Regression in R

Jan-Philipp Kolb

16 Mai, 2019

Why a part on simple regression

- some machine learning concepts are based on regression
- I would like to remind you how simple regression in R works.
- I also want to show the constraints
- In a next step we will learn, how to coop with that

Variables of the `mtcars` dataset

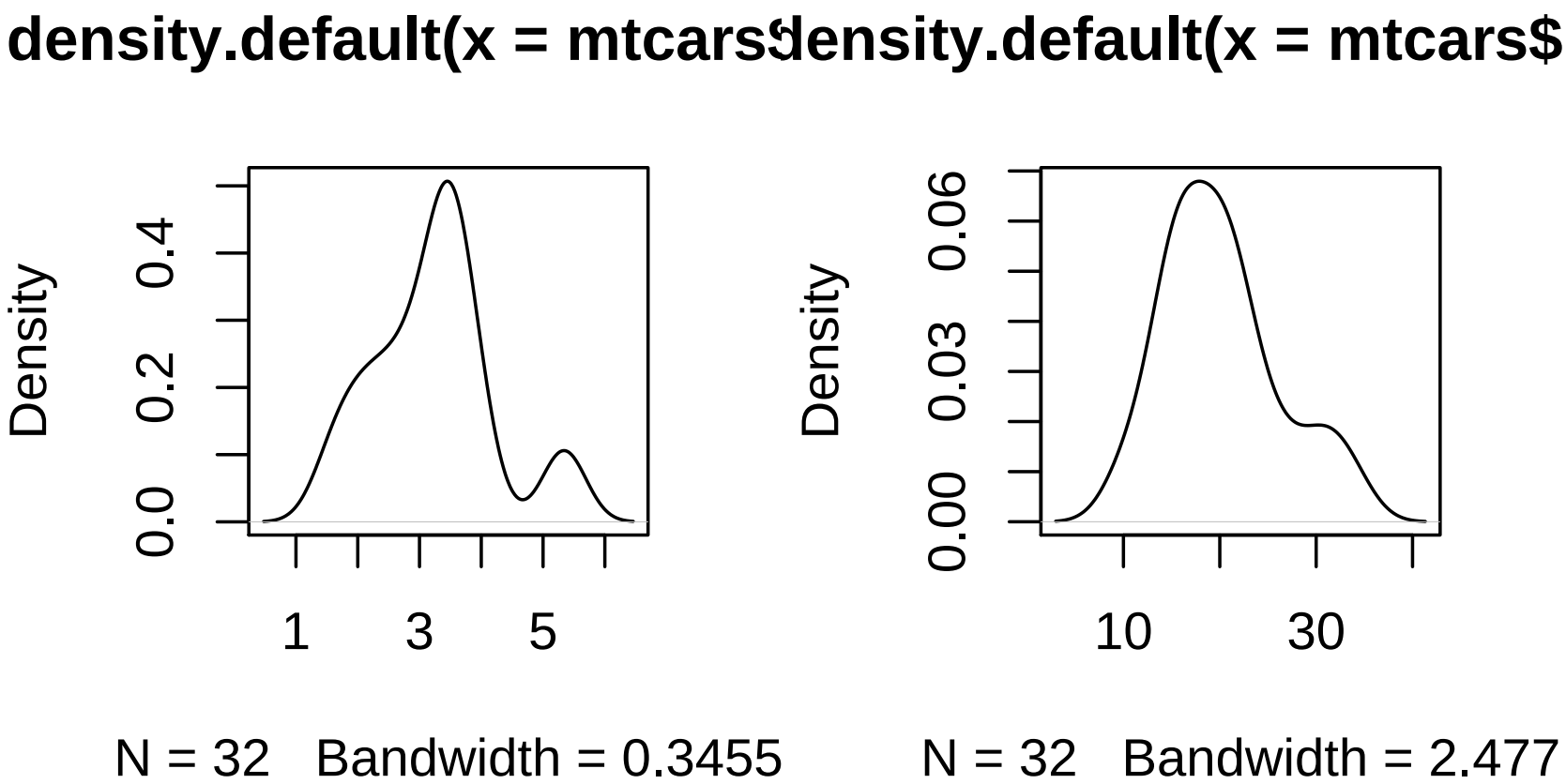
Help for the `mtcars` dataset:

- mpg - Miles/(US) gallon
- cyl - Number of cylinders
- disp - Displacement (cu.in.)
- hp - Gross horsepower
- drat - Rear axle ratio
- wt - Weight (1000 lbs)
- qsec - 1/4 mile time
- vs - Engine (0 = V-shaped, 1 = straight)
- am - Transmission (0 = automatic, 1 = manual)
- gear - Number of forward gears
- carb - Number of carburetors

Dataset mtcars

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0			
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0			
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1			
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1			
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0			
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1			
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0			
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1			
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1			
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1			
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1			
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0			
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0			
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0			
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0			
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0			

Distributions of two variables of mtcars



A simple regression model

Dependent variable - miles per gallon (mpg)

Independent variable - weight (wt)

```
##
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
##
## Coefficients:
## (Intercept)           wt
##      37.285       -5.344
```

Get the model summary

```
##
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5432 -2.3647 -0.1252  1.4096  6.8727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
## wt           -5.3445     0.5591  -9.559 1.29e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 3.046 on 30 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7446
## F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

The model formula

Model without intercept

```
##    Estimate Std. Error  t value    Pr(>|t|)
## wt 5.291624  0.5931801 8.920771 4.55314e-10
```

Adding further variables

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 39.686261  1.7149840 23.140893 3.043182e-20
## wt          -3.190972  0.7569065 -4.215808 2.220200e-04
## cyl         -1.507795  0.4146883 -3.635972 1.064282e-03
```

The command as.formula

```
## [1] "formula"
```

The command `model.matrix`

Further possibilities to specify the formula

Interaction effect

Take the logarithm

A model with interaction effect

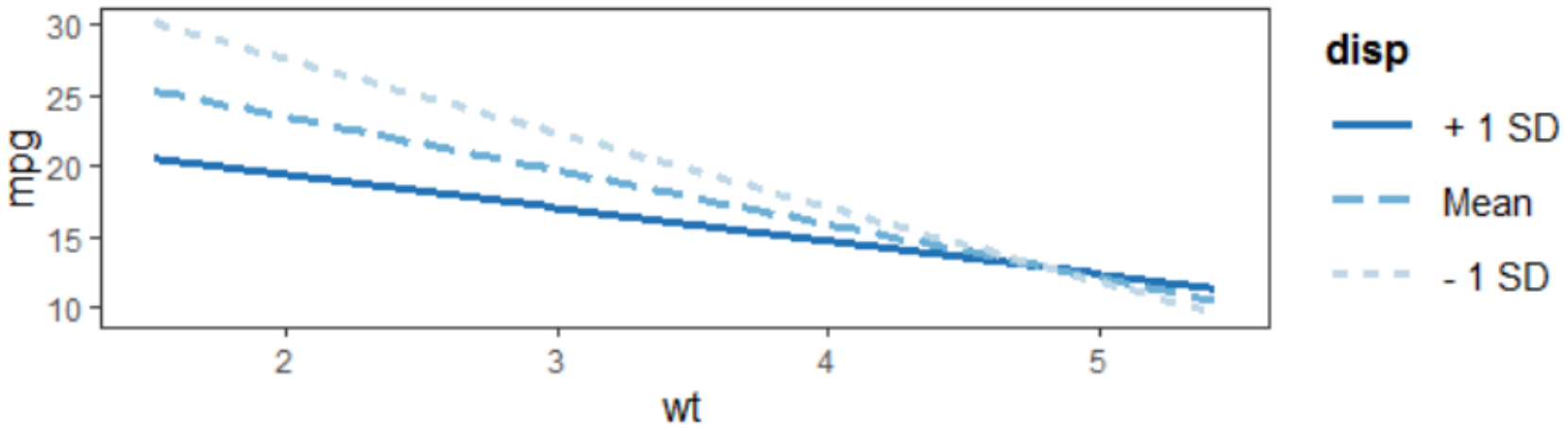
-disp - Displacement (cu.in.)

