Office Hours Week 11

Kevin Martin

11/8/2020

Agenda

- Stargazer
- Omitted Variable Bias
- Some EDA charts that I like

Stargazer

This cheat sheet from Jake Russ is amazing y'all https://www.jakeruss.com/cheatsheets/stargazer/

Bonus Fact stargazer is a tounge in cheek aknowledgement that most of the time you're "looing for stars (significance)" at the end of your statistical modeling work

Get some data to work with

We'll use the `mtcars` dataset that comes bundled with R. You can see more info about what the variables mean and where the data set comes from by typing `?mtcars` into the console. You can also equivalently type `mtcars` into the search box of the help window.

Side note there are a lot of interesting data sets that just come bundled in R. If you type the `data()` command, you can see a list of all of the data sets that are available and a brief summary of all of them.

```
## `mtcars` is a built in dataset in R. Gets used in all kinds of examples

# bonus info
# ?mtcars
# View(mtcars)
# data()
  # highlight then `Ctrl` + `Shift` + `C` to block comment and uncomment

summary(mtcars)
```

```
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0  
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
```

```
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000
##        am              gear            carb
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
##  Median :0.0000   Median :4.000   Median :2.000
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

Build some models

```
# just look at mpg as predicted by horsepower, add weight and 1/4 mile time as covariates
mod1 <- lm(mpg ~ hp , data=mtcars)
mod2 <- lm(mpg ~ hp + wt , data=mtcars)
mod3 <- lm(mpg ~ hp + wt + qsec , data=mtcars)
```

What do we expect: corr(mpg, hp): negative corr(mpg, wt): negative corr(mpg, qsec): positive

Gaze Some Stars

Formatting stargazer layout for a variety of formats

```
stargazer(mod1, mod2, mod3,
          type="text",
          se = list( sqrt(diag(vcovHC(mod1))), sqrt(diag(vcovHC(mod2))), sqrt(diag(vcovHC(mod3)))) ,
          column.labels = c("hp","wt+hp","overfit"))
```

Text layout

```
##
## ==========================================================================================
##                                             Dependent variable:
##                      --------------------------------------------------------------------
##                                                    mpg
##                                hp                 wt+hp                  overfit
##                               (1)                  (2)                    (3)
## ------------------------------------------------------------------------------------------
## hp                          -0.068***            -0.032***               -0.018
##                              (0.017)              (0.009)                (0.014)
##
## wt                                               -3.878***              -4.359***
##                                                   (0.769)                (0.950)
##
```

```
## qsec                                                             0.511
##                                                                 (0.434)
##
## Constant                30.099***           37.227***           27.611***
##                          (2.410)             (2.230)             (7.547)
##
## -----------------------------------------------------------------------------------------
## Observations               32                  32                  32
## R2                        0.602               0.827               0.835
## Adjusted R2               0.589               0.815               0.817
## Residual Std. Error    3.863 (df = 30)     2.593 (df = 29)     2.578 (df = 28)
## F Statistic         45.460*** (df = 1; 30) 69.211*** (df = 2; 29) 47.153*** (df = 3; 28)
## =========================================================================================
## Note:                                                 *p<0.1; **p<0.05; ***p<0.01
```

Important note on parsimony when we added `qsec` in, the standard errors on hp AND wt both increased. This is what happens when you have multicolinearity. This is why we tend to like parsimonious models with few extra covariates.

Side note: It's a little bit of a pain in the butt to pull out the coefficients from the Heteroskedastic Consistent vcov matrix you can see below for the standard errors.

The standard error is 0.0166 and that is a nice standard

Fancier output formats (latex and html) You **CAN** output to **latex** or **html**. Warning up top, they don't render in Rstudio. They DO render in the knitted output depending on the specific format you're looking at.

NOTICE the `results = 'asis'` up top in the code block header. (all glory to this answer on stackoverflow https://stackoverflow.com/a/30423627/1992108)

