

R: User Guide

September 2016

Table of Contents

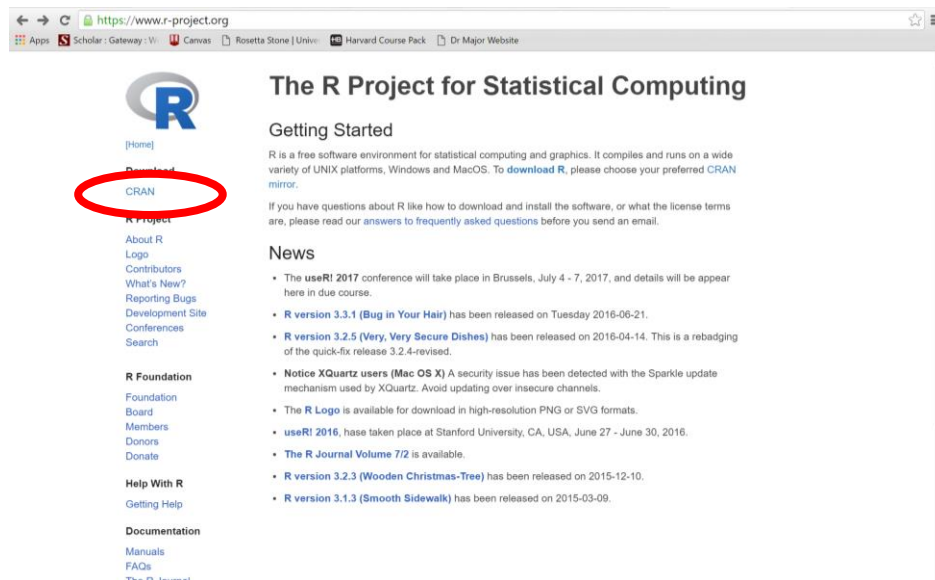
Administration	2
Downloading R	2
Downloading RStudio	5
Importing Data	7
Linear Model Using All Metrics	8
Part I	8
Part II: Plots	9
Part III.....	12
Linear Model Stepwise Fit	13
Linear Model Using Stepwise Metrics	14
Part I	14
Part II: Plots	15

Administration

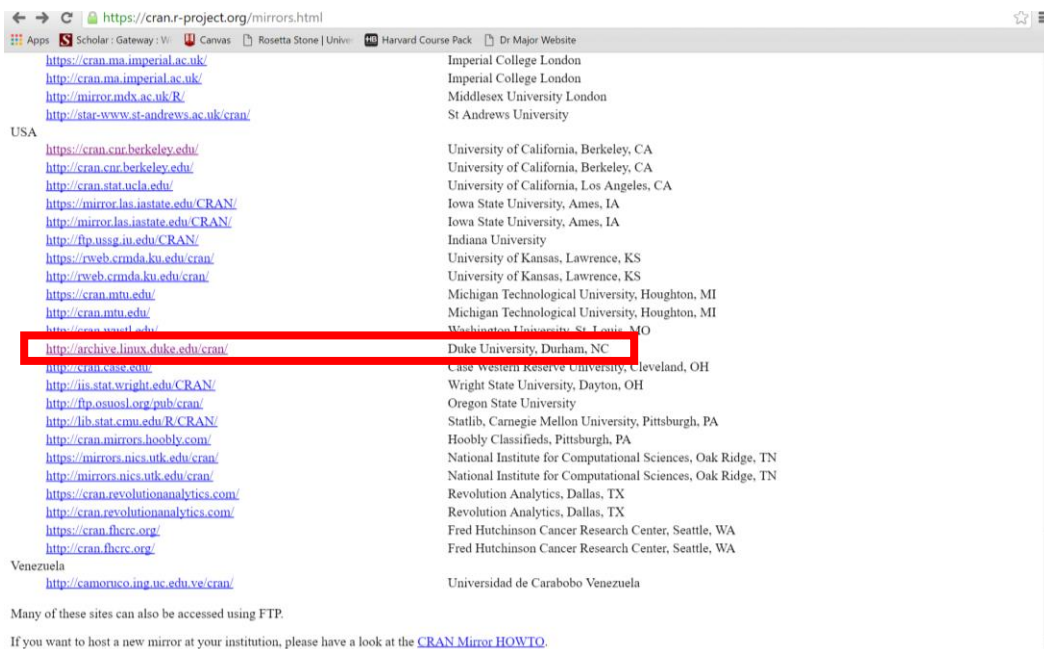
In order to perform this exercise, two software packages will be downloaded, R itself as well as RStudio – a developmental environment for you to work in.

Downloading R

1. R is an open source software. To download it, go to <https://www.r-project.org/>
**Note: Make sure you have the administrative privileges on your computer.
2. On the left side menu click on “CRAN”



3. Scroll down to find USA, and choose a location closest to you. We will choose Duke University in Durham, NC.



4. Choose your operating system

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages. **Windows and Mac** users most likely want one of these versions of R:

- [Download R for Linux](#)
- [Download R for \(Mac\) OS X](#)
- [Download R for Windows](#)

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

5. For Mac, choose the latest version of r to install. Click on the link and wait till the Installer program downloads.

archive.linux.duke.edu/cran/

R for Mac OS X

This directory contains binaries for a base distribution and packages to run on Mac OS X (release 10.6 and above). Mac OS 8.6 to 9.2 (and Mac OS 10.1) are no longer supported but you can find the last supported release of R for these systems (which is R 1.7.1) [here](#). Releases for old Mac OS systems (through Mac OS X 10.5) and PowerPC Macs can be found in the [old](#) directory.

Note: CRAN does not have Mac OS X systems and cannot check these binaries for viruses. Although we take precautions when assembling binaries please use the normal precautions with downloaded executables.