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 44.08199770 3.123062627 14.114990 2.955567e-14
## wt          -6.49567966 1.313382622 -4.945763 3.216705e-05
## disp        -0.05635816 0.013238696 -4.257078 2.101721e-04
## wt:disp      0.01170542 0.003255102  3.596022 1.226988e-03
```

Exploring interactions

```
library(jtools)
interact_plot(m3d, pred = "wt", modx = "disp")
```

- With a continuous moderator (in our case `disp`) you get three lines — 1 standard deviation above and below the mean and the mean itself.



Example: object orientation

- m3 is now a special regression object
- Various functions can be applied to this object

```
##          Mazda RX4     Mazda RX4 Wag         Datsun 710     Horn
##           22.27914          21.46545           26.25203
## Hornet Sportabout           Valiant
##           16.64696          19.59873
```

```
##          Mazda RX4     Mazda RX4 Wag         Datsun 710     Horn
##         -1.2791447        -0.4654468         -3.4520262
## Hornet Sportabout           Valiant
##          2.0530424        -1.4987281
```

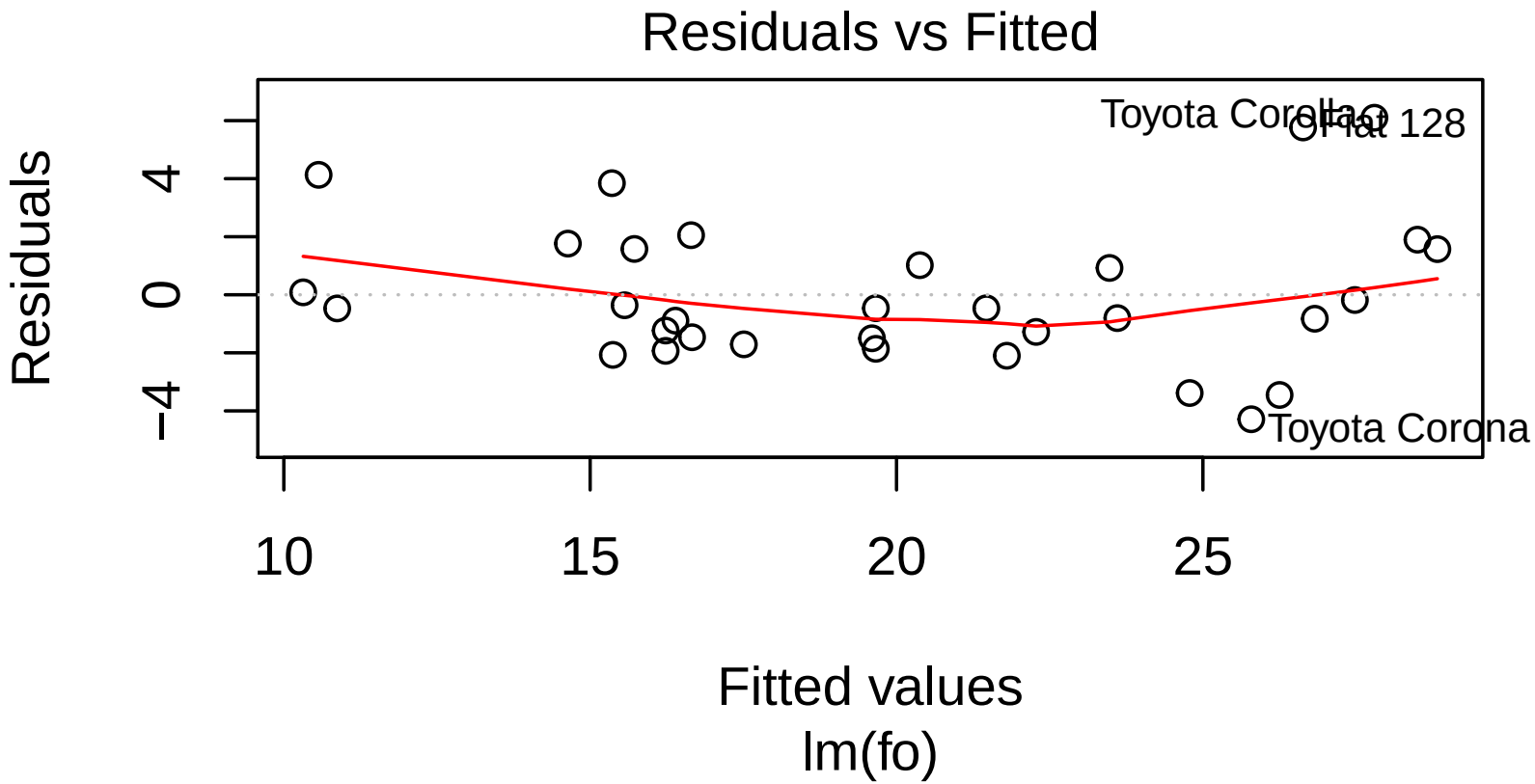
Make model prediction