```
## the results='asis' is important here
## the latex will render in the knitted pdf.
stargazer(mod1,mod2, mod3,
          type="latex",
          se = list( sqrt(diag(vcovHC(mod1))),sqrt(diag(vcovHC(mod2))) ,sqrt(diag(vcovHC(mod3)))) ,
          column.labels = c("hp","wt+hp","overfit"))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fri, Apr 02, 2021 - 04:06:50 PM

```
## the html will render if knitted to markdown (or html of course)
  # this will look BAAAD in the PDF.
stargazer(mod1,mod2, mod3,
          type="html",
          se = list( sqrt(diag(vcovHC(mod1))),sqrt(diag(vcovHC(mod2))) ,sqrt(diag(vcovHC(mod3)))) ,
          column.labels = c("hp","wt+hp","overfit"))
```

Dependent variable:

mpg

hp

wt+hp

Table 1:

	Dependent variable:		
	mpg		
	hp	wt+hp	overfit
	(1)	(2)	(3)
hp	−0.068***	−0.032***	−0.018
	(0.017)	(0.009)	(0.014)
wt		−3.878***	−4.359***
		(0.769)	(0.950)
qsec			0.511
			(0.434)
Constant	30.099***	37.227***	27.611***
	(2.410)	(2.230)	(7.547)
Observations	32	32	32
R^2	0.602	0.827	0.835
Adjusted R^2	0.589	0.815	0.817
Residual Std. Error	3.863 (df = 30)	2.593 (df = 29)	2.578 (df = 28)
F Statistic	45.460*** (df = 1; 30)	69.211*** (df = 2; 29)	47.153*** (df = 3; 28)
Note:			*p<0.1; **p<0.05; ***p<0.01

overfit

(1)

(2)

(3)

hp

-0.068***

-0.032***

-0.018

(0.017)

(0.009)

(0.014)

wt

-3.878***

-4.359***

(0.769)

(0.950)

qsec

0.511

(0.434)

Constant

30.099***

37.227***

27.611***

(2.410)

(2.230)

(7.547)

Observations

32

32

32

R2

0.602

0.827

0.835

Adjusted R2

0.589

0.815

0.817

Residual Std. Error

3.863 (df = 30)

2.593 (df = 29)

2.578 (df = 28)

F Statistic

45.460*** (df = 1; 30)

69.211*** (df = 2; 29)

47.153*** (df = 3; 28)

Note:

p<0.1; ***p<0.05;*** p<0.01

Direction of Omitted Variable Bias

We can't tell the exact size of the omitted variable bias, but we can tell the direction if we know the **direction of the relationship between the omitted variable and the included input variable (I have labeled it α_1)** as well as the **direction of the relationship between the omitted variable and the output variable (I have labeled it α_2)**

This website has a farily nice table (link). See the header "Predicting the Direction of Omitted Variable Bias"

- $\alpha_1 = sign(cor(x_{omit}, x_{iclude}))$
- $\alpha_2 = sign(cor(x_{omit}, y))$
  - This is **technically incorrect** and will not hold true all the time.
  - You should use $sign(\beta_2)$ here where β_2 is the coefficient associated with your omitted variable if it were included.
  - Often times, $\alpha_2 = sign(\beta_2)$ but not all the time. There are a few examples at the end of the document where the assumption that $\alpha_2 = sign(\beta_2)$ doesn't hold
- $\alpha_{dir} = \alpha_1 * \alpha_2$
  - As noted above, this should technically be $\alpha_{dir} = \alpha_1 * sign(\beta_2)$

We get the direction of the bias by multiplying α_1 by α_2.

