Machine Learning
Regression

1

Regression models

- dataset $\mathcal{D}$ contains m observations and $n+1$ attributes
- n independent attributes (explanatory, predictors) and 1 dependent attribute (target, response)
- observations $\mathbf{x}_i, i \in \mathcal{M}$ are points in a n dimensional space, the target attribute is denoted as y_i
- $\mathbf{X}$ is the $m \times n$ matrix of data, and $\mathbf{y}$ is the target vector
- Y, X_j are random variables, $f : \mathbb{R}^n \rightarrow \mathbb{R}$

$$Y = f(X_1, X_2, \dots X_n).$$

2

Regression models

- spurious correlation
- the hypothesis space should be simple
  - linear

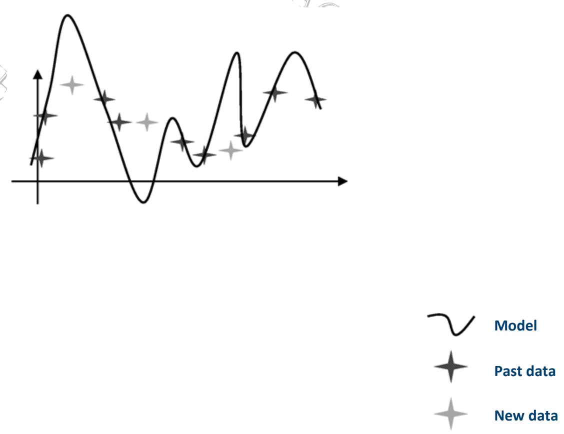
$$Y = w_1 X_1 + w_2 X_2 + \cdots + w_n X_n + b = \sum_{j=1}^{n} w_j X_j + b.$$

  - quadratic $\quad Y = b + wX + dX^2 \qquad Z = X^2$

$$Y = b + wX + dZ.$$

  - exponential $\quad Y = e^{b+wX} \qquad Z = \log Y. \qquad Z = b + wX.$

Accuracy vs. generalization



Simple linear regression

- **deterministic model**

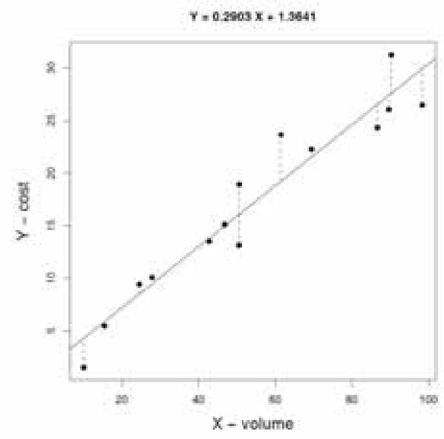
$$Y = wX + b.$$

- **probabilistic model**

$$Y = wX + b + \varepsilon,$$

Volume X (ton)	Cost Y (€×10³)	Volume X (ton)	Cost Y (€×10³)
10.11	1.53	42.87	13.51
50.56	13.14	61.53	23.65
90.28	31.24	24.60	9.43
15.50	5.47	46.85	15.12
69.52	22.27	50.63	18.94
98.40	26.47	89.68	26.06
86.66	24.32	27.91	10.08

Simple linear regression



Least squares (simple) linear regression

- residuals

$$e_i = y_i - f(x_i) = y_i - wx_i - b, \quad i \in \mathcal{M}.$$

- least squares regression: minimize the sum of squared residuals

$$\mathrm{SSE} = \sum_{i=1}^{m} e_i^2 = \sum_{i=1}^{m} [y_i - f(x_i)]^2 = \sum_{i=1}^{m} [y_i - wx_i - b]^2.$$

7

Least squares (simple) linear regression

- find the minimum

$$\frac{\partial\, \mathrm{SSE}}{\partial b} = -2 \sum_{i=1}^{m} [y_i - wx_i - b] = 0,$$

$$\frac{\partial\, \mathrm{SSE}}{\partial w} = -2 \sum_{i=1}^{m} x_i [y_i - wx_i - b] = 0.$$

- normal equation (linear system depending from the coefficients)

$$\begin{pmatrix} m & \sum_{i=1}^{m} x_i \\ \sum_{i=1}^{m} x_i & \sum_{i=1}^{m} x_i^2 \end{pmatrix} \begin{pmatrix} b \\ w \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{m} y_i \\ \sum_{i=1}^{m} x_i y_i \end{pmatrix}$$