```
## [1] 21.0 21.0 22.8 21.4 18.7 18.1
```

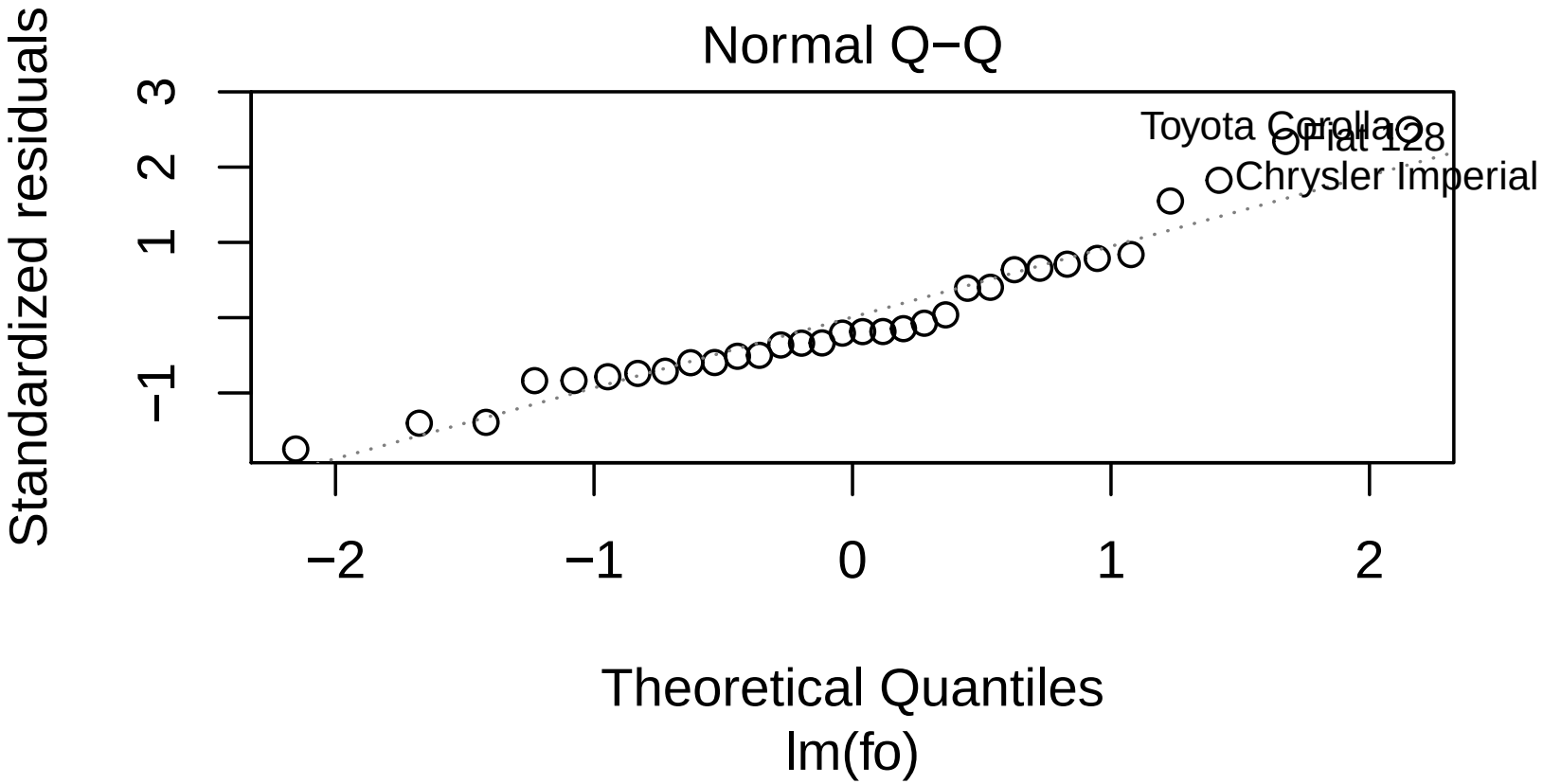
```
##          Mazda RX4      Mazda RX4 Wag          Datsun 710    Horn
##           23.28261           21.91977           24.88595
## Hornet Sportabout            Valiant
##           18.90014           18.79325
```

Residual plot - model assumptions violated?

- the case if a deviation pattern from line

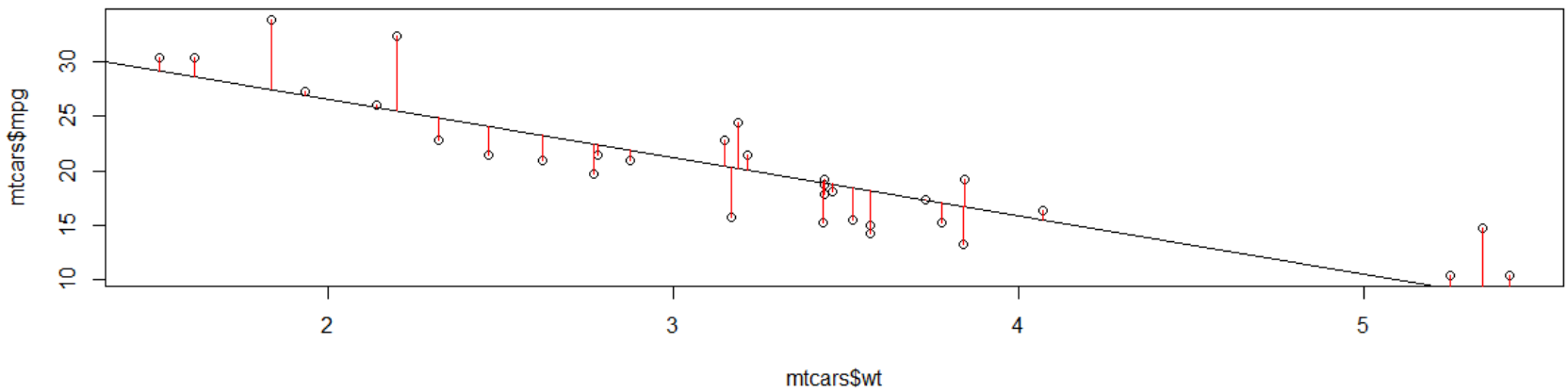


Residual plot

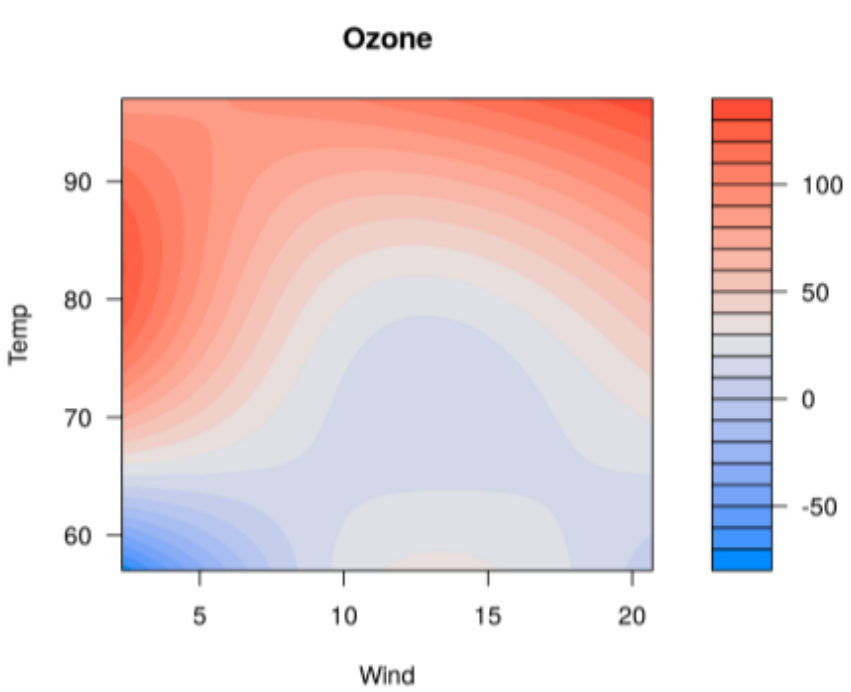


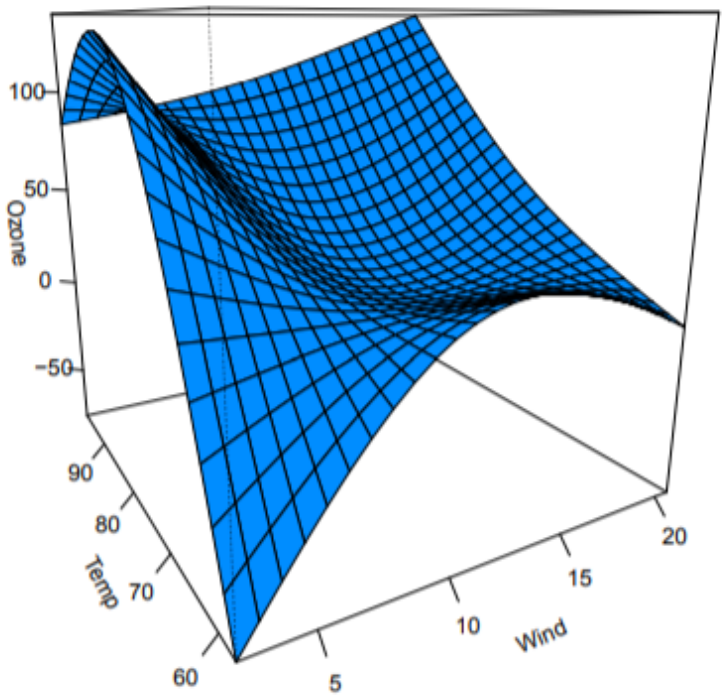
- If the residuals are normally distributed, they should be on the same line.

Regression diagnostic with base-R



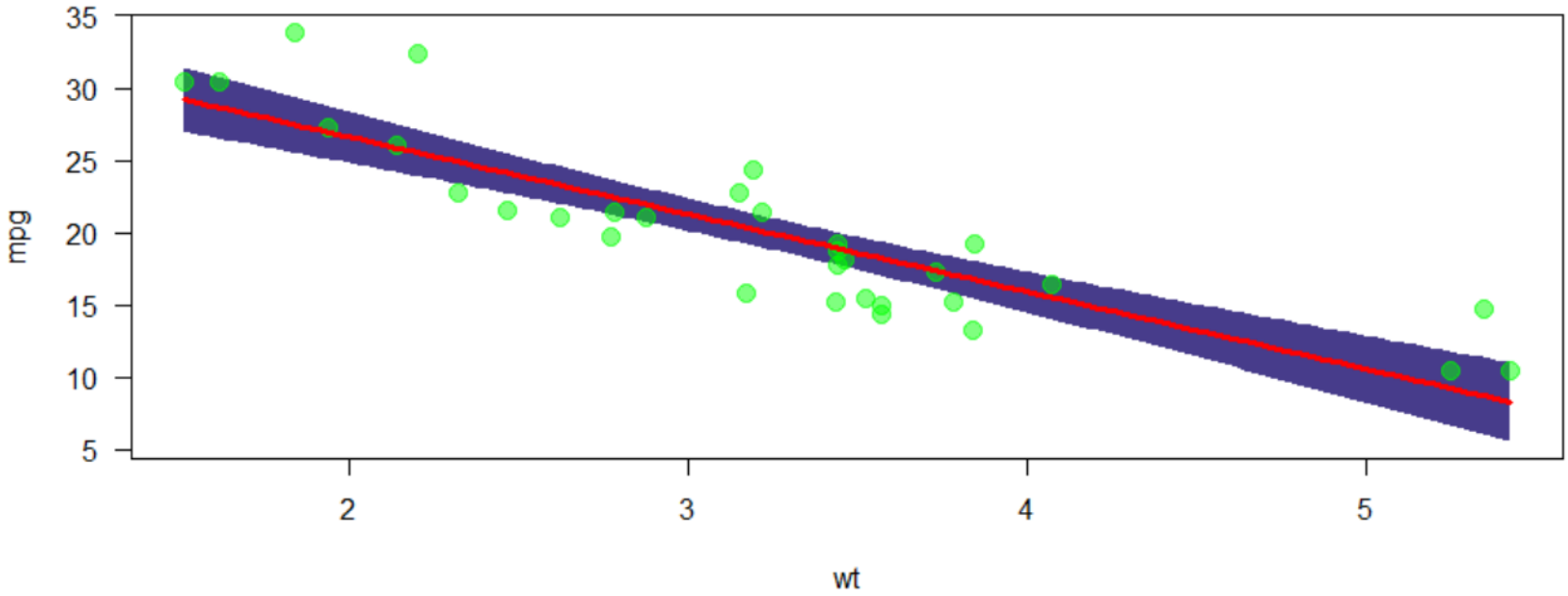
The visreg-package





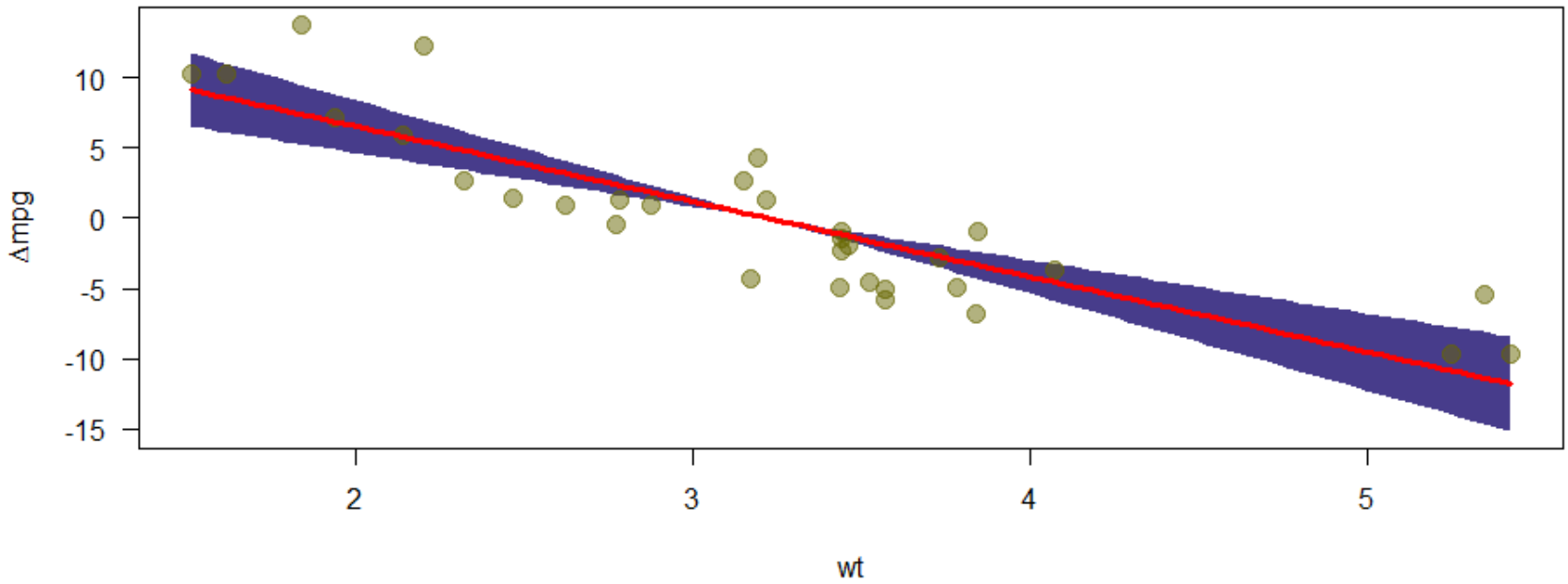
The visreg-package

- The default-argument for `type` is `conditional`.
