PRINCETON UNIVERSITY

Getting Started in Linear Regression using R

(with some examples in Stata)

(ver. 0.1-*Draft*)

Oscar Torres-Reyna

Data Consultant

otorres@princeton.edu

http://dss.princeton.edu/training/

R	Stata

Using dataset "Prestige"*

Used in the regression models in the following pages

R

```
# Dataset is in the following library

library(car)

# If not installed type

install.packages("car")

# Type help(Prestige) to access the codebook
```

- ✓ *education*. Average education of occupational incumbents, years, in 1971.
- ✓ *income*. Average income of incumbents, dollars, in 1971.
- ✓ *women. Percentage* of incumbents who are women.
- ✓ *prestige*. Pineo-Porter prestige score for occupation, from a social survey conducted in the mid-1960s.
- ✓ *census* .Canadian Census occupational code.
- ✓ *type*. Type of occupation. A factor with levels (note: out of order): bc, Blue Collar; prof, Professional, Managerial, and Technical; wc, White Collar.

Stata

```
/* Stata version here */

use http://www.ats.ucla.edu/stat/stata/examples/ara/Prestige, clear

/* Renaming/recoding variables to match the dataset's R version*/

rename educat education
rename percwomn women
rename occ_code census
recode occ_type (2=1 "bc")(4=2 "wc")(3=3 "prof")(else=.), gen(type) label(type)
label variable type "Type of occupation"
drop occ_type
replace type=3 if occtitle=="PILOTS"
gen log2income=log10(income)/log10(2)
```

*Fox, J. and Weisberg, S. (2011) *An R Companion to Applied Regression*, Second Edition, Sage.

NOTE: The R content presented in this document is mostly based on an early version of Fox, J. and Weisberg, S. (2011) *An R Companion to Applied Regression*, Second Edition, Sage; and from class notes from the ICPSR's workshop *Introduction to the R Statistical Computing Environment* taught by John Fox during the summer of 2010.

R	Stata
Linear regression	

`# R automatically process the log base 2 of income in the equation`
 	
`reg1 <- lm(prestige ~ education + log2(income) + women, data=Prestige)`
 	
`summary(reg1)`
 	
(See output next page)
	`/* You need to create the log base 2 of income first, type: */`
	
	`gen log2income=log10(income)/log10(2)`
	
	`/* Then run the regression */`
	
	`regress prestige education log2income women`

Linear regression (robust, controlling for heteroskedasticity)	

`reg1.robust <- rlm(prestige ~ education + log2(income) + women, data=Prestige)`
 	
`summary(reg1.robust)`
 	
(See output next page)
	`regress prestige education log2income women, robust`

Predicted values/Residuals	

`# After running the regression`
 	
`prestige_hat <- fitted(reg1) # predicted values`
`as.data.frame(prestige_hat)`
 	
`Prestige_resid <- residuals(reg1) # residuals`
`as.data.frame(prestige_resid)`
	`/* After running the regression */`
	
	`predict prestige_hat   /* Predicted values */`
	
	`predict prestige_resid /* Residuals */`

NOTE: For output interpretation (linear regression) please see http://dss.princeton.edu/training/Regression101.pdf

R	Stata

Linear regression (output)

```
> reg1 <- lm(prestige ~ education + log2(income) + women, data=Prestige)
> summary(reg1)

Call:
lm(formula = prestige ~ education + log2(income) + women, data = Prestige)

Residuals:
     Min       1Q   Median       3Q      Max
-17.3639  -4.4293  -0.1010   4.3160  19.1793

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -110.9658    14.8429  -7.476 3.27e-11 ***
education       3.7305     0.3544  10.527  < 2e-16 ***
log2(income)    9.3147     1.3265   7.022 2.90e-10 ***
women           0.0469     0.0299   1.568     0.12
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.093 on 98 degrees of freedom
Multiple R-squared: 0.8351,     Adjusted R-squared:  0.83
F-statistic: 165.4 on 3 and 98 DF,  p-value: < 2.2e-16
```

```
. regress prestige education log2income women

      Source |       SS       df       MS              Number of obs =     102
-------------+------------------------------           F(  3,    98) =  165.43
       Model |  24965.5409     3  8321.84695           Prob > F      =  0.0000
    Residual |  4929.88524    98  50.3049514           R-squared     =  0.8351
-------------+------------------------------           Adj R-squared =  0.8300
       Total |  29895.4261   101  295.994318           Root MSE      =  7.0926

------------------------------------------------------------------------------
    prestige |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   education |   3.730508    .354383    10.53   0.000     3.027246    4.433769
  log2income |   9.314667   1.326515     7.02   0.000     6.682241    11.94709
       women |   .0468951   .0298989     1.57   0.120    -.0124382    .1062285
       _cons |  -110.9658   14.84293    -7.48   0.000    -140.4211   -81.51052
------------------------------------------------------------------------------
```


NOTE: For output interpretation (linear regression) please see http://dss.princeton.edu/training/Regression101.pdf

R	Stata
Dummy regression with no interactions (analysis of covariance, fixed effects)	

`reg2 <- lm(prestige ~ education + log2(income) + type, data = Prestige)`
 	
`summary(reg2)`
 	
`(See output next page)`
 	
`# Reordering factor variables`
 	
`Prestige$type <- with(Prestige, factor(type, levels=c("bc", "wc", "prof")))`
	`Stata 11.x*`
	