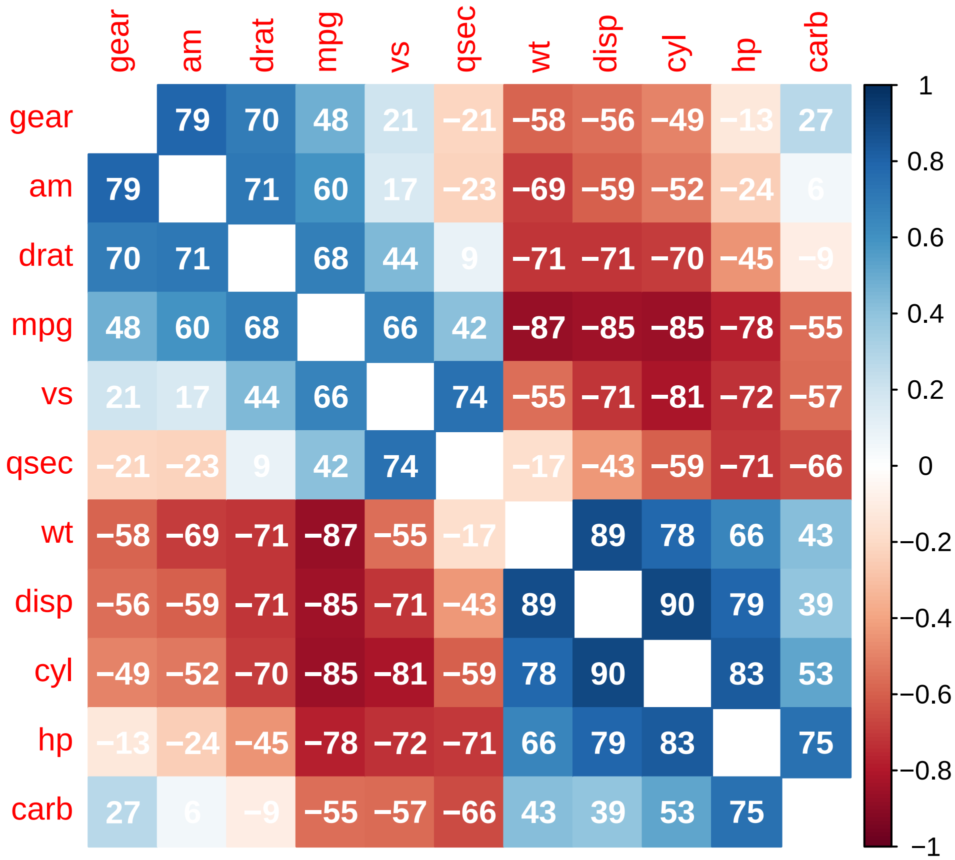
- **If α_{dir} is positive**
  - The **coefficient** associated with the included variable in the shortened equation is **larger** than it would be if the omitted variable were included.
    - *Larger means more positive in this case. It does NOT mean greater magnitude*
  - Adding in the omitted variable will push the coefficient on the included variable in the negative direction.
- **If α_{dir} is negative**
  - The **coefficient** associated with the included variable in the shortened equation is **smaller** than it would be if the omitted variable were included.
    - *Smaller means more negative in this case. It does NOT mean lesser magnitude*
  - Adding in the omitted variable will push the coefficient on the included variable in the positive direction.

More concrete models to view omitted variable biase

We're going to use the `mtcars` data that we introduced in the stargazer discussion to build some models and view the effects of omittiing (then including) variables.

We see a correlation plot of the variables in the `mtcars` data below. Blue circles indicate positive correlation, red circles indicate negative correlation.

```
## corrplot takes a correlation matrix as an arument
  # needs the corrplot package
corrplot::corrplot(cor(mtcars),method = "color",order="AOE",
                   diag=FALSE, addCoef.col = "white", addCoefasPercent = TRUE)
```



Estimator Negatively Biased Away from Zero In the case below, we have an estimator that is biased in the negative direction. Since the coefficient that it is associated with is negative as well we would say it is biased away from zero. We break down the components of the omitted variable bias below.

- α_1: `wt` **positively** correlated with `hp`
- α_2: `wt` **negatively** correlated with `mpg`
- $\alpha_{dir}(estimate)$: **negative** overall

We see that including the ommited variable reduces the negative bias of the coefficient on `hp`. It becomes less negative and moves towards zero.

```
sprintf("a1 is %d", (a1     <- sign(cor(mtcars$wt,mtcars$hp))))
```

```
## [1] "a1 is 1"
```

```
sprintf("a2 is %d", (a2     <- sign(cor(mtcars$wt,mtcars$mpg))))
```

```
## [1] "a2 is -1"
```

```
sprintf("adir(estimate) is %d",(adir <- a1*a2))
```

```
## [1] "adir(estimate) is -1"
```

```r
stargazer(mod1, mod2, type = "text")
```

```
##
## ===================================================================
##                                   Dependent variable:
##                         -------------------------------------------
##                                           mpg
##                                  (1)                   (2)
## -------------------------------------------------------------------
## hp                            -0.068***             -0.032***
##                                (0.010)               (0.009)
##
## wt                                                  -3.878***
##                                                      (0.633)
##
## Constant                      30.099***             37.227***
##                                (1.634)               (1.599)
##
## -------------------------------------------------------------------
## Observations                     32                    32
## R2                              0.602                 0.827
## Adjusted R2                     0.589                 0.815
## Residual Std. Error       3.863 (df = 30)       2.593 (df = 29)
## F Statistic           45.460*** (df = 1; 30) 69.211*** (df = 2; 29)
## ===================================================================
## Note:                                  *p<0.1; **p<0.05; ***p<0.01
```