As of 2016/03/01 package binaries for R versions older than 2.12.0 are only available from the [CRAN archive](#) so users of such versions should a CRAN mirror setting accordingly.

R 3.3.1 "Bug in Your Hair" released on 2016/06/21

Please check the MD5 checksum of the downloaded image to ensure that it has not been tampered with or corrupted during the mirroring process:

example type
md5 R-3.3.1.pkg
in the *Terminal* application to print the MD5 checksum for the R-3.3.1.pkg image. On Mac OS X 10.7 and later you can also validate the signature:
`pkgutil --check-signature R-3.3.1.pkg`

Files:

R 3.3.1 binary for Mac OS X 10.9 (Mavericks) and higher, signed package. Contains R 3.3.1 framework, R.app GUI 1.68 in 64-bit for Intel Macs, Tcl/Tk 8.6.0 X11 libraries and Texinfo 5.2. The latter two components are optional and can be omitted when choosing "custom install", it is only needed if you want to use the `tc1tk` package or build package documentation from sources.

Note: the use of X11 (including `tc1tk`) requires [XQuartz](#) to be installed since it is no longer part of OS X. Always re-install XQuartz when upgrading your OS X to a new major version.

R 3.2.1 legacy binary for Mac OS X 10.6 (Snow Leopard) - 10.8 (Mountain Lion), signed package. Contains R 3.2.1 framework, R.app GUI 1.66 in 64-bit for Intel Macs. This package contains the R framework, 64-bit GUI (R.app), Tcl/Tk 8.6.0 X11 libraries and Texinfo 5.2. GNU Fortran is **NOT** included (needed if you want to compile packages from sources that contain FORTRAN code) please see [the tools directory](#).
NOTE: the binary support for OS X before Mavericks is being phased out, we do not expect further releases!

R-3.3.1.pkg
MD5-checksum: 1281f220e7692c1622ab49d4c066cd811b06bd19
SHA1-hash: b6e91d912bac22a324f0c651c7ef9063eece0d0
(ca. 71MB)

R-3.2.1-snowleopard.pkg
MD5-checksum: 31669d01314496b375f00c0f89146565
SHA1-hash: b6e91d912bac22a324f0c651c7ef9063eece0d0
(ca. 68MB)

CRAN
[Mirrors](#)
[What's new?](#)
[Task Views](#)
[Search](#)

About R
[R Homepage](#)
[The R Journal](#)

Software
[R Sources](#)
[R Binaries](#)
[Packages](#)
[Other](#)

Documentation
[Manuals](#)
[FAQs](#)
[Contributed](#)

For Windows, click on “Installing for the first time” link

ce.edu/cran/

Canvas Rosetta Stone | Univer Harvard Course Pack Dr Major Website

R for Windows

Subdirectories:

- [base](#) Binaries for base distribution (managed by Duncan Murdoch). This is what you want to [install R for the first time](#).
- [contrib](#) Binaries of contributed CRAN packages (for R \geq 2.11.x; managed by Uwe Ligges). There is also information on [third party software](#) available for CRAN Windows services and corresponding environment and make variables.
- [old.contrib](#) Binaries of contributed CRAN packages for outdated versions of R (for R $<$ 2.11.x; managed by Uwe Ligges).
- [Rtools](#) Tools to build R and R packages (managed by Duncan Murdoch). This is what you want to build your own packages on Windows, or to build R itself.

Please do not submit binaries to CRAN. Package developers might want to contact Duncan Murdoch or Uwe Ligges directly in case of questions / suggestions related to Windows binaries.

You may also want to read the [R FAQ](#) and [R for Windows FAQ](#).

Note: CRAN does some checks on these binaries for viruses, but cannot give guarantees. Use the normal precautions with downloaded executables.

Click on the top link to download the Installer.

R-3.3.1 for Windows (32/64 bit)

[Download R 3.3.1 for Windows](#) (70 megabytes, 32/64 bit)

[Download instructions for this version](#)

[New features in this version](#)

If you want to double-check that the package you have downloaded exactly matches the package distributed by R, you can compare the [md5sum](#) of the .exe to the [true fingerprint](#). You will need a version of md5sum for windows: both [graphical](#) and [command line versions](#) are available.

Frequently asked questions

- [Does R run under my version of Windows?](#)
- [How do I update packages in my previous version of R?](#)
- [Should I run 32-bit or 64-bit R?](#)

Please see the [R FAQ](#) for general information about R and the [R Windows FAQ](#) for Windows-specific information.

Other builds

- Patches to this release are incorporated in the [r-patched snapshot build](#).
- A build of the development version (which will eventually become the next major release of R) is available in the [r-devel snapshot build](#).
- [Previous releases](#)

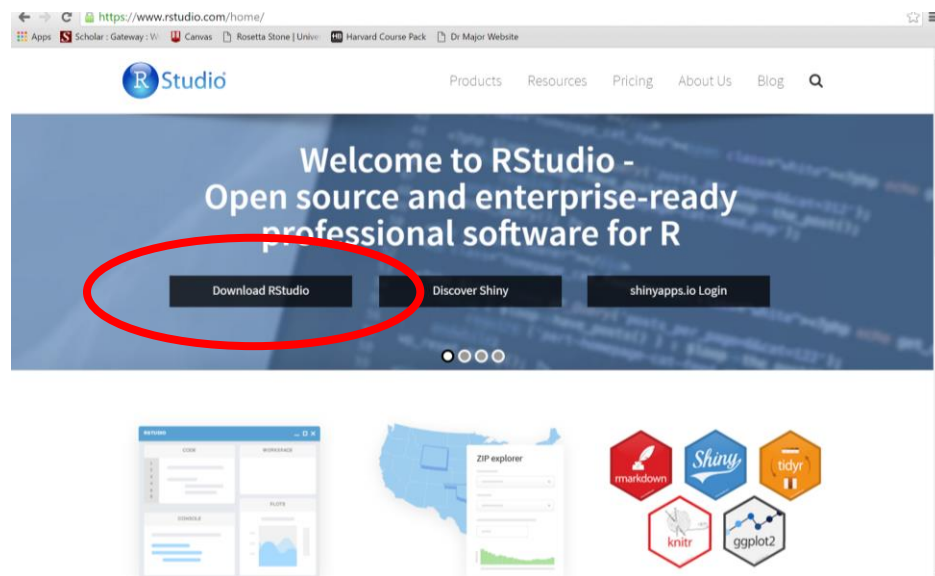
Note to webmasters: A stable link which will redirect to the current Windows binary release is [<CRAN MIRROR>/bin/windows/base/release.htm](#).