	`regress prestige education log2income  i.type`
	
	`Stata 10.x`
	
	`xi: regress prestige education log2income  i.type`
	
	*See http://www.stata.com/help.cgi?whatsnew10to11

Dummy regression with no interactions (interpretation, see output next page)	

	bc	wc	prof
Intercept	-81.2	-81.2-1.44 = -82.64	-81.2 + 6.75 = -74.45
log2(income)	7.27	7.27	7.27
education	3.28	3.28	3.28

NOTE: "type" is a categorical or factor variable with three options: bc (blue collar), prof (professional, managerial, and technical) and wc (white collar). R automatically recognizes it as factor and treat it accordingly. In Stata you need to identify it with the "i." prefix (in Stata 10.x or older you need to add "xi:")
NOTE: For output interpretation (linear regression) please see http://dss.princeton.edu/training/Regression101.pdf
NOTE: For output interpretation (fixed effects) please see http://dss.princeton.edu/training/Panel101.pdf

R	Stata

Dummy regression with interactions (output)

```
> reg2 <- lm(prestige ~ education + log2(income) + type, data = Prestige)
> summary(reg2)

Call:
lm(formula = prestige ~ education + log2(income) + type, data = Prestige)

Residuals:
    Min      1Q  Median      3Q     Max 
-13.511  -3.746   1.011   4.356  18.438 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -81.2019    13.7431  -5.909 5.63e-08 ***
education      3.2845     0.6081   5.401 5.06e-07 ***
log2(income)   7.2694     1.1900   6.109 2.31e-08 ***
typewc        -1.4394     2.3780  -0.605   0.5465    
typeprof       6.7509     3.6185   1.866   0.0652 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 6.637 on 93 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared: 0.8555,     Adjusted R-squared: 0.8493 
F-statistic: 137.6 on 4 and 93 DF,  p-value: < 2.2e-16
```

```
. regress prestige education log2income  i.type

      Source |       SS       df       MS              Number of obs =      98
-------------+------------------------------           F(  4,    93) =  137.64
       Model |  24250.5893     4  6062.64731           Prob > F      =  0.0000
    Residual |  4096.2858     93  44.0460839           R-squared     =  0.8555
-------------+------------------------------           Adj R-squared =  0.8493
       Total |  28346.8751    97  292.235825           Root MSE      =  6.6367

------------------------------------------------------------------------------
    prestige |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   education |   3.284486    .608097     5.40   0.000     2.076926    4.492046
  log2income |   7.269361   1.189955     6.11   0.000     4.906346    9.632376
             |
        type |
          2  |  -1.439403   2.377997    -0.61   0.546    -6.161635    3.282828
          3  |   6.750887   3.618496     1.87   0.065    -.434729     13.9365
             |
       _cons |  -81.20187   13.74306    -5.91   0.000    -108.4929   -53.91087
------------------------------------------------------------------------------
```

R	Stata
Dummy regression with interactions	

`reg3 <- lm(prestige ~ type*(education + log2(income)), data = Prestige)`
 	
`summary(reg3)`
 	
`(See output next page)`
 	
`# Other ways to run the same model`
 	
`reg3a <- lm(prestige ~ education + log2(income) + type + log2(income):type + education:type, data = Prestige)`
 	
`reg3b <- lm(prestige ~ education*type + log2(income)*type, data = Prestige)`
	`Stata 11.x*`
	
	`regress prestige i.type##c.education i.type##c.log2income`
	
	`Stata 10.x`
	
	`xi: regress prestige i.type*education i.type*log2income`
	
	`*See http://www.stata.com/help.cgi?whatsnew10to11`

Dummy regression with interactions (interpretation, see output next page)	

	bc	wc	prof
Intercept	-120.05	-120.05 +30.24 = -89.81	-120.05 + 85.16 = -34.89
log2(income)	11.08	11.08-5.653 = 5.425	11.08 - 6.536 = 4.542
education	2.34	2.34 + 3.64 = 5.98	2.34 + 0.697 = 3.037

NOTE: "type" is a categorical or factor variable with three options: bc (blue collar), prof (professional, managerial, and technical) and wc (white collar). R automatically recognizes it as factor and treat it accordingly. In Stata you need to identify it with the "i." prefix (in Stata 10.x or older you need to add "xi:")

NOTE: For output interpretation (linear regression) please see http://dss.princeton.edu/training/Regression101.pdf

NOTE: For output interpretation (fixed effects) please see http://dss.princeton.edu/training/Panel101.pdf

