,122.533)
Observations	84
R ²	0.743
Adjusted R ²	0.740
Residual Std. Error	4,323.446 (df = 82)
F Statistic	237.256 ^{***} (df = 1; 82)
Note:	$p < 0.1$; $p < 0.05$; $p < 0.01$

$P < .05$, so we reject the null hypothesis. Corruption is strongly associated with lower GDP.

We might object that there's a lag built in to this model, in that present corruption might affect future gdp, but not present gdp. However, corruption perceptions are themselves composites of corruption perceptions over lifetimes, with perhaps a strong recency bias but at least the potential to encode older information.

Let's examine the regression coefficients graphically:

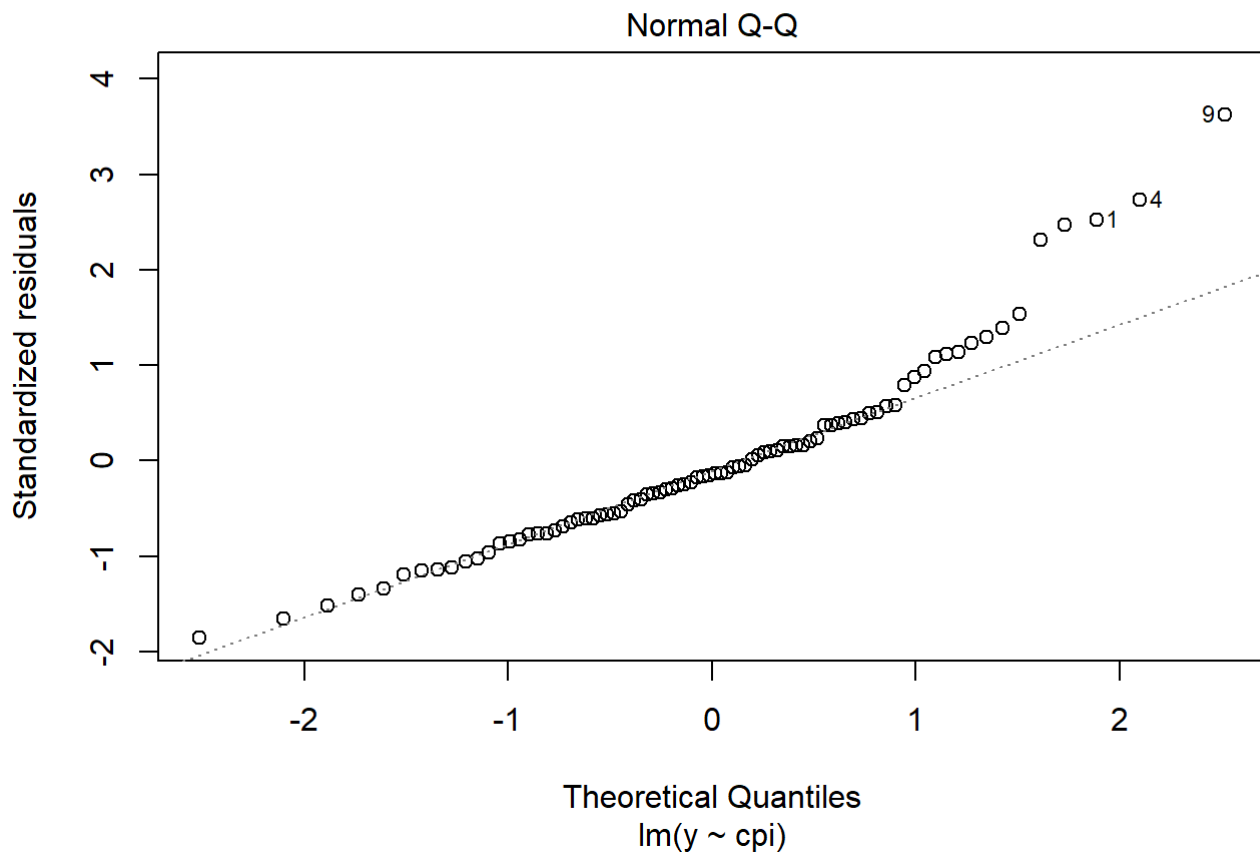
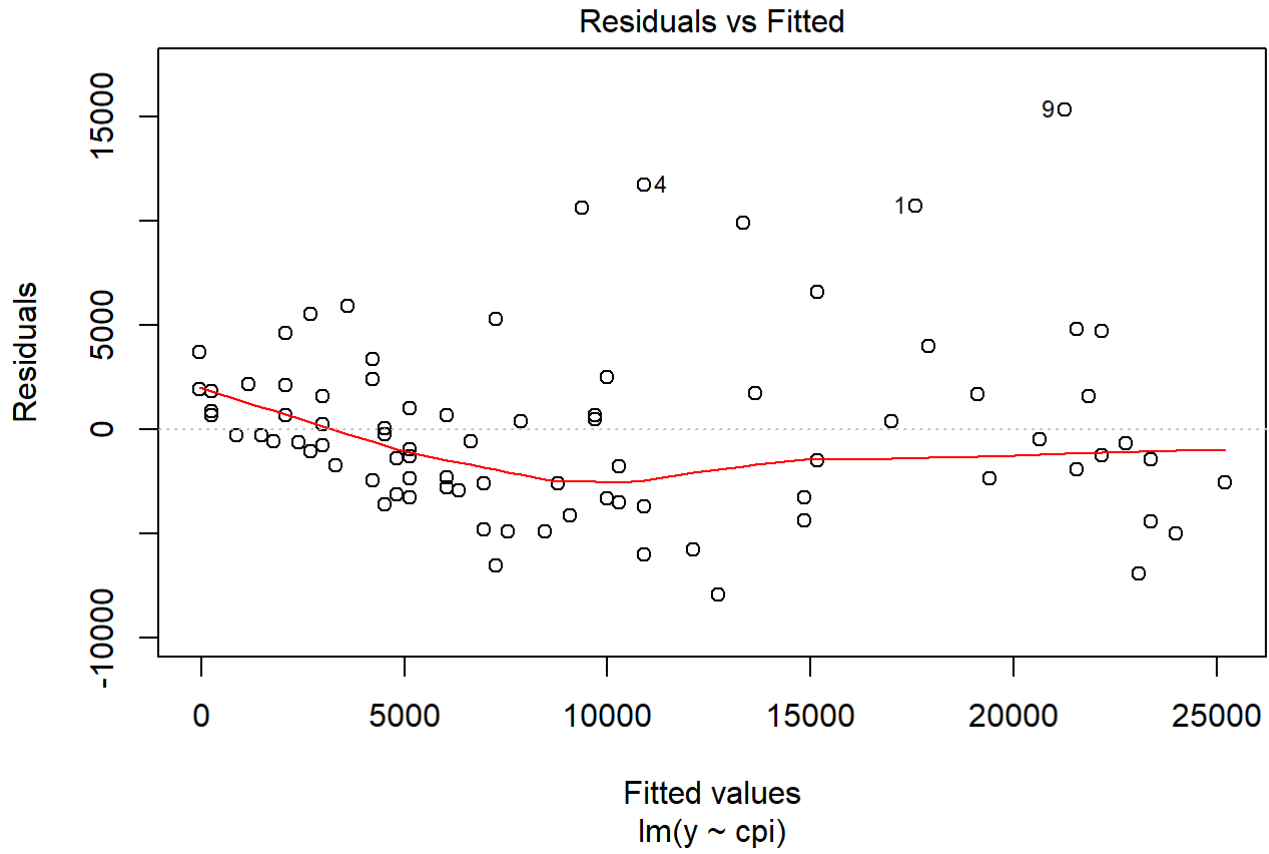
```
dwplot(reg20)
```