Last change: 2016-06-21, by Duncan Murdoch

- Run the Installer. Follow the directions from the Installer Wizard to install the software. The developers recommend for you to keep all the default settings.
 - **Note: R is available for both 32-bit and 64-bit systems. You will have the option to choose if you want to install one or both of the versions. Developers recommend to go with 64-bit version if your system is 64-bit, and 32-bit if your system is 32-bit.

Downloading RStudio

1. Go to <https://www.rstudio.com/home>
2. Click on download



3. Click to download the desktop version

The screenshot shows the "Choose Your Version of RStudio" page with the following details:

RStudio is a set of integrated tools designed to help you be more productive with R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management. [Learn More](#)

	RStudio Desktop (Free License)	RStudio Desktop (Commercial License)	RStudio Server (Free License)	RStudio Server Pro (Commercial License)
Integrated Development Environment for R	✓	✓	✓	✓
Priority support		✓		✓
Access via Web Browser			✓	✓
Enterprise Security and Access Controls				✓
Project Sharing				✓
Access to Multiple Versions of R				✓
Multiple Concurrent Sessions				✓
License	AGPL	Commercial	AGPL	Commercial
Pricing	FREE	\$995/yr*	FREE	\$9,995/yr*
	DOWNLOAD	BUY NOW	DOWNLOAD	DOWNLOAD
	Learn More	Learn More	Learn More	Learn More (45 Day Free Trial)

- Choose an appropriate installer for your system and run it after it is downloaded.

Installers for Supported Platforms

Installers	Size	Date	MD5
RStudio 0.99.903 - Windows Vista/7/8/10	77.1 MB	2016-07-18	716f28f2143c5e21f4acea5752e284f8
RStudio 0.99.903 - Mac OS X 10.6+ (64-bit)	60 MB	2016-07-18	d14a1585b5a5ac0839507b9c04d460d6
RStudio 0.99.903 - Ubuntu 12.04+/Debian 8+ (32-bit)	81.6 MB	2016-07-18	761eae80b0ba4d4cd9051a802a2c44e2
RStudio 0.99.903 - Ubuntu 12.04+/Debian 8+ (64-bit)	88.3 MB	2016-07-18	98ea59d3db00e0083d3e4053514f764d
RStudio 0.99.903 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (32-bit)	81 MB	2016-07-18	ce2ea1023d99175cb909def0fe66eba7
RStudio 0.99.903 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (64-bit)	81.9 MB	2016-07-18	152f247255e86904cf3354afbc7b3b99

Zip/Tarballs

Zip/tar archives	Size	Date	MD5
RStudio 0.99.903 - Windows Vista/7/8/10	110.6 MB	2016-07-18	53817c5703a5fefbba513e6d05133e1d
RStudio 0.99.903 - Ubuntu 12.04+/Debian 8+ (32-bit)	82.3 MB	2016-07-18	bc2c16be996ed08200f1fde7b9e2b93a
RStudio 0.99.903 - Ubuntu 12.04+/Debian 8+ (64-bit)	89.2 MB	2016-07-18	44c418d506e395c70416df458b0788b2
RStudio 0.99.903 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (32-bit)	81.6 MB	2016-07-18	c85a4e536fb71189744fba7aec9e35b5
RStudio 0.99.903 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (64-bit)	82.8 MB	2016-07-18	ad5761417fa07cc4db7dfb91aa535b5a

- Click through the Installer Wizard. You can keep all the default settings.
 ** Note: RStudio might not create a desktop icon, but you can locate it by using search option on your computer.

Importing Data

1. Open your data file in Excel.
2. Make sure your active tab is Student_Data
3. Go to Save As, and save the file as ComputationsCSV.csv (*.csv – comma delimited).

**Note: .csv means that values are separated by commas (thus csv – comma separated values).
4. You will be informed that .csv format does not support workbooks. Click OK as you will only need your student data from the active sheet.
5. In RStudio, type the following command

```
> Comp<-read.csv(file.choose(),header=TRUE)
```

6. This command will bring up a pop up window, where you can find our ComputationsCSV.csv file. Click Open to import the data.
7. Type the following command:

```
> attach(Comp)
```

Linear Model Using All Metrics

Part I

Goal:

```

> fm<- lm(Num_Semem_To ~ AvgGPACHg_To + Base_GPA + AvgDelta_AttPass_To + AvgQualPts_To + AvgProGpa_To + Age + Race, data = Comp)
> coef(fm)
      (Intercept)      AvgGPACHg_To      Base_GPA AvgDelta_AttPass_To
    -0.157679954      0.373707866     -0.350222467      0.419670446
    AvgQualPts_To      AvgProGpa_To      Age      RaceBlack
     0.002312124      1.275342790     0.088069632     0.425442264
    RaceHispanic      RaceIndian      Racewhite
    -0.784484172      4.640008082     -0.584089551
>
> formula(fm)
Num_Semem_To ~ AvgGPACHg_To + Base_GPA + AvgDelta_AttPass_To +
  AvgQualPts_To + AvgProGpa_To + Age + Race
> |

```

Directions:

1. Make sure you have performed the steps above to import the data.
2. Assign variable fm with the following command

```

> fm<- lm(Num_Semem_To~AvgGPACHg_To+Base_GPA+AvgDelta_AttPass_To+AvgQualPts_To+AvgProGpa_To+Age+Race, data=Comp)

```

3. Next, enter command:

```

> coef(fm)

```

Press Enter.