- Scatterplot of `mpg` and `wt` plus regression line and confidence bands



Visualisation with visreg

- Second argument - Specification covariate for visualisation
- plot shows the effect on the expected value of the response by moving the x variable away from a reference point on the x-axis (for numeric variables, the mean).

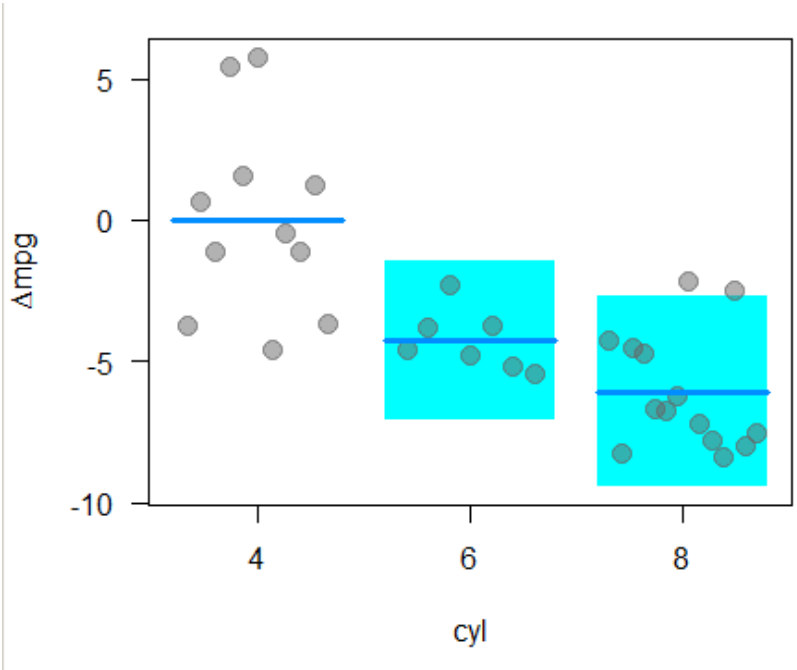


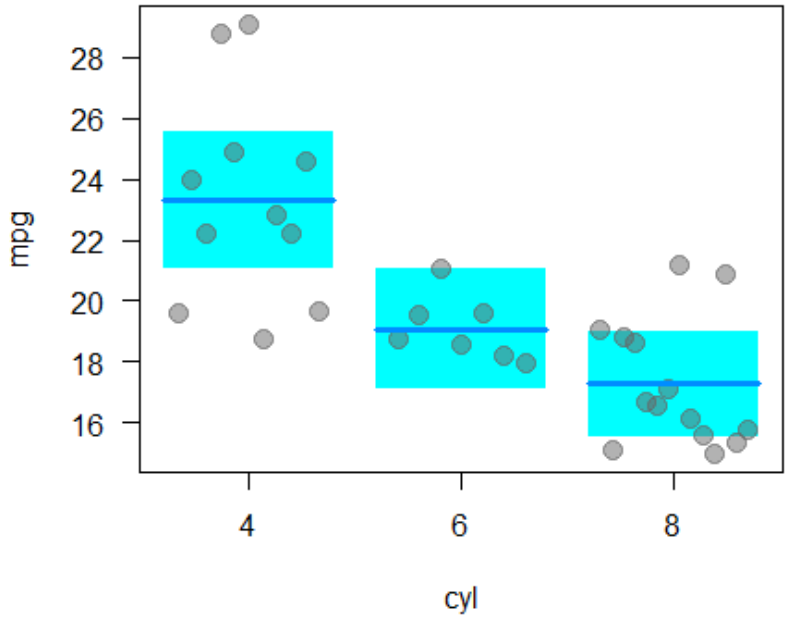
Regression with factors

- The effects of factors can also be visualized with `visreg`:

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 33.990794  1.8877934 18.005569 6.257246e-17
## cyl6        -4.255582  1.3860728 -3.070244 4.717834e-03
## cyl8        -6.070860  1.6522878 -3.674214 9.991893e-04
## wt          -3.205613  0.7538957 -4.252065 2.130435e-04
```