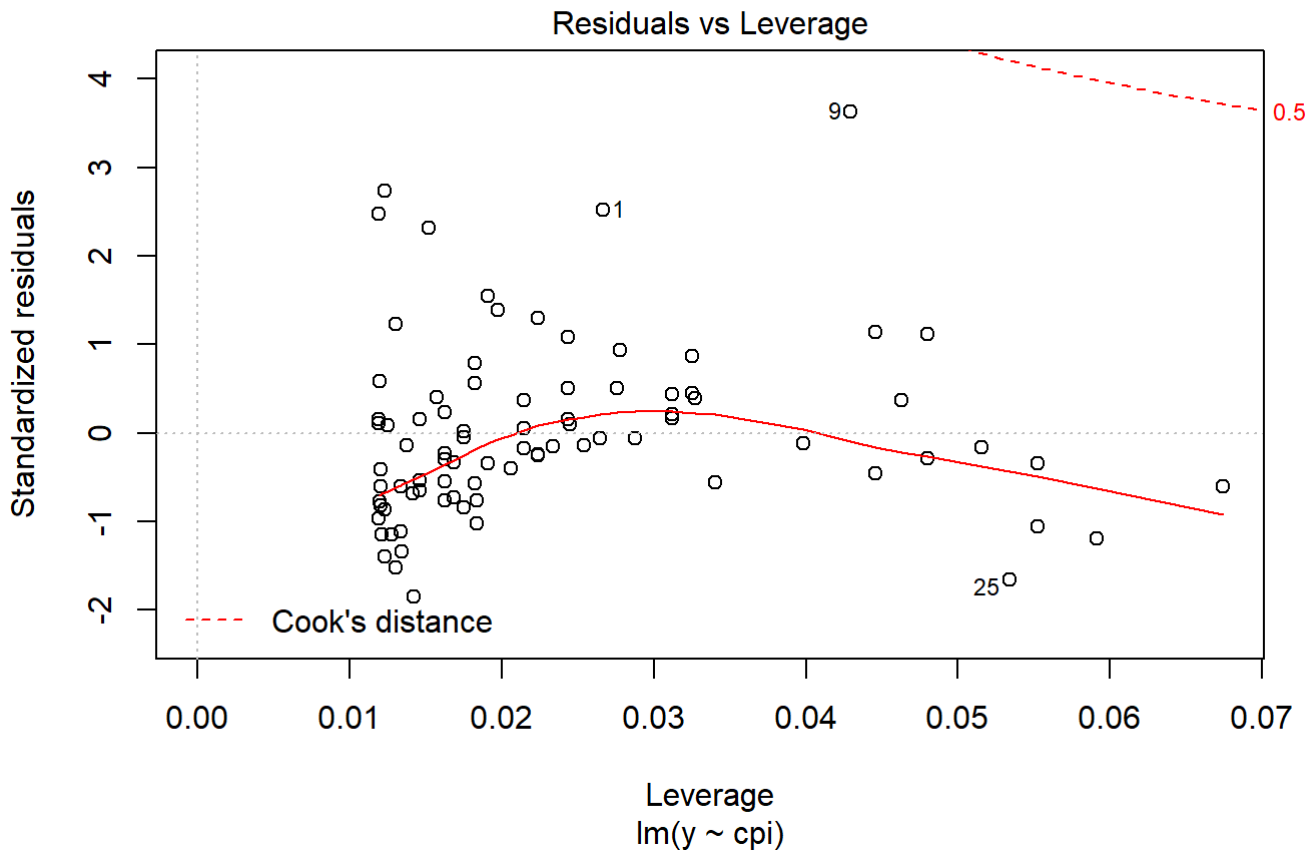