R	Stata

Dummy regression with interactions (output)

```
> reg3 <- lm(prestige ~ type*(education + log2(income)), data = Prestige)
> summary(reg3)

Call:
lm(formula = prestige ~ type * (education + log2(income)), data = Prestige)

Residuals:
    Min      1Q  Median      3Q     Max 
-13.970  -4.124   1.206   3.829  18.059 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)           -120.0459    20.1576  -5.955 5.07e-08 ***
typewc                  30.2412    37.9788   0.796  0.42800    
typeprof                85.1601    31.1810   2.731  0.00761 ** 
education                2.3357     0.9277   2.518  0.01360 *  
log2(income)            11.0782     1.8063   6.133 2.32e-08 ***
typewc:education         3.6400     1.7589   2.069  0.04140 *  
typeprof:education       0.6974     1.2895   0.541  0.58998    
typewc:log2(income)     -5.6530     3.0519  -1.852  0.06730 .  
typeprof:log2(income)   -6.5356     2.6167  -2.498  0.01434 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.409 on 89 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared: 0.871,	Adjusted R-squared: 0.8595 
F-statistic: 75.15 on 8 and 89 DF,  p-value: < 2.2e-16
```

```
. regress prestige i.type##c.education i.type##c.log2income

      Source |       SS       df       MS              Number of obs =      98
-------------+------------------------------           F(  8,    89) =   75.15
       Model |  24691.4782     8  3086.43477           Prob > F      =  0.0000
    Residual |  3655.3969     89  41.0718753           R-squared     =  0.8710
-------------+------------------------------           Adj R-squared =  0.8595
       Total |  28346.8751    97  292.235825           Root MSE      =  6.4087

------------------------------------------------------------------------------
    prestige |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        type |
          2  |   30.24117   37.97878     0.80   0.428    -45.22186    105.7042
          3  |   85.16011     31.181     2.73   0.008     23.20414    147.1161
             |
   education |   2.335673    .927729     2.52   0.014      .492295    4.179051
             |
       type# |
 c.education |
          2  |   3.640038   1.758948     2.07   0.041     .1450456     7.13503
          3  |   .6973987   1.289508     0.54   0.590    -1.864827    3.259624
             |
  log2income |   11.07821   1.806298     6.13   0.000     7.489136    14.66729
             |
       type# |
c.log2income |
          2  |  -5.653036   3.051886    -1.85   0.067    -11.71707     .410996
          3  |  -6.535558   2.616708    -2.50   0.014     -11.7349   -1.336215
             |
       _cons |  -120.0459    20.1576    -5.96   0.000    -160.0986   -79.99318
------------------------------------------------------------------------------
```

R

Diagnostics for linear regression (residual plots, see next page for the graph)

```
library(car)

reg1 <- lm(prestige ~ education + income + type,
data = Prestige)

residualPlots(reg1)

           Test stat Pr(>|t|)
education     -0.684    0.496
income        -2.886    0.005
type              NA       NA
Tukey test    -2.610    0.009

# Using ‘income’ as is.
# Variable ‘income’ shows some patterns.

# Other options:

residualPlots(reg1, ~ 1, fitted=TRUE) #Residuals
   vs fitted only

residualPlots(reg1, ~ education, fitted=FALSE) #
   Residuals vs education only
```

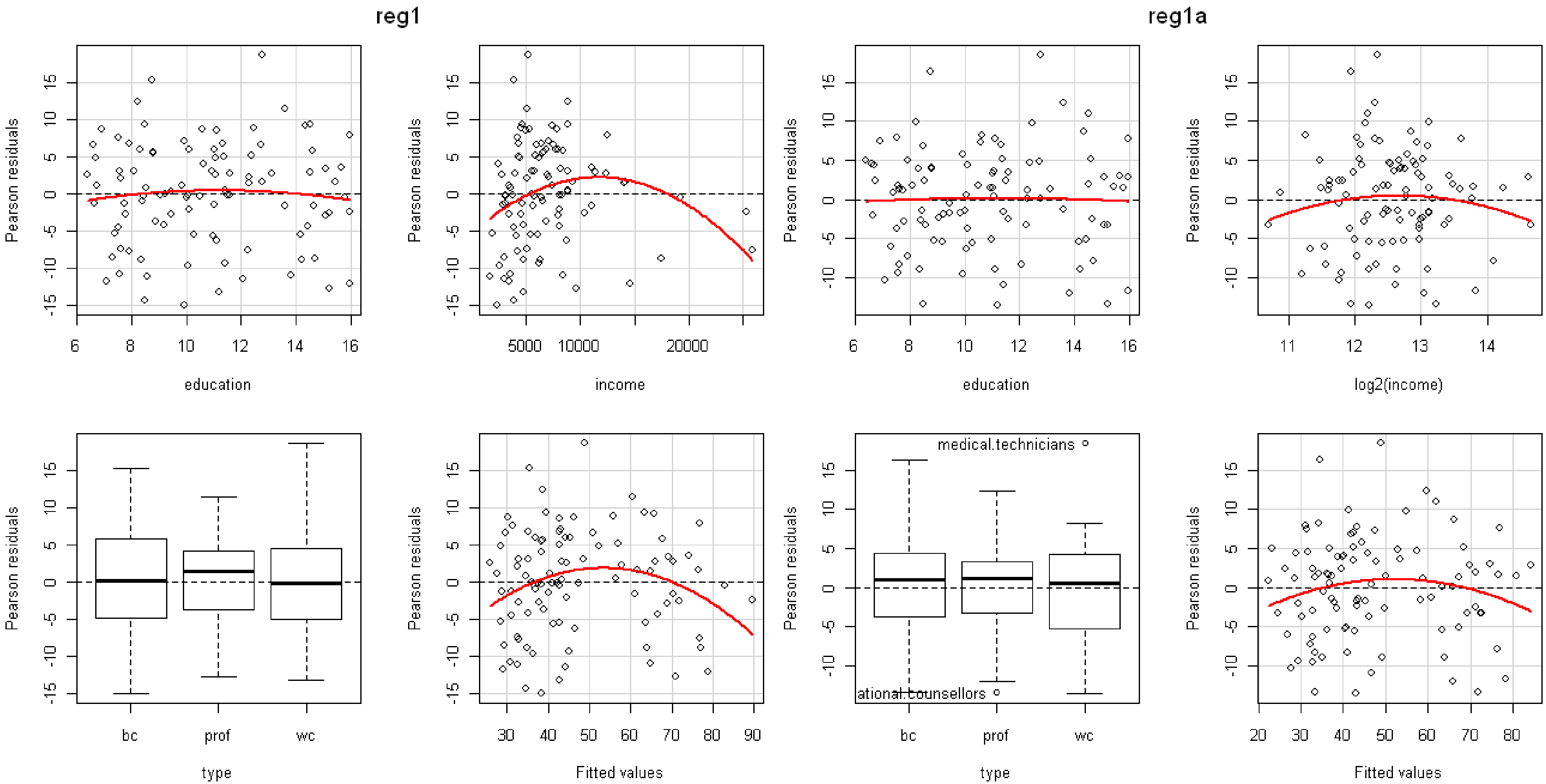
```
library(car)

reg1a <- lm(prestige ~ education + log2(income) +
type, data = Prestige)

residualPlots(reg1a)

             Test stat Pr(>|t|)
education       -0.237    0.813
log2(income)    -1.044    0.299
type                NA       NA
Tukey test      -1.446    0.148

# Using ‘log2(income)’.
# Model looks ok.
```

```
# What to look for: No patterns, no problems.
# All p’s should be non-significant.
# Model ok if residuals have mean=0 and variance=1 (Fox,316)
# Tukey test null hypothesis: model is additive.
```