Effects of factors



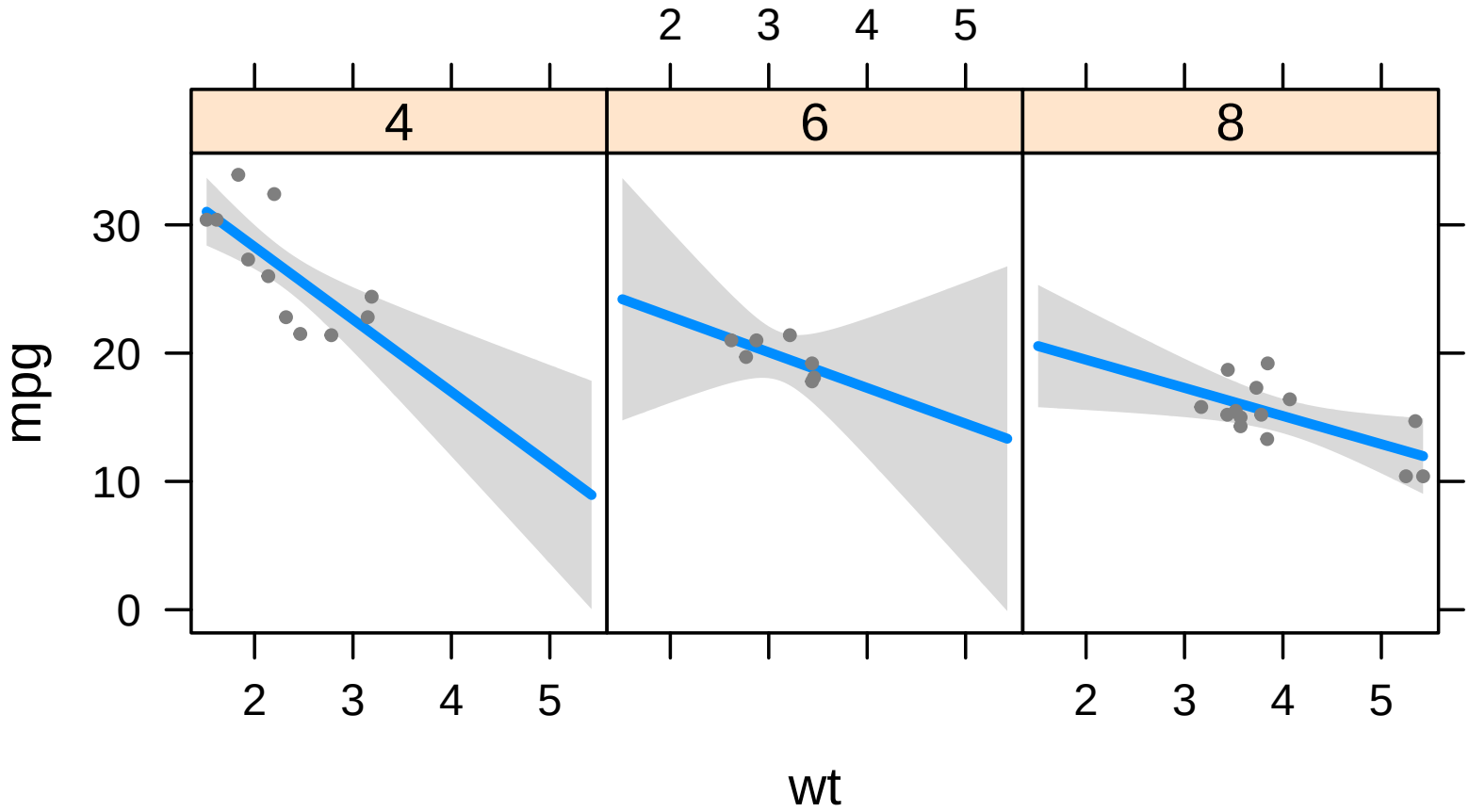


The command `model.matrix`

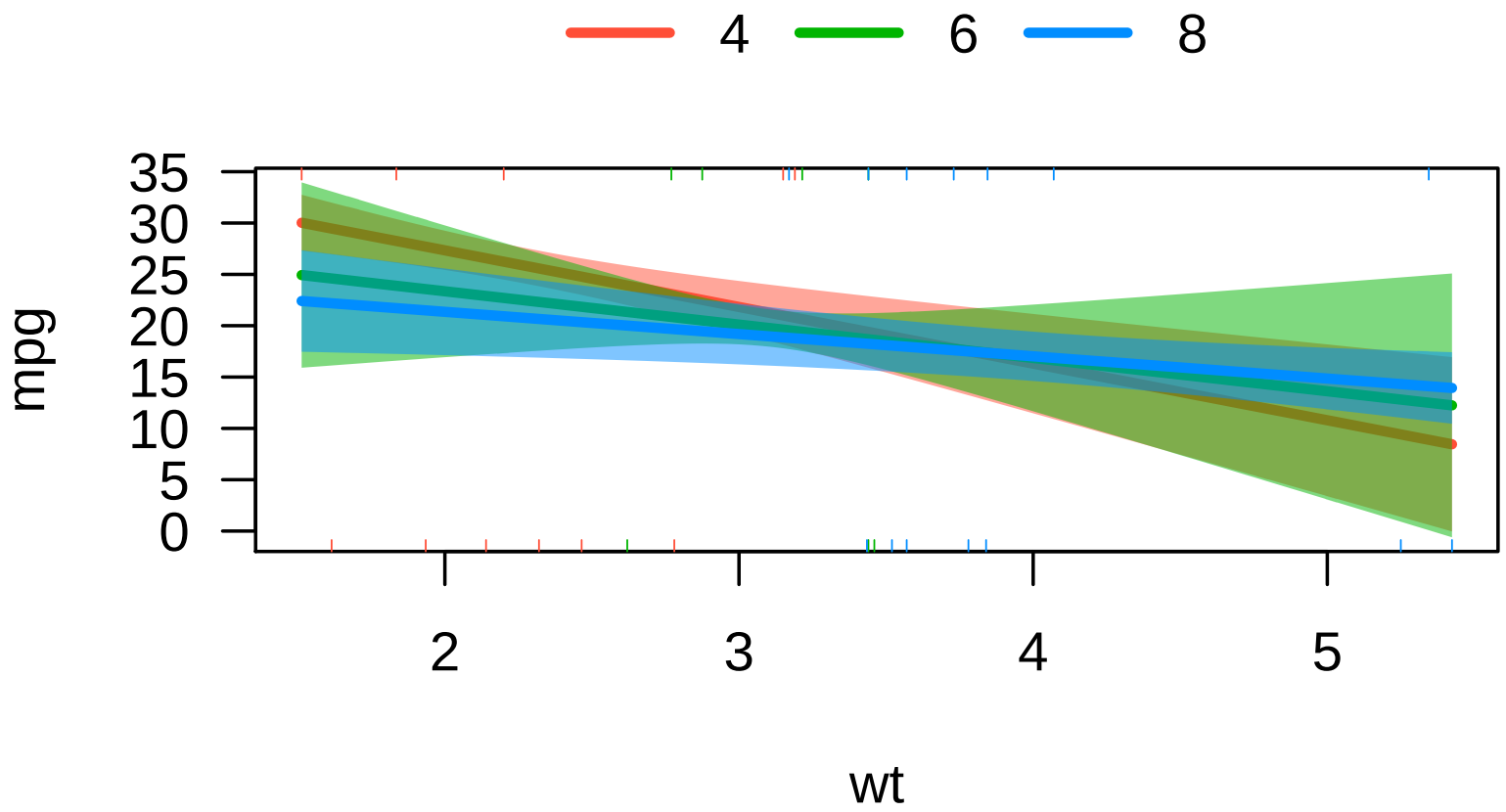
The package visreg - Interactions

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  39.571196   3.193940 12.3894599 2.058359e-12
## cyl6        -11.162351   9.355346 -1.1931522 2.435843e-01
## cyl8        -15.703167   4.839464 -3.2448150 3.223216e-03
## wt           -5.647025   1.359498 -4.1537586 3.127578e-04
## cyl6:wt       2.866919   3.117330  0.9196716 3.661987e-01
## cyl8:wt       3.454587   1.627261  2.1229458 4.344037e-02
```

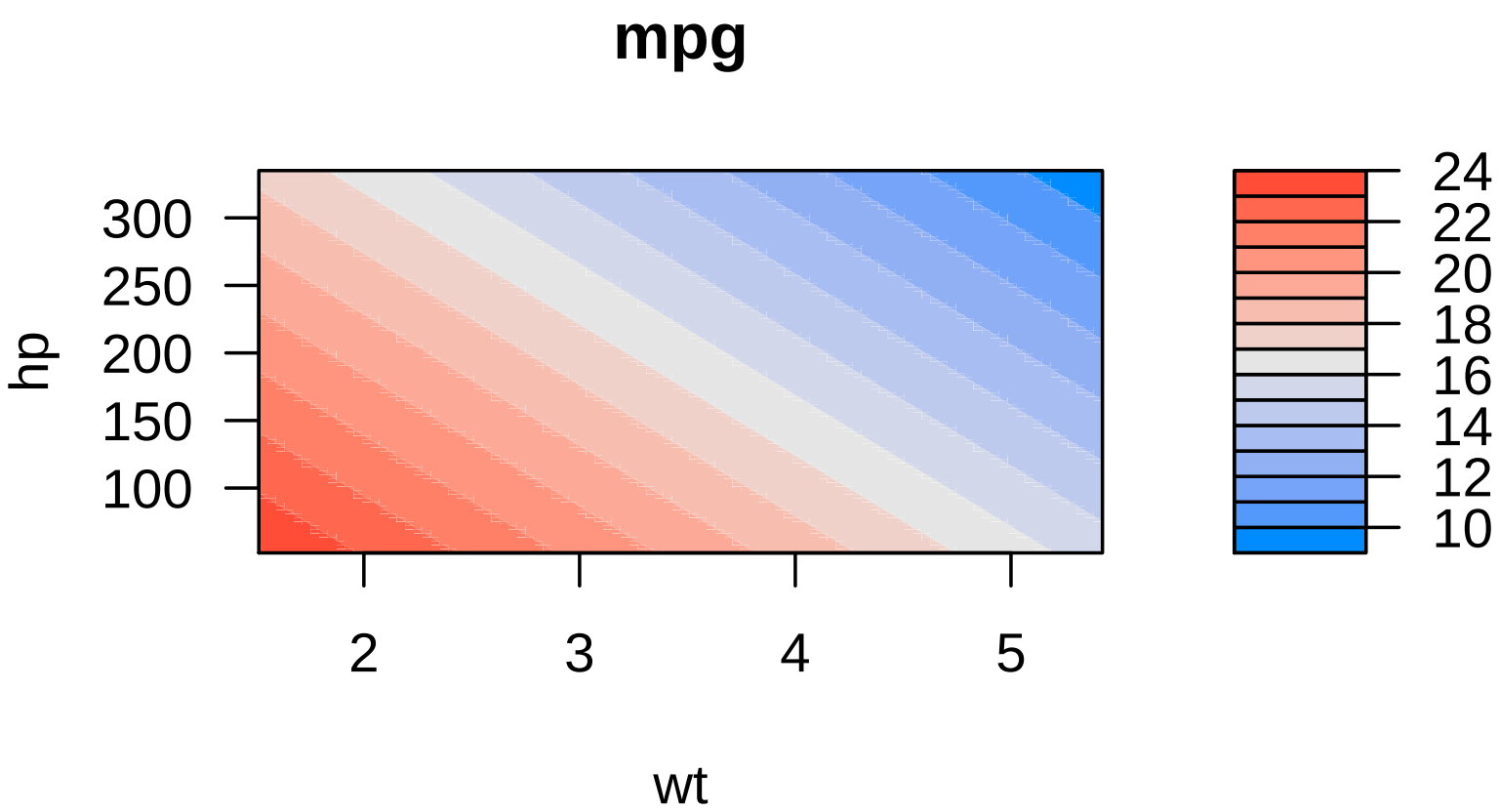
Control of the graphic output with `layout`.



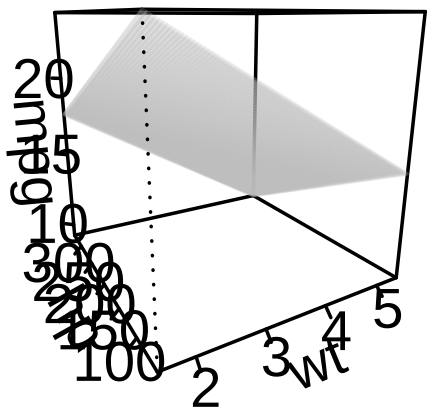
The package visreg - Interactions overlay



The package visreg - visreg2d