4. Enter command:

```

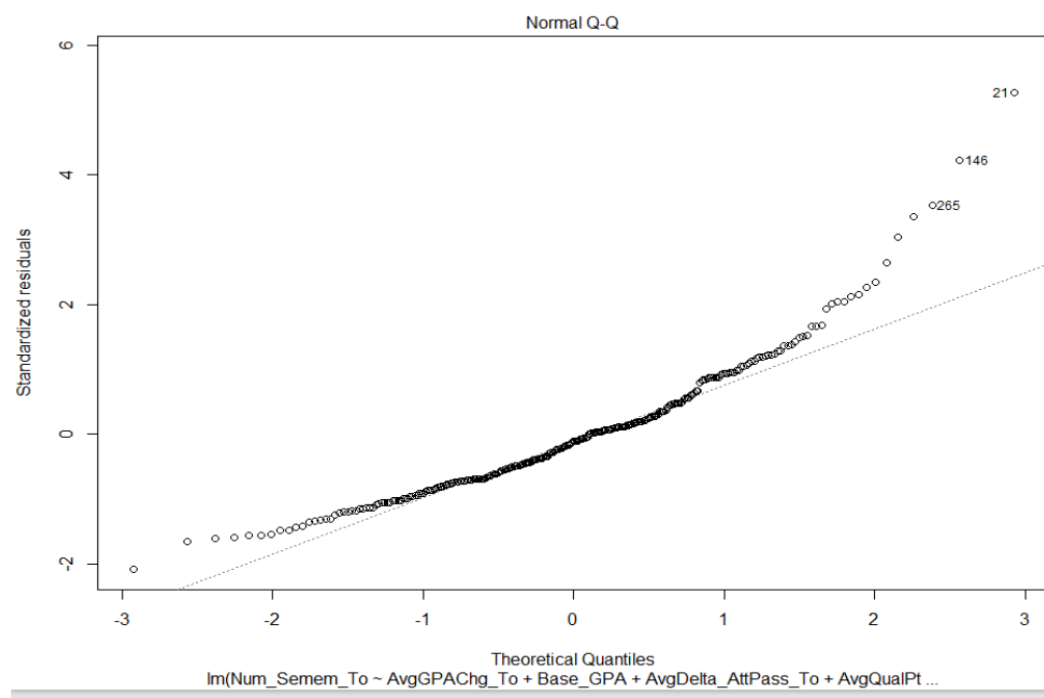
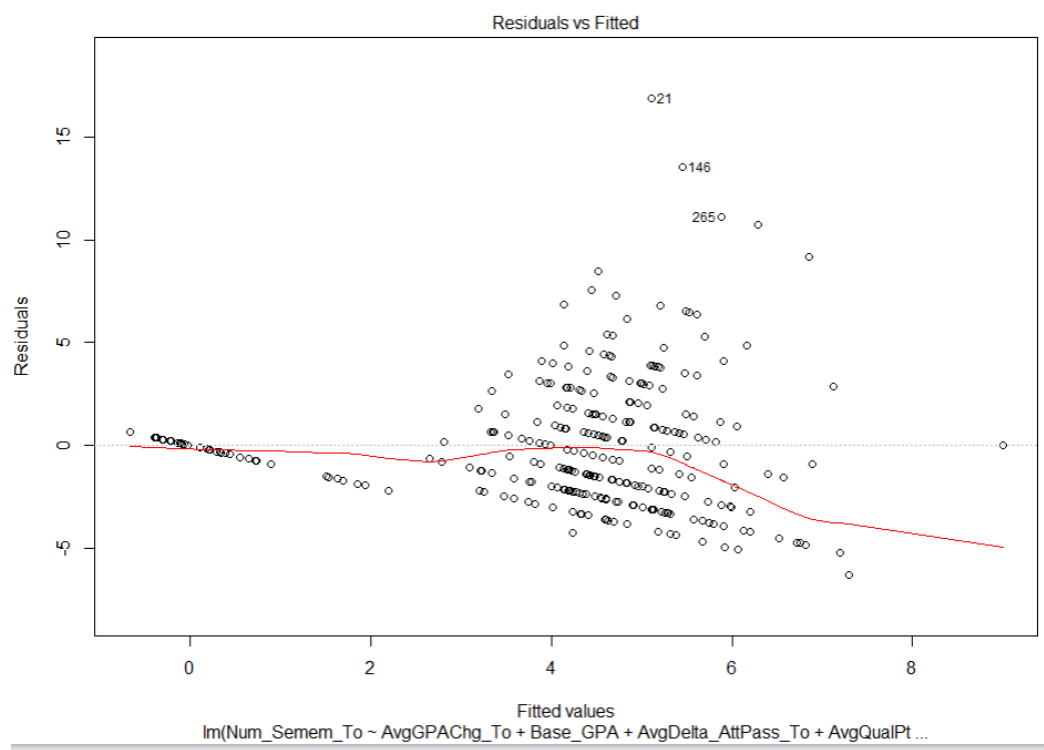