NOTE: For Stata version please see http://dss.princeton.edu/training/Regression101.pdf

R

Diagnostics for linear regression (residual plots graph)



R

Influential variables - Added-variable plots (see next page for the graph)

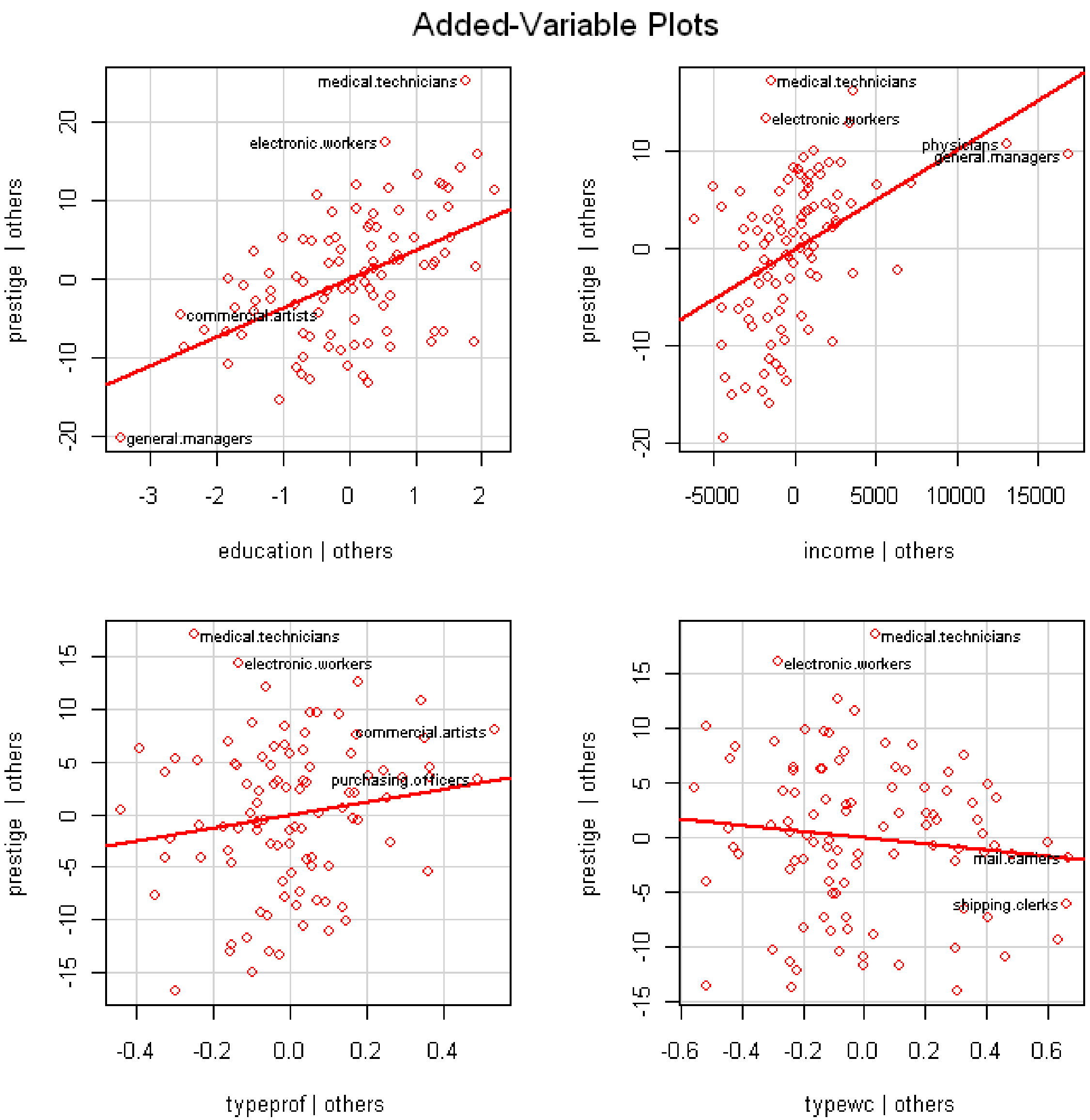
```
library(car)

reg1 <- lm(prestige ~ education + income + type, data = Prestige)

avPlots(reg1, id.n=2, id.cex=0.7)

# id.n – id most influential observation
# id.cex – font size for id.
```

```
# Graphs outcome vs predictor variables holding the rest constant (also called partial-regression plots)
# Help identify the effect (or influence) of an observation on the regression coefficient of the predictor variable
```