The package visreg - surface



Producing nice table output with Package stargazer

Example HTML output:

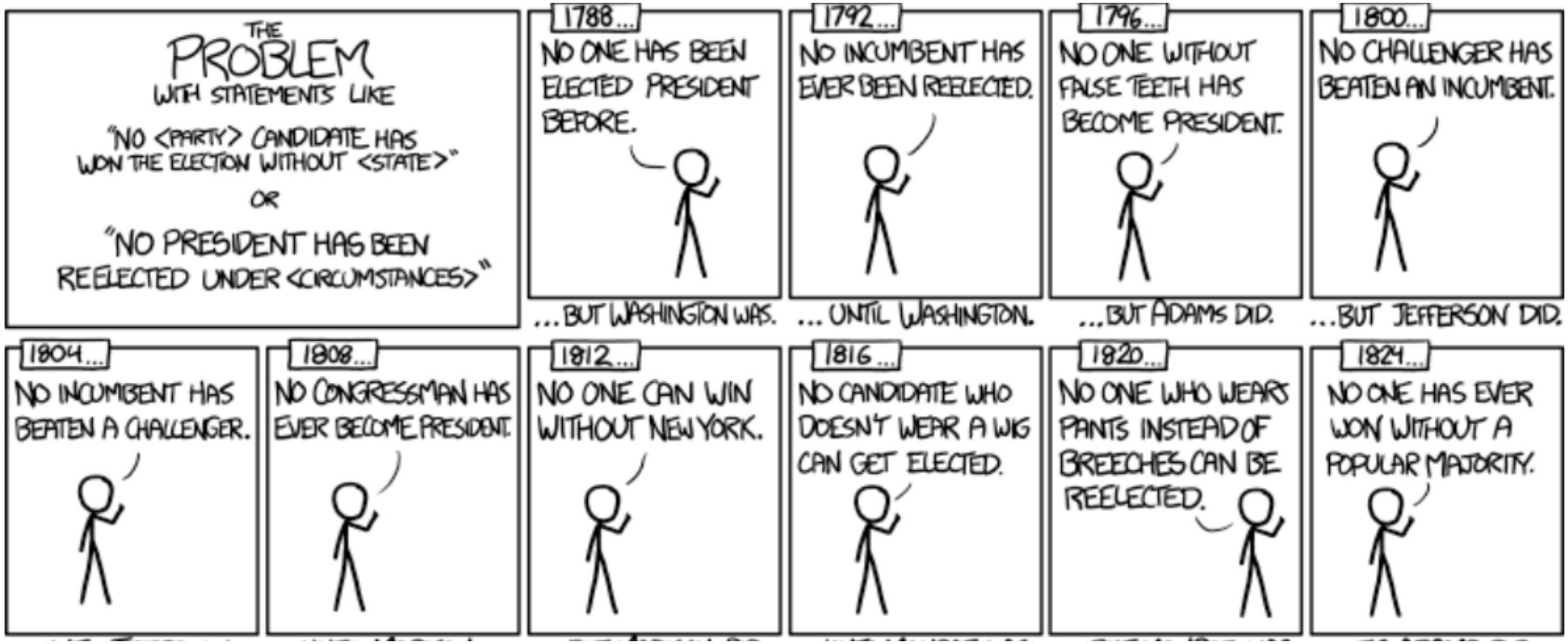
	Dependent variable:
	mpg
wt	-3.125***
	(0.911)
cyl	-1.510***
	(0.422)
am	0.176
	(1.304)
Constant	39.418***
	(2.641)
Observations	32
R^2	0.830

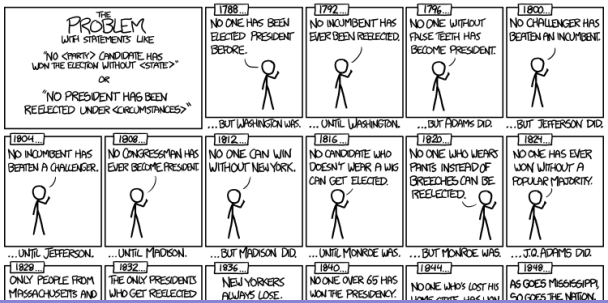
Exercise

- Install the package `AmesHousing` and create a processed Version of the Ames housing data with the variables `Sale_Price`, `Gr_Liv_Area` and `TotRms_AbvGrd`
- Create a suitable regression model with `Sale_Price` as dependent and `Gr_Liv_Area` and `TotRms_AbvGrd` as independent variables. What do you think?

The problem - Overfitting

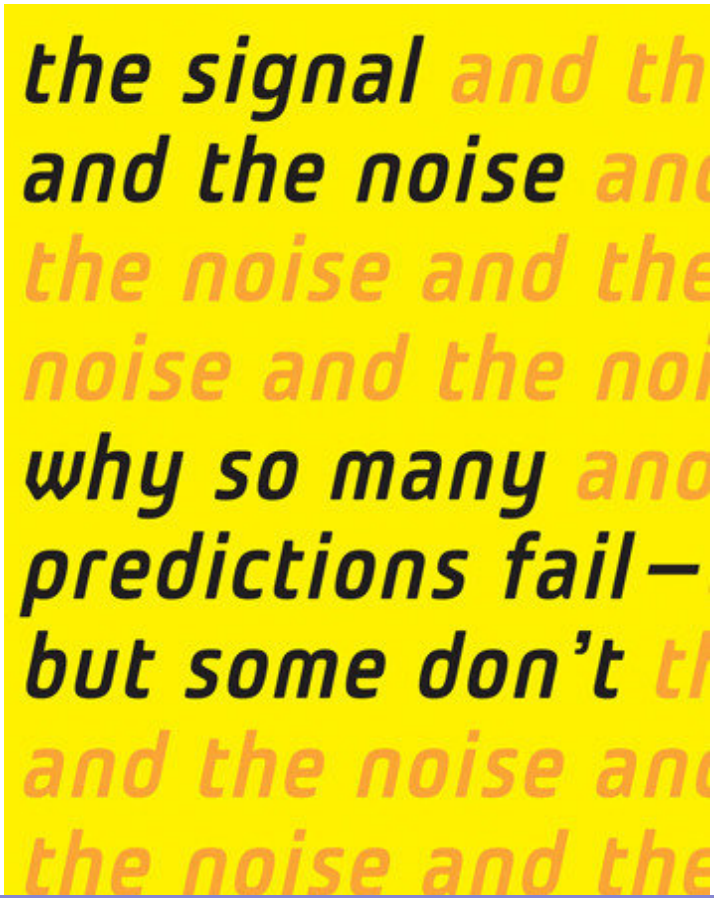
- Our model doesn't generalize well from our training data to unseen data.