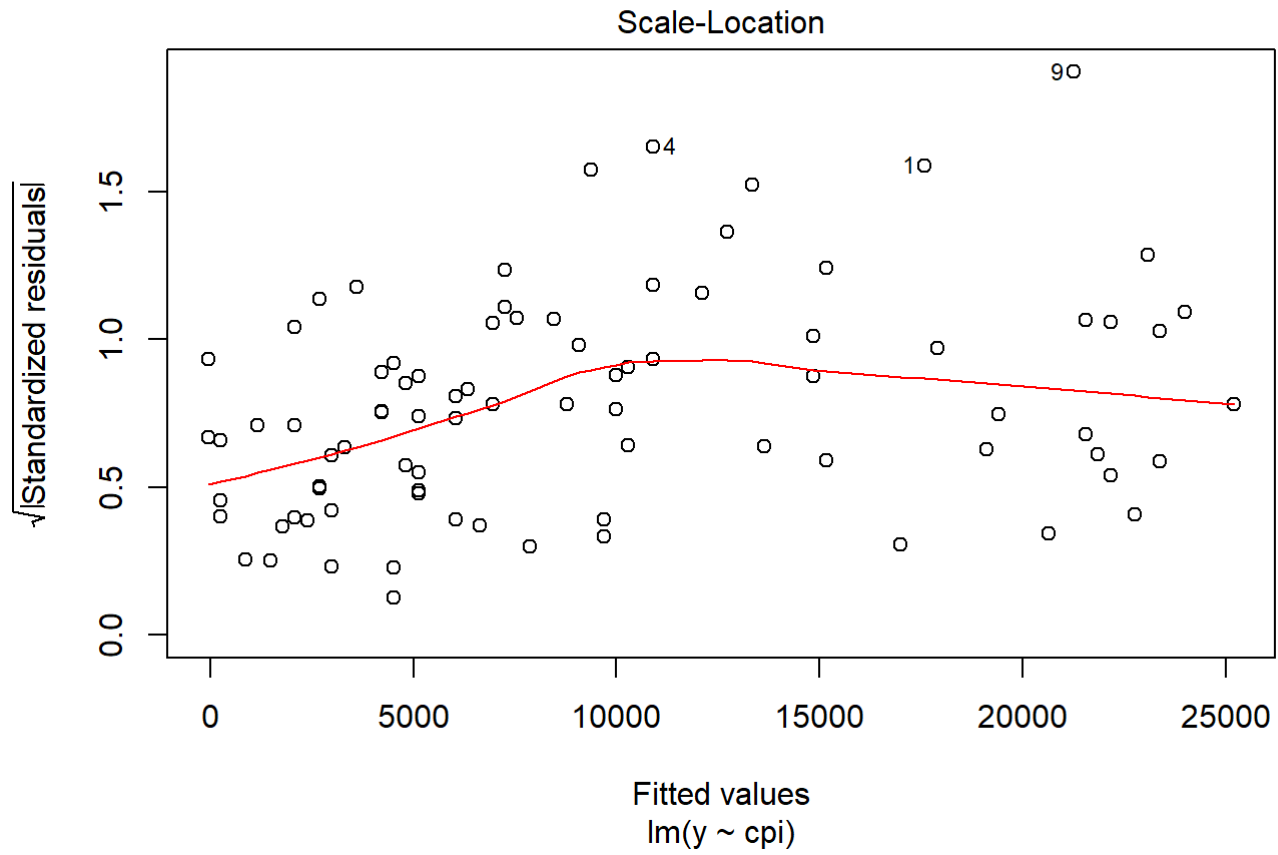