NOTE: For Stata version please see http://dss.princeton.edu/training/Regression101.pdf

R

Added-variable plots – Influential variables (graph)



R

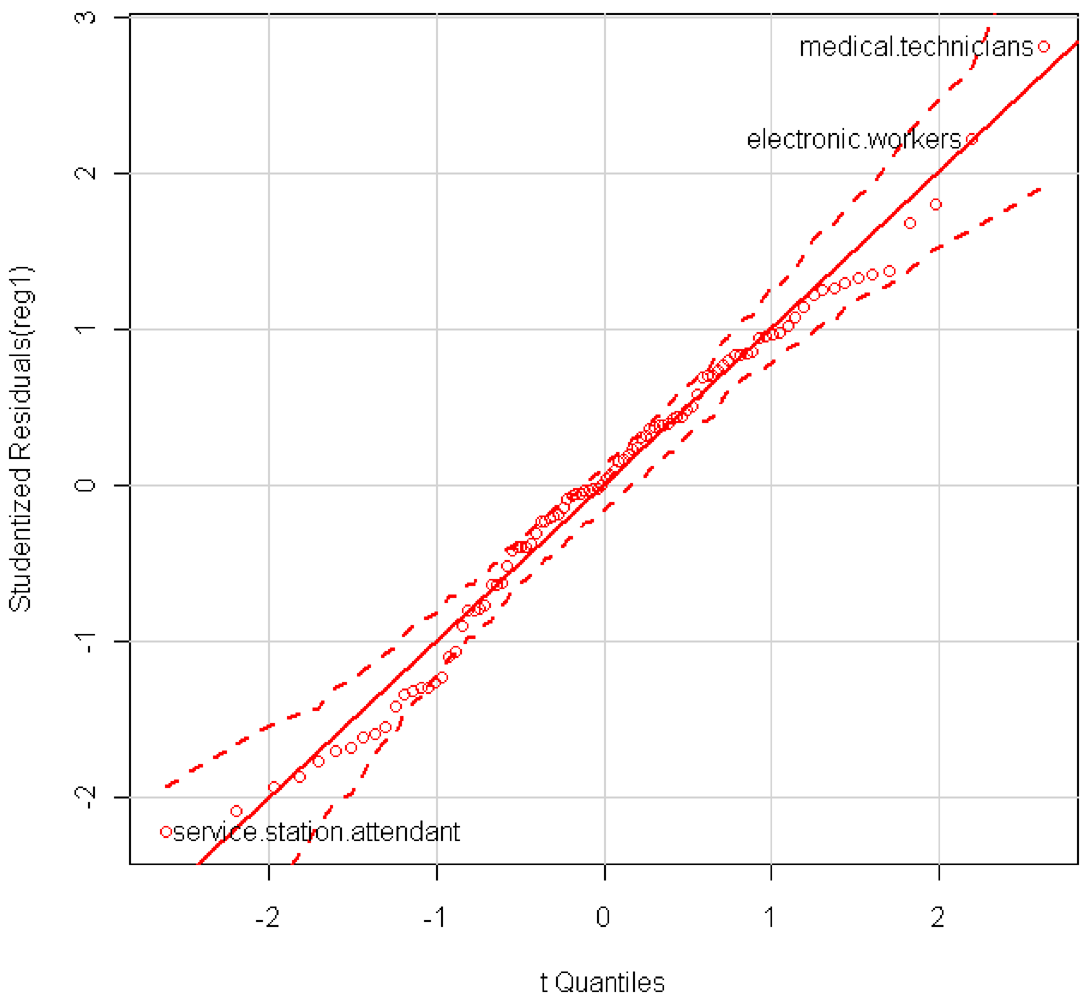
Outliers – QQ-Plots (see next page for the graph)

```
library(car)

reg1 <- lm(prestige ~ education + income + type, data = Prestige)

qqPlot(reg1, id.n=3)

[1] "medical.technicians"       "electronic.workers"
[3] "service.station.attendant"

# id.n – id observations with high residuals
```

NOTE: For Stata version please see http://dss.princeton.edu/training/Regression101.pdf

R

Added-variable plots – Influential variables (graph)



R

Outliers – Bonferonni test

```
library(car)

reg1 <- lm(prestige ~ education + income + type, data = Prestige)

outlierTest(reg1)

     No Studentized residuals with Bonferonni p < 0.05
     Largest |rstudent|:
                          rstudent unadjusted p-value Bonferonni p
     medical.technicians 2.821091          0.0058632      0.57459

# Null for the Bonferonni adjusted outlier test is the observation is an outlier. Here observation
      related to 'medical.technicians' is an outlier.
```