Estimator Positively Biased Away from Zero In the case below, we have an estimator that is biased in the positive direction. Since the coefficient that it is associated with is positive as well we would say it is biased away from zero. We break down the components of the omitted variable bias below.

- α_1: `disp` **positively** correlated with `invq`
- α_2: `disp` **positively** correlated with `hp`
- $\alpha_{dir}(estimate)$: **positive** overall

We see that including the ommited variable reduces the positive bias of the coefficient on `invq`. It becomes less positive and moves towards zero.

```r
## this variable is created out of the ether to get all positively correlated variables.
invq = 1/mtcars$qsec

sprintf("a1 is %d", (a1    <- sign(cor(mtcars$disp,invq))))
```

```
## [1] "a1 is 1"
```

```r
sprintf("a2 is %d", (a2    <- sign(cor(mtcars$disp,mtcars$hp))))
```

```
## [1] "a2 is 1"
```

```r
sprintf("adir(estimate) is %d",(adir <- a1*a2))
```

```
## [1] "adir(estimate) is 1"
```

```r
mod1 <- lm(data=mtcars, hp ~ invq)
mod2 <- lm(data=mtcars, hp ~ invq + disp)
stargazer(mod1, mod2, type = "text")
```

```
##
## =================================================================
##                                  Dependent variable:
##                        ------------------------------------------
##                                           hp
##                                (1)                  (2)
## -----------------------------------------------------------------
## invq                      9,058.347***         6,050.442***
##                           (1,485.284)          (1,038.113)
##
## disp                                              0.320***
##                                                   (0.047)
##
## Constant                  -365.715***          -269.449***
##                             (84.420)             (55.212)
##
## -----------------------------------------------------------------
## Observations                  32                   32
## R2                          0.554                0.828
## Adjusted R2                 0.539                0.816
## Residual Std. Error    46.570 (df = 30)     29.436 (df = 29)
## F Statistic         37.194*** (df = 1; 30) 69.593*** (df = 2; 29)
## =================================================================
## Note:                                 *p<0.1; **p<0.05; ***p<0.01
```

Estimator Positively Biased Towards Zero (SOME ASSUMPTIONS BREAK) In the case below, we have an estimator that is biased in the Positive direction. Since the coefficient that it is associated with is negative we would say it is biased towards zero. We break down the components of the omitted variable bias below.

We see that including the ommited variable reduces the positive bias of the coefficient on `cyl`. It becomes more negative and moves away from zero.

Directional Estimator Breakdown Our estimated bias direction based on correlations is:

- α_1: `vs` **negatively** correlated with `cyl`
- α_2: `vs` **positively** correlated with `mpg`
- $\alpha_{dir}(estimate)$: **negative** overall

Our correct bias direction which is based on β_2 is:

- α_1: `vs` **negatively** correlated with `cyl`
- β_2: is a **negative** coefficient in the full regression
- $\alpha_{dir}(correct)$: **positive** overall

What's the Lesson? Obviously, this breakdown of our directional bias estimator is distressing. You would like to think that you can at least predict the direction of your bias on omitted variables. Obviously in the real world you probably can't actually see if $sign(\beta_2)$ is different than α_2 because that would require finding β_2, which would require running the regression including the omitted variable. If you could do that, then you wouldn't need this whole exercise on estimating the bias associated with omitting the variable in the first place.