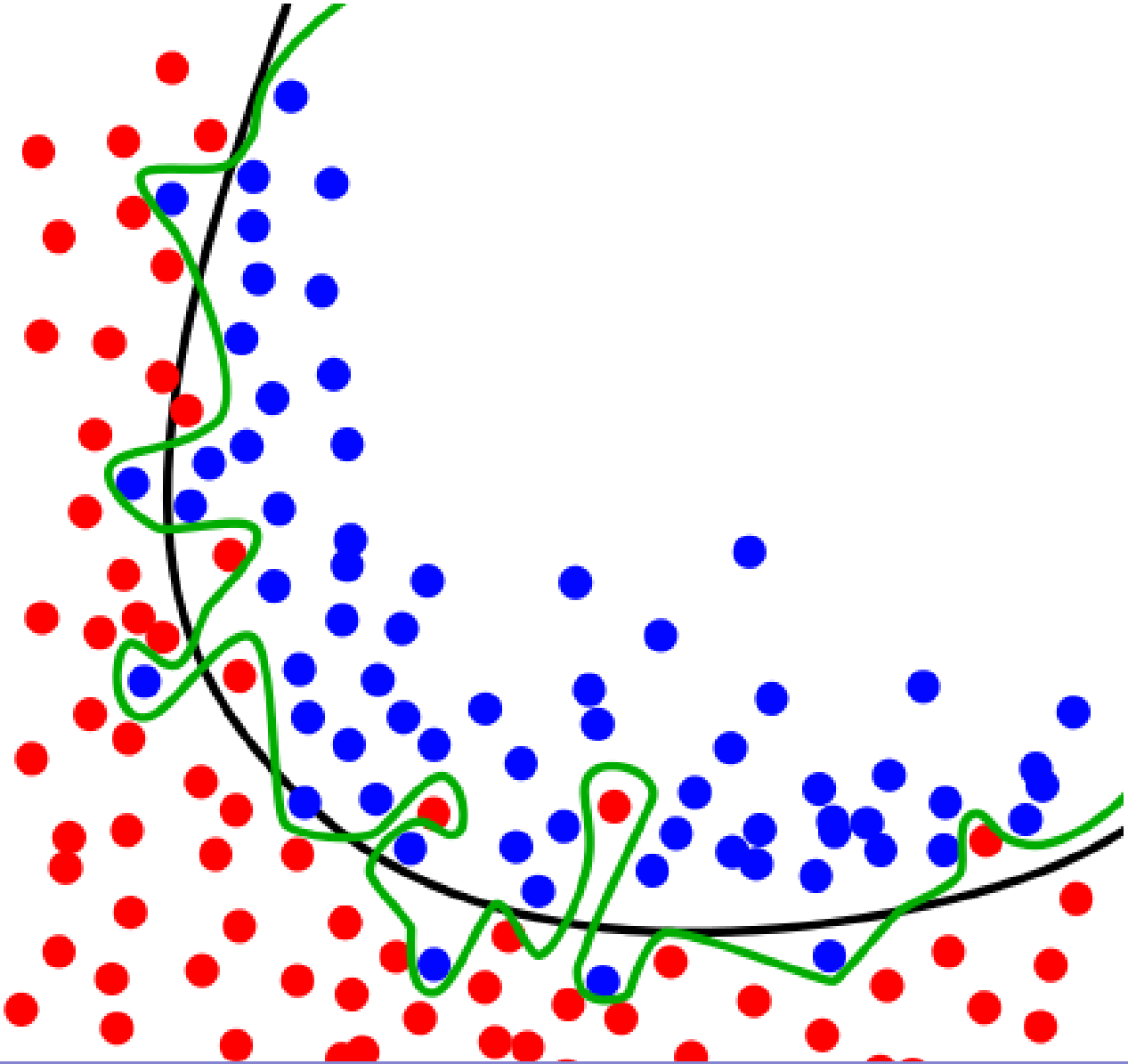


The Signal and the Noise

- In predictive modeling, you can think of the "signal" as the true underlying pattern that you wish to learn from the data.
- "Noise," on the other hand, refers to the irrelevant information or randomness in a dataset.



Overfitting.



What can be done against overvitting

- Cross Validation
- Train with more data
- Remove features
- Regularization - e.g. ridge and lasso regression
- Ensembling - e.g. bagging and boosting

Cross-validation

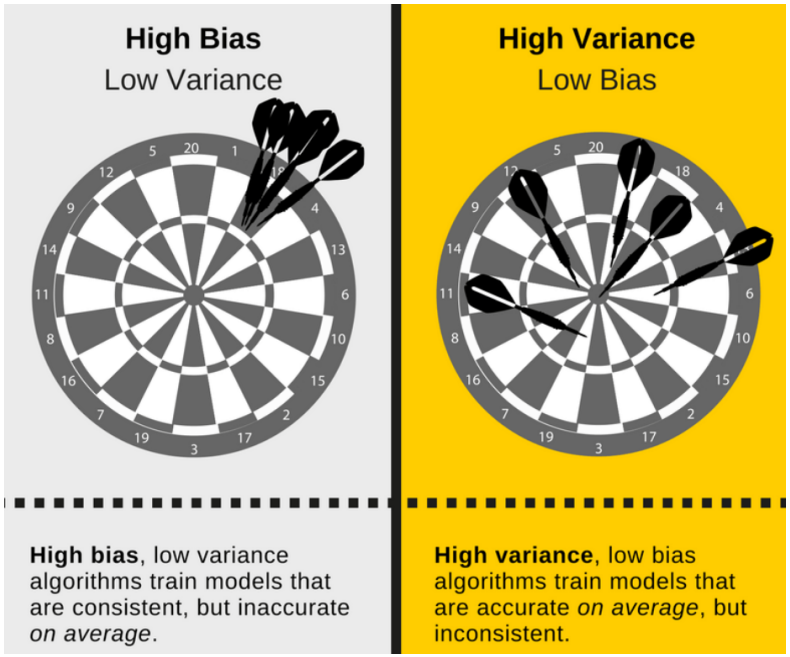
- Cross-validation is a powerful preventative measure against overfitting.
- Use your initial training data to generate multiple mini train-test splits. Use these splits to tune your model.

Cross Validation in R