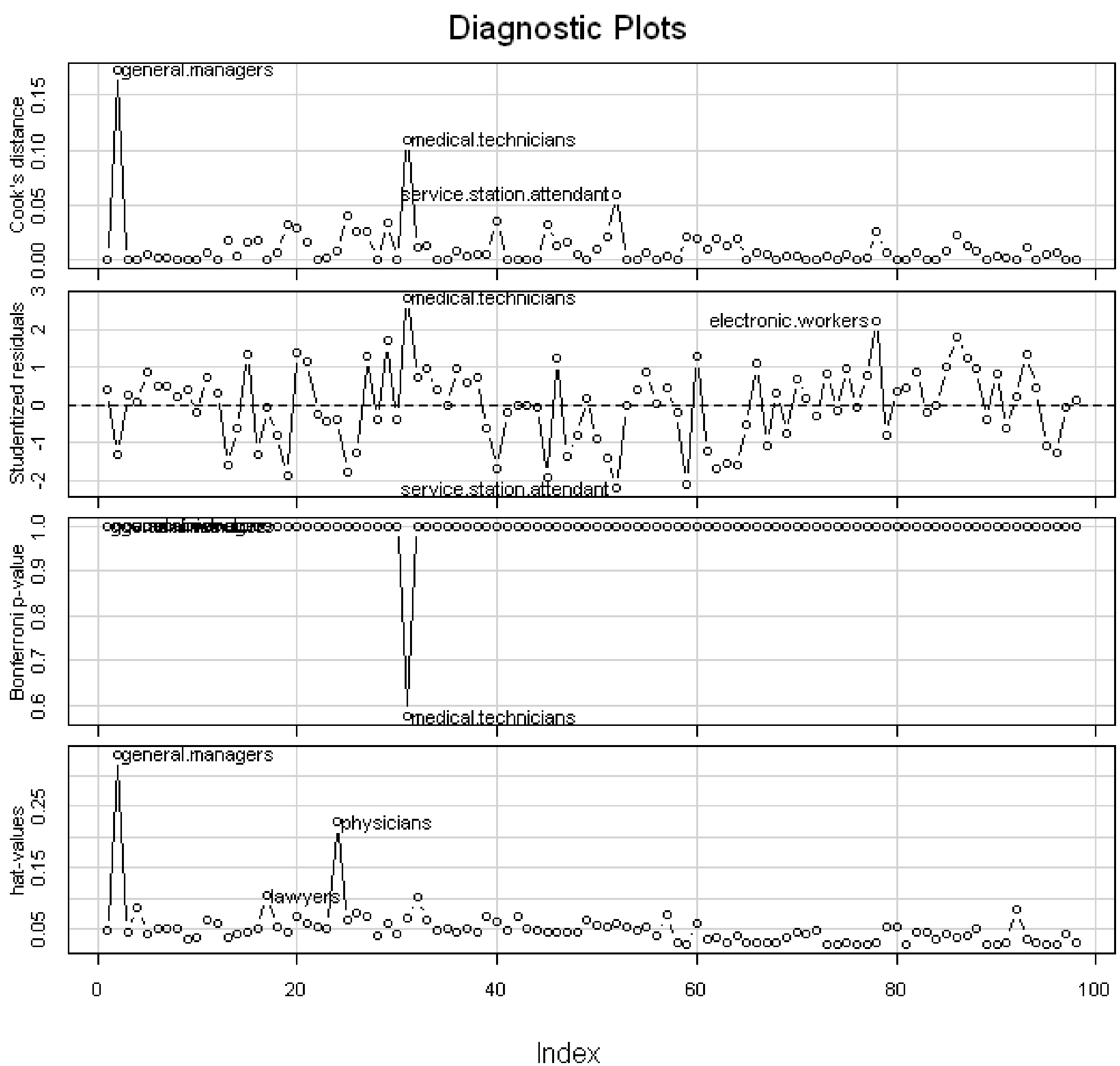
High leverage (*hat*) points (graph next page)

```
library(car)

reg1 <- lm(prestige ~ education + income + type, data = Prestige)

influenceIndexPlot(reg1, id.n=3)

# Cook's distance measures how much an observation influences the overall model or predicted values
# Studentizided residuals are the residuals divided by their estimated standard deviation as a way to
      standardized
# Bonferroni test to identify outliers
# Hat-points identify influential observations (have a high impact on the predictor variables)

NOTE:  If an observation is an outlier and influential (high leverage) then that observation can change the fit
       of the linear model, it is advisable to remove it. To remove a case(s) type
reg1a <- update(prestige.reg4, subset=rownames(Prestige) != "general.managers")
reg1b <- update(prestige.reg4, subset= !(rownames(Prestige) %in% c("general.managers","medical.technicians")))
```

NOTE: For Stata version please see http://dss.princeton.edu/training/Regression101.pdf

R

High leverage (*hat*) points (graph)