This should just drive home the fact that causality is hard. Even basic things like predicting direction on omitted variable bias are frought with counter-intuitive examples. You really need an experiment to determine causality. If that interests you, might I suggest the w241 course.

```
sprintf("a1 is %d", (a1    <- sign(cor(mtcars$vs,mtcars$cyl))))
```

```
## [1] "a1 is -1"
```

```
sprintf("a2 is %d", (a2    <- sign(cor(mtcars$vs,mtcars$mpg))))
```

```
## [1] "a2 is 1"
```

```
sprintf("adir(estimate) is %d",(adir <- a1*a2))
```

```
## [1] "adir(estimate) is -1"
```

```
mod1 <- lm(data=mtcars, mpg ~ cyl)
mod2 <- lm(data=mtcars, mpg ~ cyl + vs)
mod3 <- lm(data=mtcars, mpg ~ vs)
stargazer(mod1, mod2, mod3, type = "text")
```

```
##
## ==========================================================================================
##                                                Dependent variable:
##                     ----------------------------------------------------------------------
##                                                        mpg
##                               (1)                     (2)                     (3)
## ------------------------------------------------------------------------------------------
## cyl                        -2.876***               -3.091***
##                             (0.322)                 (0.558)
##
## vs                                                  -0.939                  7.940***
##                                                     (1.978)                 (1.632)
##
## Constant                   37.885***               39.625***               16.617***
##                             (2.074)                 (4.225)                 (1.080)
##
## ------------------------------------------------------------------------------------------
## Observations                  32                      32                      32
## R2                          0.726                   0.728                   0.441
## Adjusted R2                 0.717                   0.710                   0.422
## Residual Std. Error     3.206 (df = 30)         3.248 (df = 29)         4.581 (df = 30)
## F Statistic         79.561*** (df = 1; 30) 38.866*** (df = 2; 29) 23.662*** (df = 1; 30)
## ==========================================================================================
## Note:                                                        *p<0.1; **p<0.05; ***p<0.01
```

EDA charts that I like

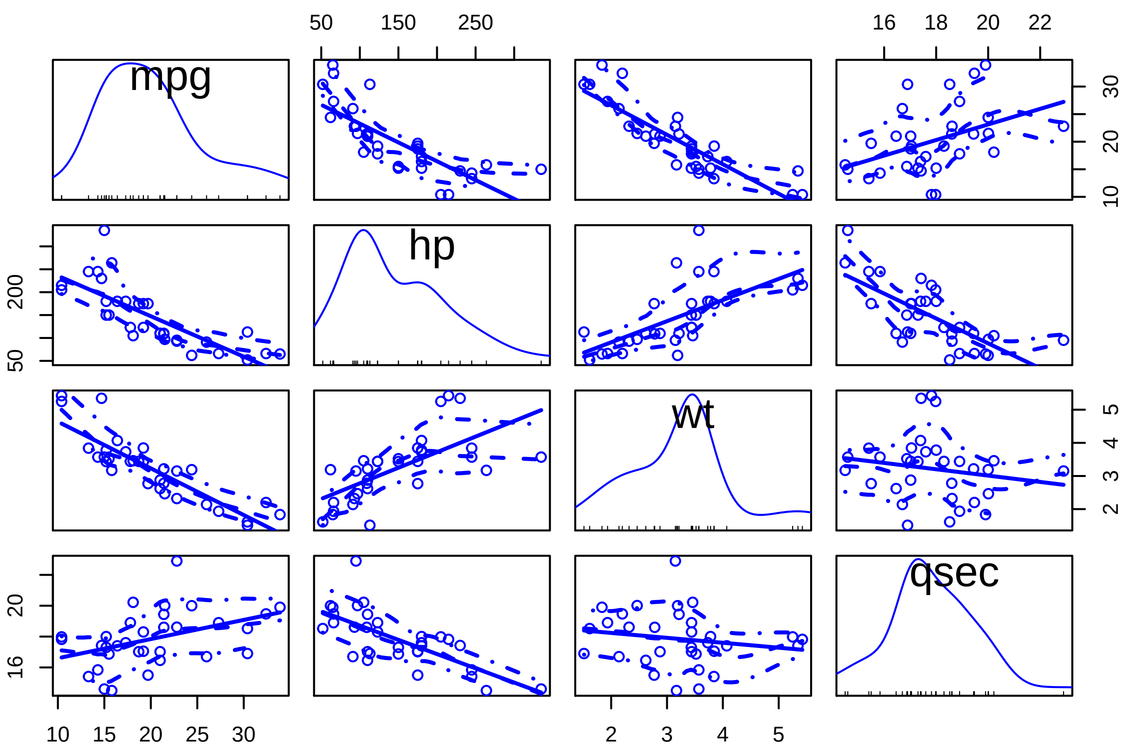
Corrplot

I like `corrplot()` I used it above. Very quick and easy way to check correlations. I'm not sure if the default color template has accessibility issues for the color blind. It looks like it might, but you should be able to add numbers or use `method="ellipse"` to account for that.