We will now apply the diagnostic tools we learned last week: residual plots, leverages, Studentized residuals, DFBETAs, Cook's distance, and multicollinearity. We want to know whether any of the core regression assumptions (linearity, normality and fixity) are violated.

First we will examine the residual plots:

```
plot(reg20)
```



In the residuals vs. fitted plot, we observe moderate linearity spread horizontally and fairly evenly, with outliers at 1, 4 and 9 (USA, Belgium, Luxembourg). In addition, we observe homoskedasticity, as error terms appear to be uncorrelated with x .

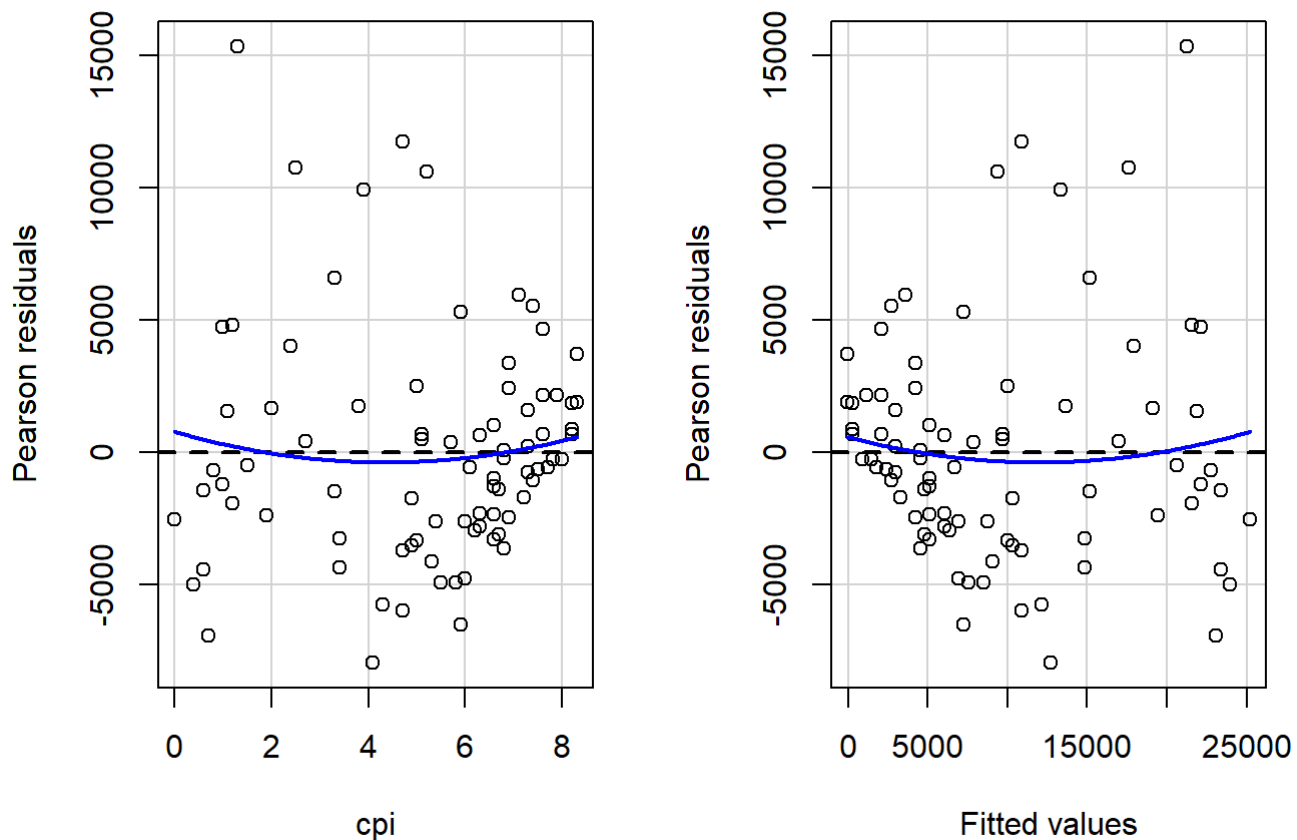
The q-q plot verifies that residuals are normally distributed, with the same notable outliers at 1, 4 and 9 (USA, Belgium, Luxembourg).

The scale-location plot shows an even spread of residuals along the range of predictors, with a nearly-horizontal fitted line, indicating homoskedasticity.