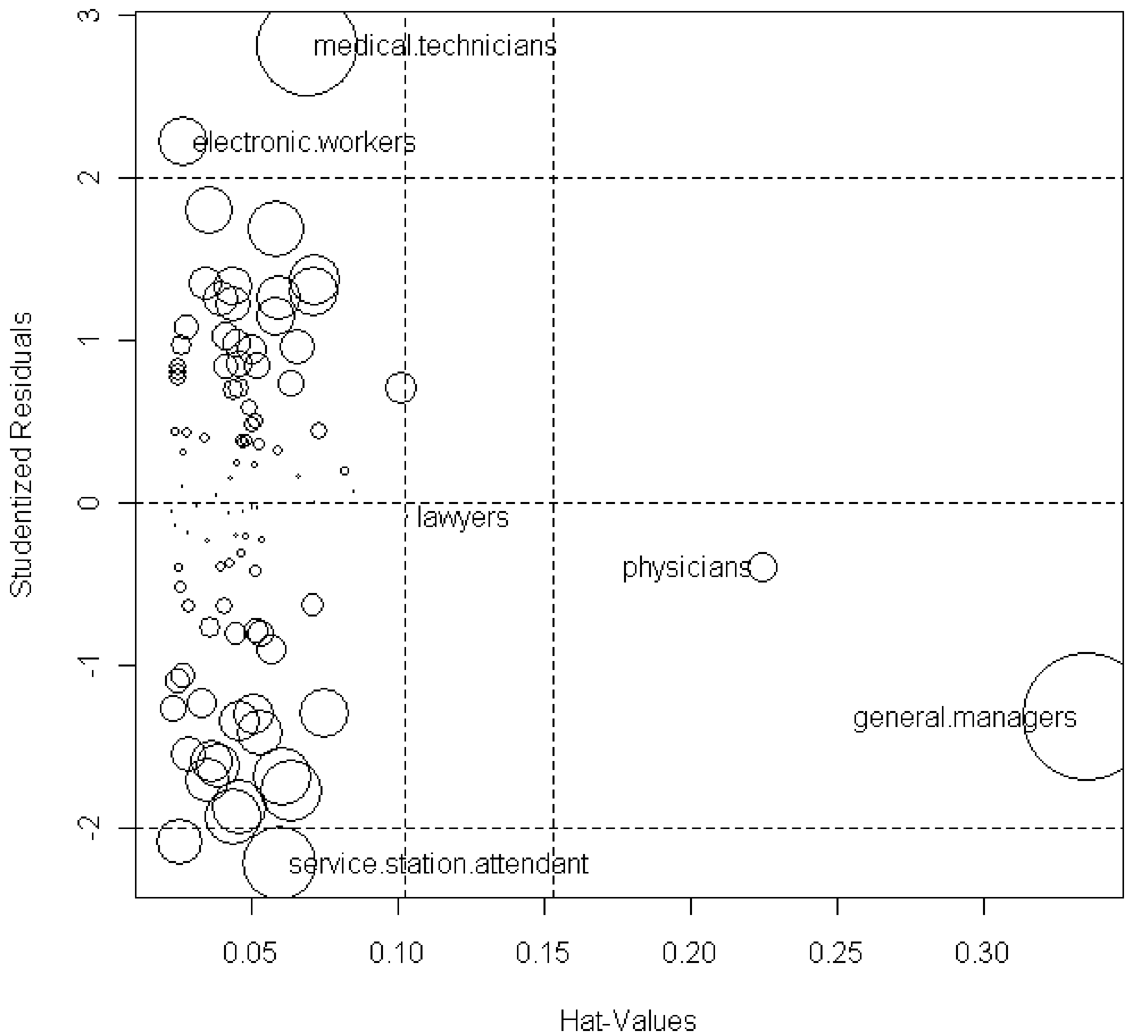


R
Influence Plots (see next page for a graph)

`library(car)`
 	
`reg1 <- lm(prestige ~ education + income + type, data = Prestige)`
 	
`influencePlot(reg1, id.n=3)`
 	
`# Creates a bubble-plot combining the display of Studentized residuals, hat-values, and Cook's distance (represented in the circles).`

R

Influence plot



R

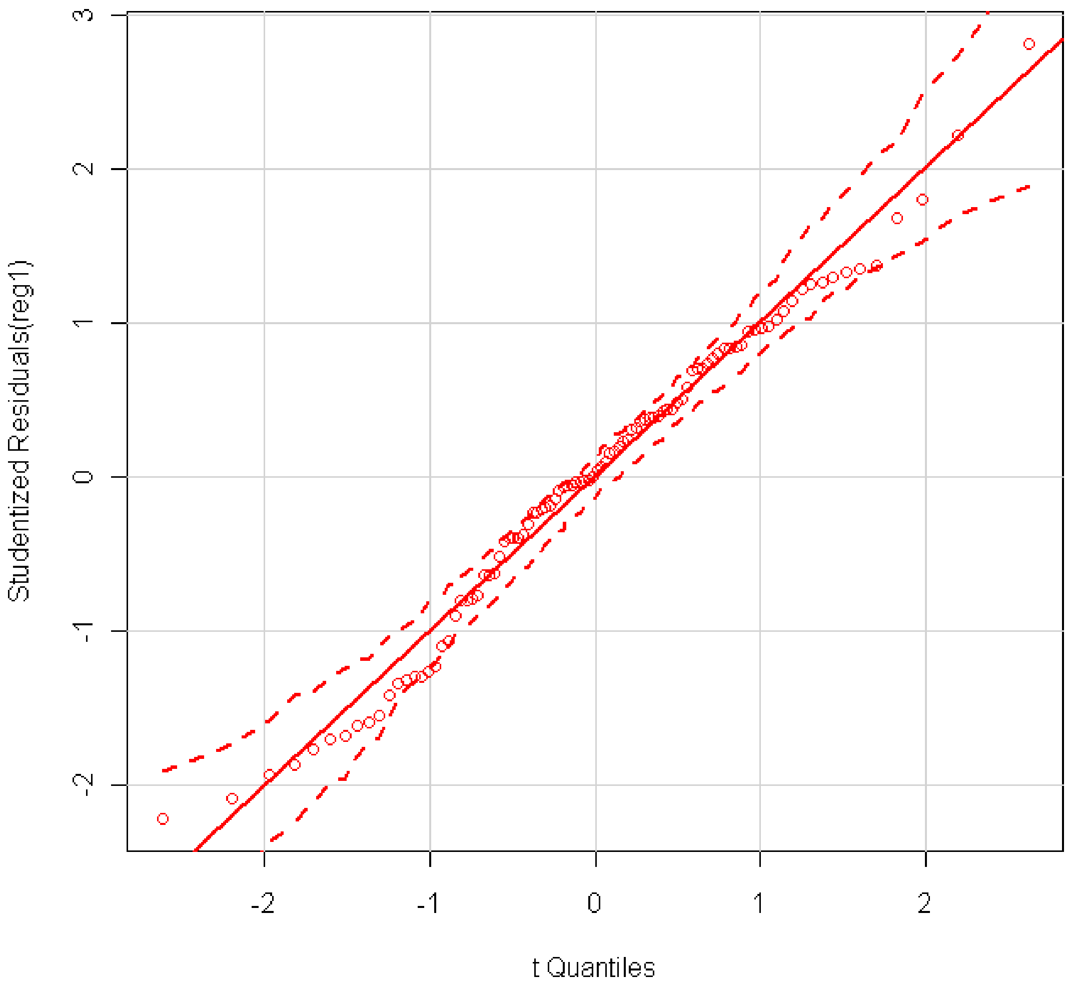
Testing for normality (see graph next page)

```
library(car)

reg1 <- lm(prestige ~ education + income + type, data = Prestige)

qqPlot(reg1)

# Look for the tails, points should be close to the line or within the confidence intervals.
# Quantile plots compare the Studentized residuals vs a t-distribution
# Other tests: shapiro.test(), mshapiro.test() in library(mvnormtest)-library(ts)
```