> formula(fm)

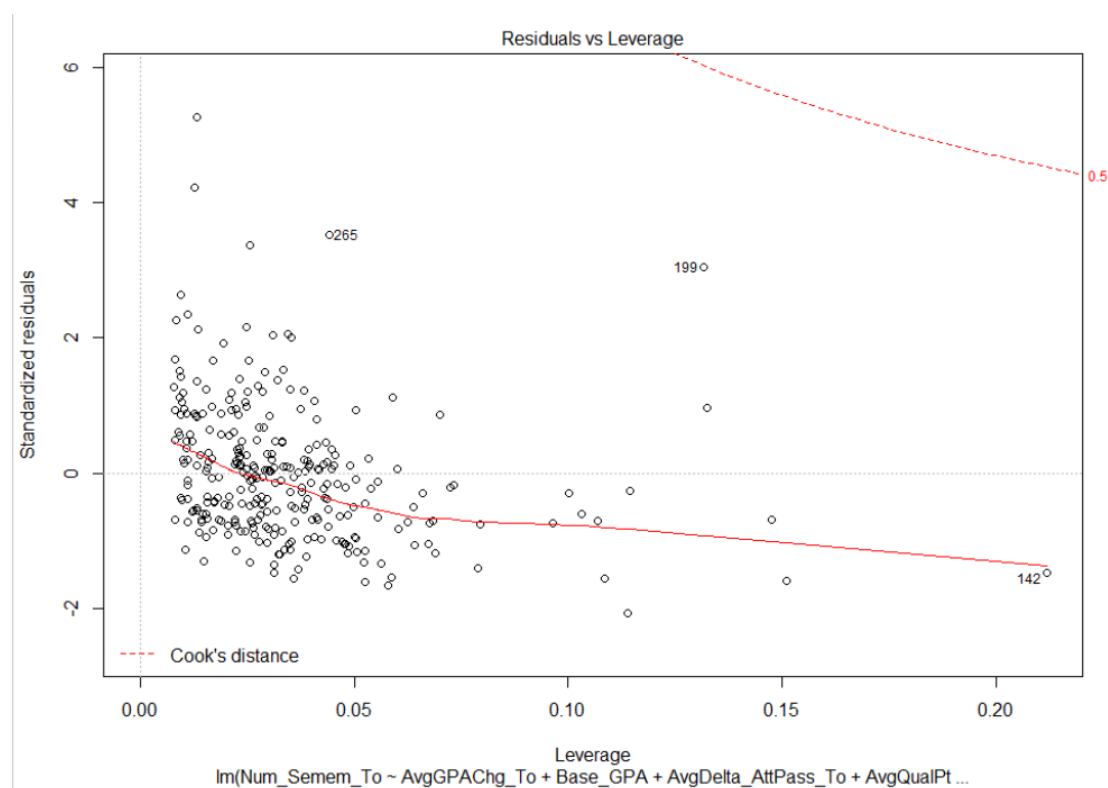
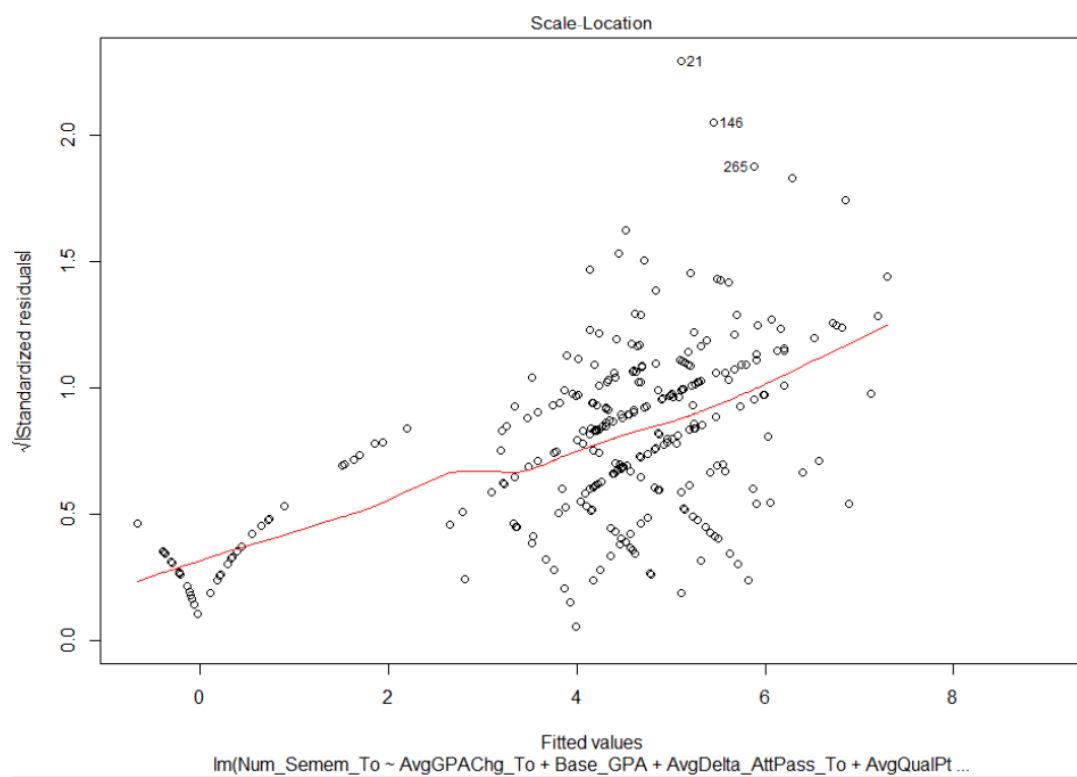
```