```
## Linear Regression
##
## 47 samples
##  5 predictor
##
## No pre-processing
## Resampling: Leave-One-Out Cross-Validation
## Summary of sample sizes: 46, 46, 46, 46, 46, 46, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   7.738618  0.6128307  6.116021
##
```

K-FOLD CROSS VALIDATION

The bias variance tradeoff



Good literature for linear regression in R

Useful PDF document:

J H Maindonald - **Using R for Data Analysis and Graphics Introduction, Code and Commentary**

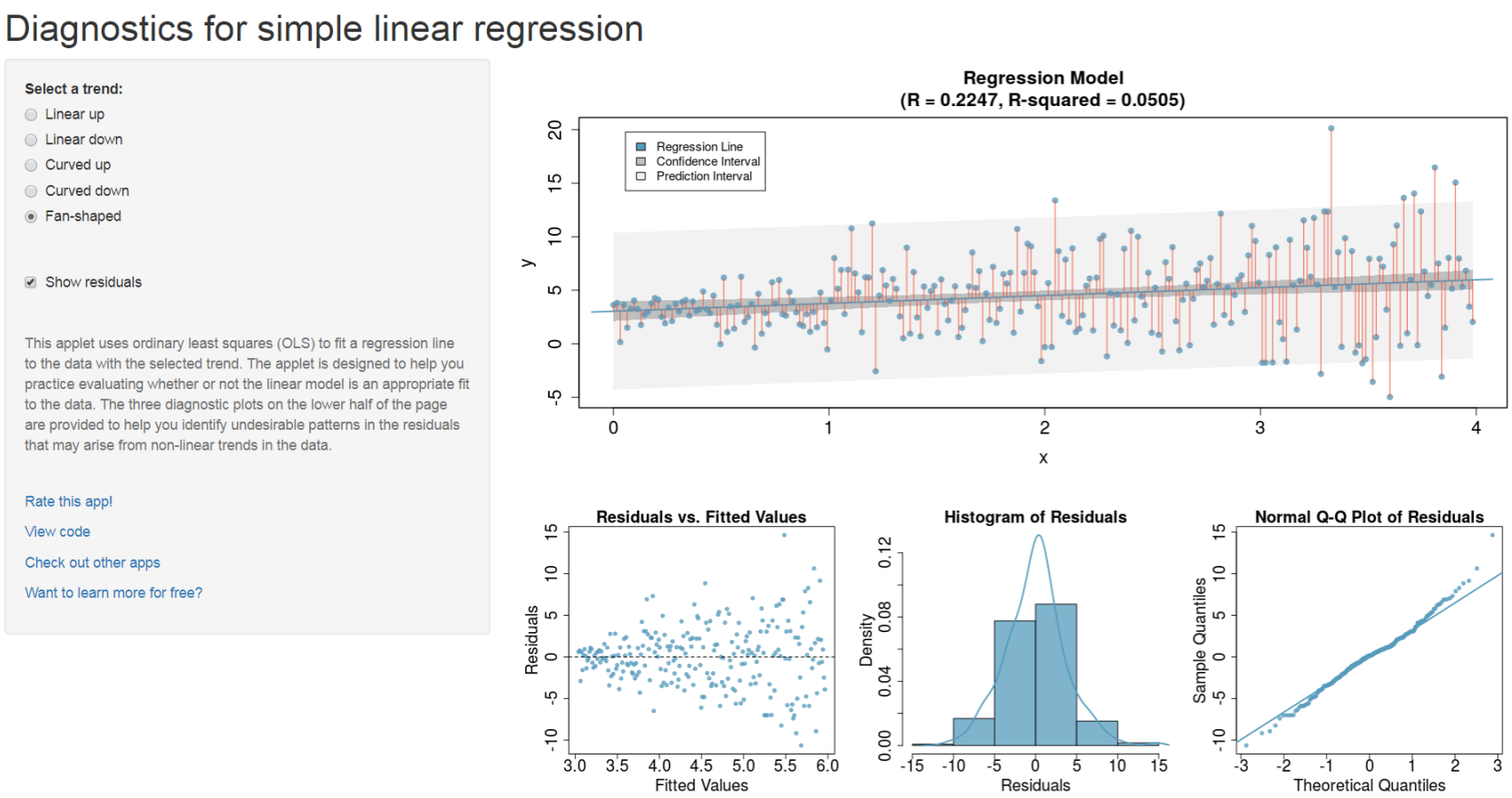
- Introduction to R
- Data analysis
- Statistical models
- Inference concepts
- Regression with one predictor
- Multiple linear regression
- Extending the linear model
- ...

Links - linear regression

- Regression - **r-bloggers**
- The complete book of **Faraway**- very intuitive
- Good introduction on **Quick-R**
- **Multiple regression**
- **15 Types of Regression you should know**
- `ggeffects` - **Create Tidy Data Frames of Marginal Effects for 'ggplot' from Model Outputs**

Shiny App - Diagnostics for simple linear regression

https://gallery.shinyapps.io/slr_diag/



- Shiny App - Simple Linear Regression

Resources

- machine learning iteration