NOTE: For Stata version please see http://dss.princeton.edu/training/Regression101.pdf

R

Influence plot



R

Testing for heteroskedasticity

```
library(car)

reg1 <- lm(prestige ~ education + income + type, data = Prestige)

ncvTest(reg1)

     Non-constant Variance Score Test
     Variance formula: ~ fitted.values
     Chisquare = 0.09830307    Df = 1     p = 0.7538756

# Breush/Pagan and Cook/Weisberg score test for non-constant error variance. Null is constant variance
# See also residualPlots(reg1).
```

NOTE: For Stata version please see http://dss.princeton.edu/training/Regression101.pdf

R

Testing for multicolinearity

```
library(car)

reg1 <- lm(prestige ~ education + income + type, data = Prestige)

vif(reg1)

                 GVIF     Df    GVIF^(1/(2*Df))
     education 5.973932   1        2.444163
     income    1.681325   1        1.296659
     type      6.102131   2        1.571703


# A gvif> 4 suggests collinearity.

# “When there are strong linear relationships among the predictors in a regression analysis, the
      precision of the estimated regression coefficients in linear models declines compared to what it
      would have been were the predictors uncorrelated with each other” (Fox:359)
```

NOTE: For Stata version please see http://dss.princeton.edu/training/Regression101.pdf

References/Useful links

- DSS Online Training Section http://dss.princeton.edu/training/
- Princeton DSS Libguides http://libguides.princeton.edu/dss
- John Fox’s site http://socserv.mcmaster.ca/jfox/
- Quick-R http://www.statmethods.net/
- UCLA Resources to learn and use R http://www.ats.ucla.edu/stat/R/
- UCLA Resources to learn and use Stata http://www.ats.ucla.edu/stat/stata/
- DSS - Stata http://dss/online_help/stats_packages/stata/
- DSS - R http://dss.princeton.edu/online_help/stats_packages/r

References/Recommended books

- *An R Companion to Applied Regression*, Second Edition / John Fox , Sanford Weisberg, Sage Publications, 2011
- *Data Manipulation with R* / Phil Spector, Springer, 2008
- *Applied Econometrics with R* / Christian Kleiber, Achim Zeileis, Springer, 2008
- *Introductory Statistics with R* / Peter Dalgaard, Springer, 2008
- *Complex Surveys. A guide to Analysis Using R* / Thomas Lumley, Wiley, 2010
- *Applied Regression Analysis and Generalized Linear Models* / John Fox, Sage, 2008
- *R for Stata Users* / Robert A. Muenchen, Joseph Hilbe, Springer, 2010
- *Introduction to econometrics* / James H. Stock, Mark W. Watson. 2nd ed., Boston: Pearson Addison Wesley, 2007.
- *Data analysis using regression and multilevel/hierarchical models* / Andrew Gelman, Jennifer Hill. Cambridge ; New York : Cambridge University Press, 2007.
- *Econometric analysis* / William H. Greene. 6th ed., Upper Saddle River, N.J. : Prentice Hall, 2008.
- *Designing Social Inquiry: Scientific Inference in Qualitative Research* / *Gary King, Robert O. Keohane, Sidney Verba, Princeton University Press*, 1994.
- *Unifying Political Methodology: The Likelihood Theory of Statistical Inference* / Gary King, Cambridge University Press, 1989
- *Statistical Analysis: an interdisciplinary introduction to univariate & multivariate methods /* Sam Kachigan, New York : Radius Press, c1986
- *Statistics with Stata (updated for version 9) /* Lawrence Hamilton, Thomson Books/Cole, 2006