Press Enter.

Part II: Plots

Goal:





Directions:

1. Enter the following command and press enter

```
> plot(fm)
```

2. You will see the first graph pop up on the bottom right console. Click Enter to get the next graph. There will be 4 graphs in total.

****Note:** Make sure you switch the tab in the bottom right console to Plots.

3. You can scroll between graphs with forward and backward button on the Plot console

The screenshot displays the RStudio interface. The top-left pane shows a data frame with columns: Num_Semem_To, AvgGPAChg_To, Base_GPA, AvgDelta_AttPass_To, AvgQualPts_To, AvgProGpa_To, Age, and Race. The bottom-left pane shows the console output for the command `summary(fm)`, including the model call, residuals, coefficients, and statistical tests. The bottom-right pane shows the 'Plots' tab with a 'Residuals vs Leverage' plot. A red circle highlights the navigation buttons (left and right arrows) in the plot console. Red arrows point from text boxes to these buttons and the 'Plots Tab' label.

Plots Tab

Forward and Backward buttons

Residuals vs Leverage

Standardized residuals

Leverage

Im(Num_Semem_To ~ AvgGPAChg_To + Base_GPA + AvgDelta_AttPass_To + AvgQualPt ...)

--- Cook's distance

0.5

142

199

0.285

Part III

Goal:

```

> summary(fm)
Call:
lm(formula = Num_Semem_To ~ AvgGPACHg_To + Base_GPA + AvgDelta_AttPass_To +
    AvgQualPts_To + AvgProgpa_To + Age + Race, data = Comp)

Residuals:
    Min       1Q   Median       3Q      Max
-6.3005 -2.2118 -0.3521  1.5119 16.8932

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.157680   1.347873  -0.117  0.906956
AvgGPACHg_To    0.373708   0.384397   0.972  0.331790
Base_GPA       -0.350222   0.270245  -1.296  0.196058
AvgDelta_AttPass_To 0.419670   0.125148   3.353  0.000908 ***
AvgQualPts_To  0.002312   0.033839   0.068  0.945573
AvgProgpa_To   1.275343   0.385275   3.310  0.001054 **
Age            0.088070   0.033681   2.615  0.009410 **
RaceBlack      0.425442   0.783457   0.543  0.587538
RaceHispanic   -0.784484   0.759625  -1.033  0.302620
RaceIndian     4.640008   3.370043   1.377  0.169658
Racewhite     -0.584090   0.681027  -0.858  0.391811
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.229 on 281 degrees of freedom
Multiple R-squared:  0.2277, Adjusted R-squared:  0.2002
F-statistic: 8.284 on 10 and 281 DF, p-value: 8.891e-12

>

```

Directions:

1. To get to our goal screen enter the following command and press Enter:

```
> summary(fm)
```

Linear Model Stepwise Fit

Goal:

```
> step(fm)
Start: AIC=695.39
Num_Semem_To ~ AvgGPACHg_To + Base_GPA + AvgDelta_AttPass_To +
  AvgQualPts_To + AvgProgpa_To + Age + Race

- AvgQualPts_To      Df Sum of Sq  RSS   AIC
- AvgGPACHg_To       1    0.049 2930.4 693.40
- AvgGPACHg_To       1    9.857 2940.3 694.37
- Race               4   76.913 3007.3 694.96
- Base_GPA           1   17.514 2947.9 695.13
<none>              2930.4 695.39
- Age                1   71.301 3001.7 700.41
- AvgProgpa_To       1  114.270 3044.7 704.56
- AvgDelta_AttPass_To 1  117.272 3047.7 704.85

Step: AIC=693.4
Num_Semem_To ~ AvgGPACHg_To + Base_GPA + AvgDelta_AttPass_To +
  AvgProgpa_To + Age + Race

- AvgGPACHg_To       1    10.85 2941.3 692.48
- Race               4   76.90 3007.3 692.96
- Base_GPA           1   17.78 2948.2 693.16
<none>              2930.5 693.40
- Age                1   76.83 3007.3 698.96
- AvgDelta_AttPass_To 1  133.01 3063.5 704.36
- AvgProgpa_To       1   511.99 3442.4 738.42

Step: AIC=692.48
Num_Semem_To ~ Base_GPA + AvgDelta_AttPass_To + AvgProgpa_To +
  Age + Race

- Race               4   75.67 3017.0 691.89
<none>              2941.3 692.48
- Base_GPA           1   31.30 2972.6 693.57
- Age                1   74.69 3016.0 697.80
- AvgDelta_AttPass_To 1  122.23 3063.5 702.37
- AvgProgpa_To       1   518.31 3459.6 737.87

Step: AIC=691.89
Num_Semem_To ~ Base_GPA + AvgDelta_AttPass_To + AvgProgpa_To +
  Age

<none>              3017.0 691.89
- Base_GPA           1   42.65 3059.6 693.99
- Age                1   72.55 3089.5 696.83
- AvgDelta_AttPass_To 1  119.98 3136.9 701.28
- AvgProgpa_To       1   514.58 3531.6 735.88

call:
lm(formula = Num_Semem_To ~ Base_GPA + AvgDelta_AttPass_To +
  AvgProgpa_To + Age, data = Comp)

Coefficients:
(Intercept)      Base_GPA  AvgDelta_AttPass_To    AvgProgpa_To      Age
  0.04773      -0.51059      0.38004      1.29238      0.08463

> |
```

Directions:

1. To get the goal screen simply enter the following command and click Enter

```
> step(fm)
```

****Note:** Make sure your variable fm has been defined earlier.

Linear Model Using Stepwise Metrics

Part I

Goal:

```
> fm2<- lm(Num_Semem_To ~ Base_GPA + AvgDelta_AttPass_To + AvgProGpa_To + Age + Race, data = Comp)
> summary(fm2)

Call:
lm(formula = Num_Semem_To ~ Base_GPA + AvgDelta_AttPass_To +
    AvgProGpa_To + Age + Race, data = Comp)

Residuals:
    Min       1Q   Median       3Q      Max
-6.2383 -2.2117 -0.3987  1.5034 16.8750

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.12959    1.26355   0.103  0.918385
Base_GPA      -0.44122    0.25426  -1.735  0.083774 .
AvgDelta_AttPass_To  0.38520    0.11233   3.429  0.000695 ***
AvgProGpa_To   1.30550    0.18487   7.062  1.28e-11 ***
Age            0.08610    0.03212   2.681  0.007777 **
RaceBlack      0.45704    0.77402   0.590  0.555346
RaceHispanic  -0.74052    0.75492  -0.981  0.327470
RaceIndian     4.57634    3.36178   1.361  0.174506
Racewhite     -0.56579    0.67563  -0.837  0.403058
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.224 on 283 degrees of freedom
Multiple R-squared:  0.2248, Adjusted R-squared:  0.2029
F-statistic: 10.26 on 8 and 283 DF,  p-value: 1.354e-12
```

Directions:

1. Enter the following command in the command line and press Enter:

```
>
> fm2<-lm(Num_Semem_To~Base_GPA+AvgDelta_AttPass_To+AvgProGpa_To+Age+Race, data=Comp)
<
```

2. To pull up the summary of the model, enter the following and press Enter:

```
>
> summary(fm2)
```

3. This should be your final result:

```

>
> fm2<-lm(Num_Semem_To~Base_GPA+AvgDelta_AttPass_To+AvgProGpa_To+Age+Race, data=Comp)
> summary(fm2)

Call:
lm(formula = Num_Semem_To ~ Base_GPA + AvgDelta_AttPass_To +
    AvgProGpa_To + Age + Race, data = Comp)

Residuals:
    Min       1Q   Median       3Q      Max
-6.2384 -2.2117 -0.3987  1.5033 16.8750

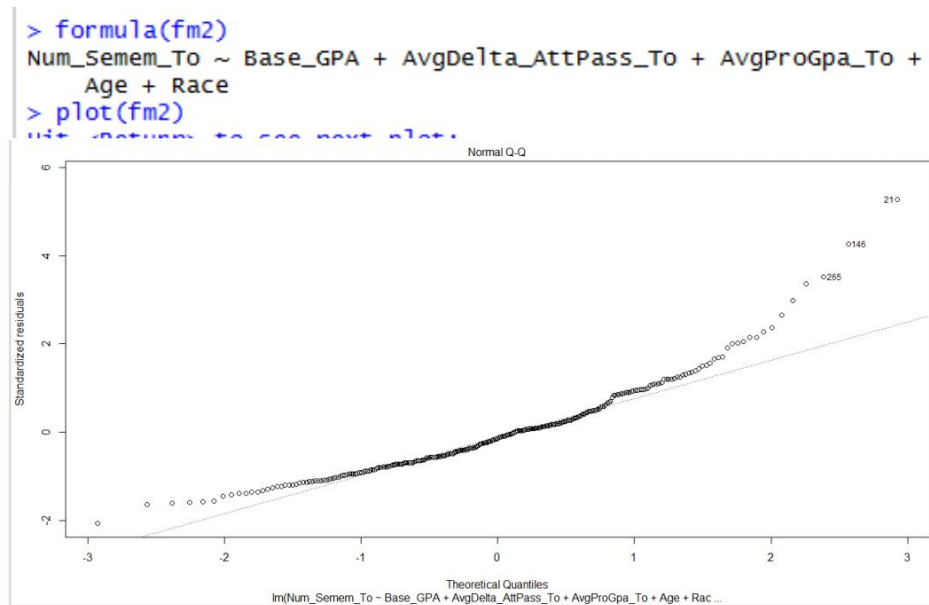
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.12958    1.26354   0.103  0.918388
Base_GPA     -0.44122    0.25426  -1.735  0.083769 .
AvgDelta_AttPass_To  0.38520    0.11233   3.429  0.000695 ***
AvgProGpa_To  1.30551    0.18487   7.062 1.27e-11 ***
Age          0.08610    0.03212   2.681  0.007777 **
RaceBlack    0.45704    0.77402   0.590  0.555341
RaceHispanic -0.74051    0.75492  -0.981  0.327473
RaceIndian   4.57634    3.36177   1.361  0.174505
RaceWhite   -0.56579    0.67562  -0.837  0.403060
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

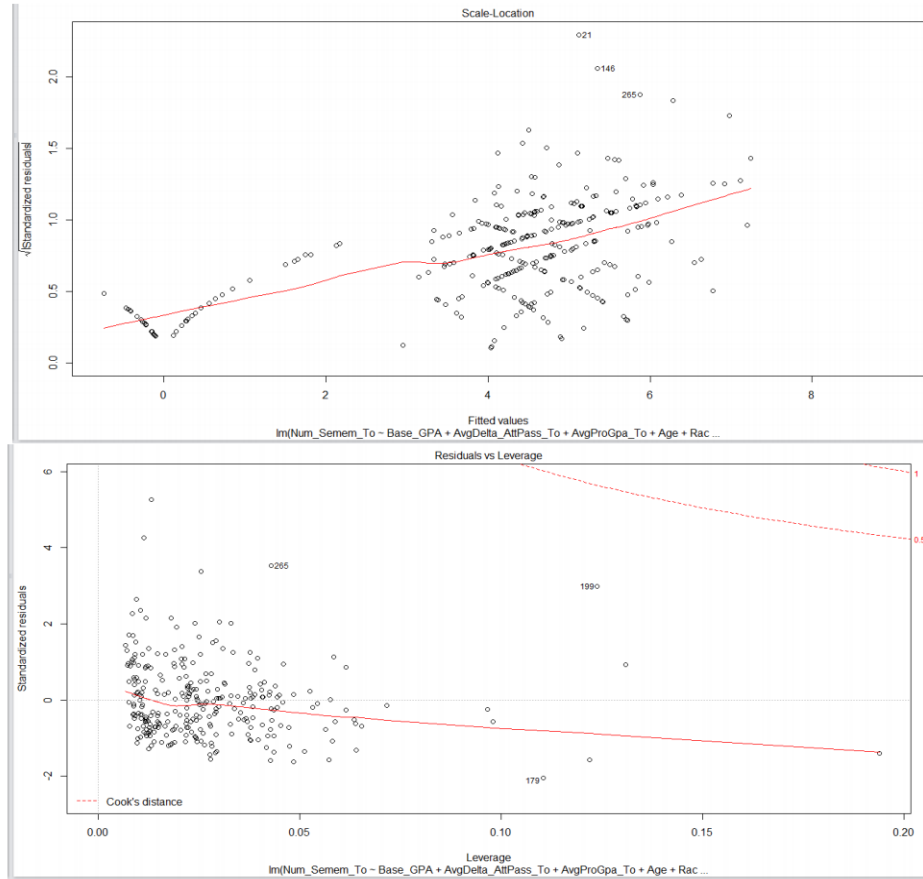
Residual standard error: 3.224 on 283 degrees of freedom
Multiple R-squared:  0.2248, Adjusted R-squared:  0.2029
F-statistic: 10.26 on 8 and 283 DF, p-value: 1.354e-12
>