Scatter Plot Matrix.

I really like `scatterplotMatrix()` you can think of it as a more advanced/dense version of the corrplot. The price you pay for the increased information density is that your plots end up being busier and harder to interpret at a glance than the comparable `corrplot()`. I personally wouldn't do it with more than about 4 variables.

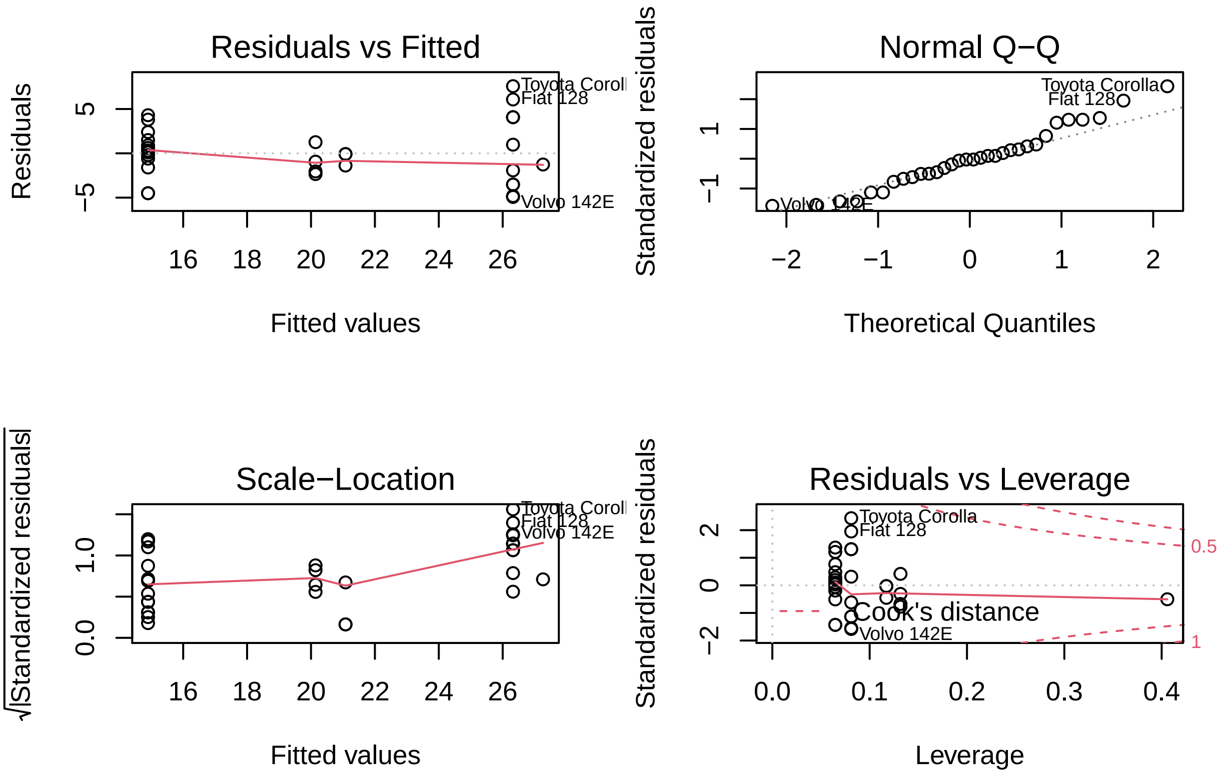
```
## Plot a scatterplot matrix
  # requires the car package
car::scatterplotMatrix(mtcars[,c("mpg","hp","wt","qsec")])
```



Plot a model

There are some really good diagnostic charts that come up if you just put your model inside a plot command. We will cover the interpretation of these charts later in the cours. You can use `par` to make a grid of charts for your diagnostics to map onto. That can make your reports cleaner.

```
par(mfrow = c(2,2), oma = c(0,0,0,0)) #oma = outside margins plot(basemodel)
plot(mod2)
```


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Volvo 142E
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Leverage
Cook's distance
0.5
1