Finally, the residuals vs. leverage plot will assist us in identifying influential cases with high leverage. Observations beyond the Cook's distance lines are relatively more influential. In our case, observations 1, 9 and 25 (USA, Luxembourg and New Zealand) may be driving our results.

Now we will examine residual plots for each variable:

```
residualPlots(reg20)
```

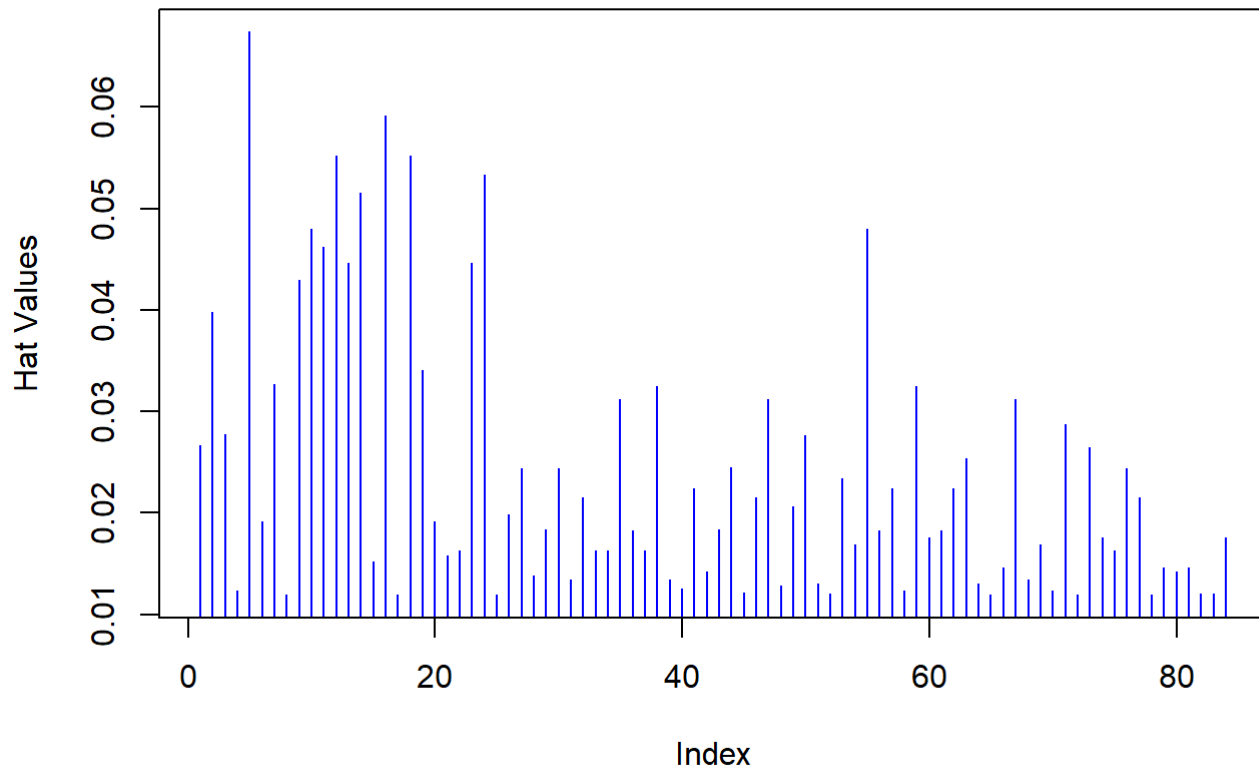


##	Test stat	Pr(> Test stat)
## cpi	0.648	0.5188
## Tukey test	0.648	0.5170

Now we will look at leverage by graphing the hat values (high-leverage points, outliers in x):