```

Part II: Plots

Goal:





Directions:

1. Enter the following command and click Enter:

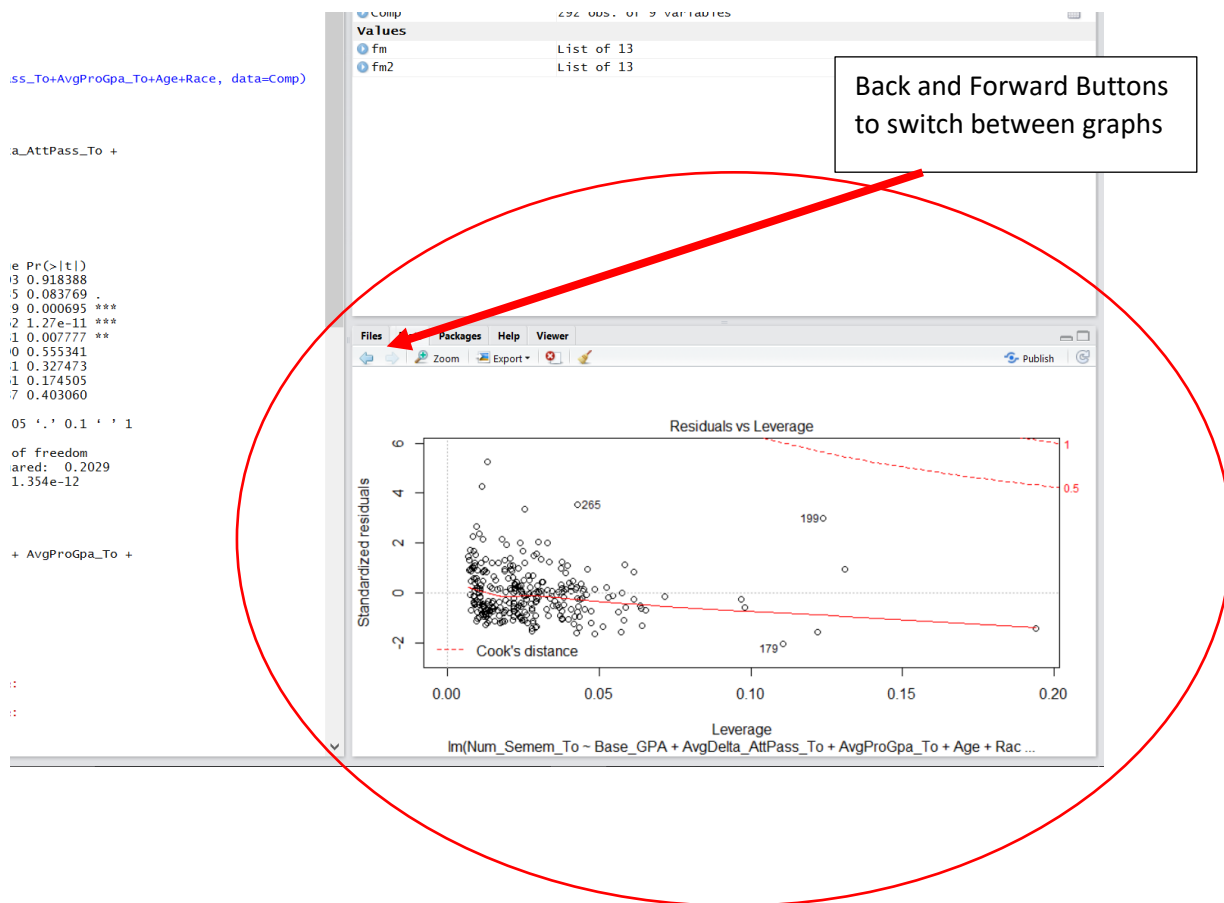
```
> formula(fm2)
```

This will display the linear model we are using.

2. Enter the following command and click Enter in order to get the plots. You can see them in the bottom right console.

```
> plot(fm2)
```

3. In order to get the next graph, click Enter and repeat this action again.



You are now done with your R tutorial.