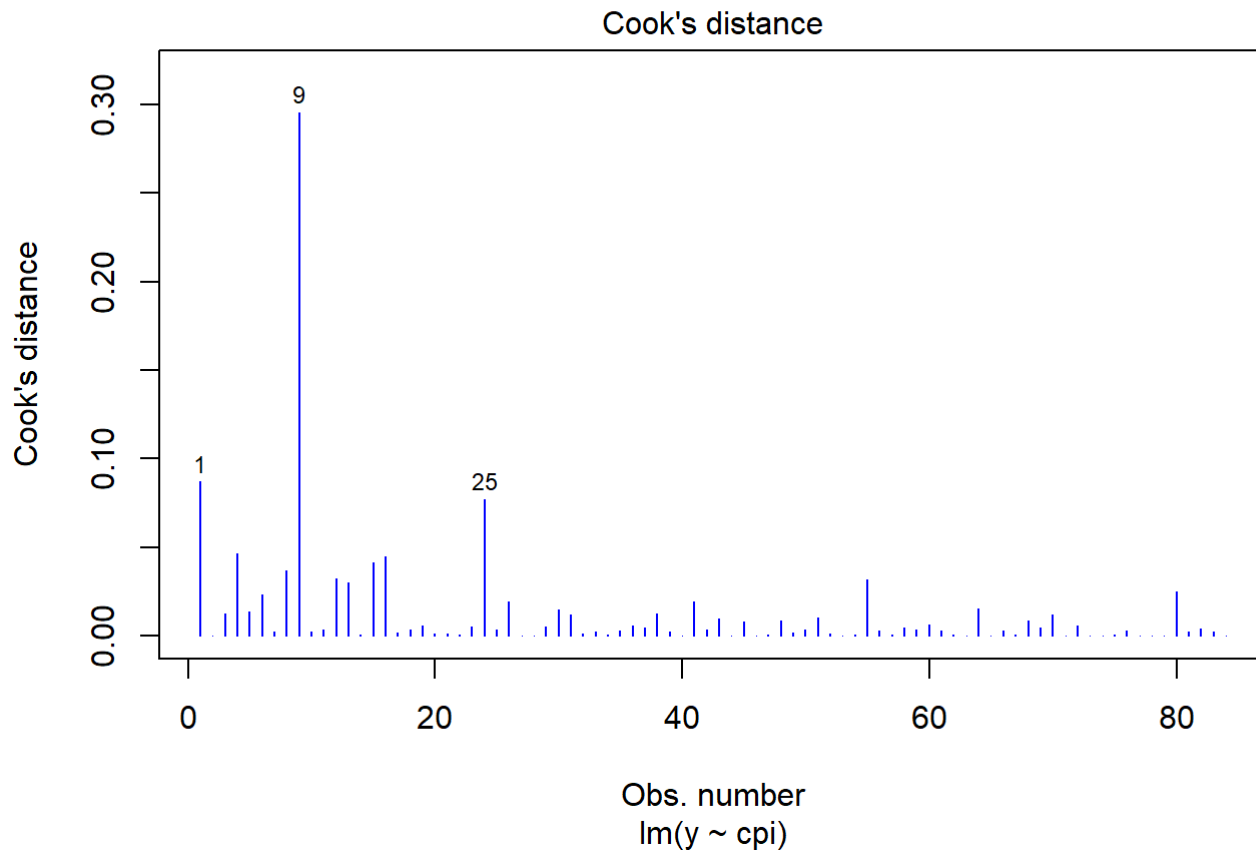
```
plot(hatvalues(reg20), type='h', col="blue", ylab="Hat Values", main="Hat values")  
p = 6; n = 86  
abline(h = 2*p/n, lty=2) # Threshold for suspects
```

Hat values



Next we will examine Cook's distance:

```
plot(reg20, which=4, col="blue")
```



Here, observations 1, 9 and 25 (USA, Luxembourg and New Zealand) are significant outliers.

Finally, we will examine the DFBetas:

```
par(mfrow=c(1,2))
for (j in 1:2){
  plot(abs((dfbetas(reg20))[,j]), col=4, type='h', ylab='DFBETAS')
  abline(h = 2/sqrt(n), lty=2) # Threshold for suspects
}
```