8

Least squares (simple) linear regression

$$\hat{w} = \frac{\sigma_{xy}}{\sigma_{xx}},$$
$$\hat{b} = \bar{\mu}_y - \hat{w}\bar{\mu}_x,$$

$$\bar{\mu}_x = \frac{\sum_{i=1}^{m} x_i}{m}, \qquad \bar{\mu}_y = \frac{\sum_{i=1}^{m} y_i}{m}.$$

$$\sigma_{xy} = \sum_{i=1}^{m}(x_i - \bar{\mu}_x)(y_i - \bar{\mu}_y).$$
$$\sigma_{xx} = \sum_{i=1}^{m}(x_i - \bar{\mu}_x)^2,$$
$$\sigma_{yy} = \sum_{i=1}^{m}(y_i - \bar{\mu}_y)^2,$$

Least squares (simple) linear regression

- **prediction**

$$\hat{Y} = \hat{f}(X) = \hat{b} + \hat{w}X = \bar{\mu}_y + \frac{\sigma_{xy}}{\sigma_{xx}}(X - \bar{\mu}_x).$$

- **alternatively imposing a cross at the origin**

$$\hat{w} = \frac{\sum_{i=1}^{m} x_i y_i}{\sum_{i=1}^{m} x_i^2},$$
$$\hat{b} = b = 0,$$

Least squares (multiple) linear regression

- probabilistic model

$$Y = w_1 X_1 + w_2 X_2 + \cdots + w_n X_n + b + \varepsilon.$$

$$\mathbf{e} = (e_1, e_2, \ldots, e_m)$$

$$\mathbf{w} = (b, \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n)$$

- extend matrix X by a vector with all components = 1

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{e}.$$

Least squares linear regression

- sum of squared residuals

$$\mathrm{SSE} = \sum_{i=1}^{m} e_i^2 = \|\mathbf{e}\|^2 = \sum_{i=1}^{m} (y_i - \mathbf{w}'\mathbf{x}_i)^2$$
$$= (\mathbf{y} - \mathbf{X}\mathbf{w})'(\mathbf{y} - \mathbf{X}\mathbf{w}).$$

- null partial derivatives

$$\frac{\partial\,\mathrm{SSE}}{\partial \mathbf{w}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{w} = \mathbf{0}.$$

- normal equation

$$\mathbf{X}'\mathbf{X}\mathbf{w} = \mathbf{X}'\mathbf{y}.$$

- minimum point

$$\hat{\mathbf{w}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Least squares linear regression

- values predicted by model

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{w}} = (\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y} = \mathbf{H}\mathbf{y}.$$

- hat matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

- residuals

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y}.$$

13

Assumptions on the residuals

- random variable ε should follow a normal distribution of mean 0 and constant variance

$$\mathrm{E}(\varepsilon_i|\mathbf{x}_i) = 0,$$
$$\mathrm{Var}(\varepsilon_i|\mathbf{x}_i) = \sigma^2,$$

- residuals ε_i e ε_k should be independent
- estimate of σ

$$\bar{\sigma}^2 = \frac{\mathrm{SSE}}{m-n-1} = \frac{\sum_{i=1}^{m}(y_i - \mathbf{w}'\mathbf{x}_i)^2}{m-n-1} = \frac{\mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y}}{m-n-1}.$$

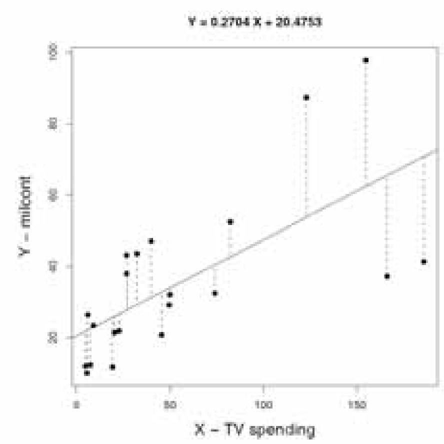
- if standard deviation σ is constant we have homoscedasticity, otherwise heteroscedasticity

14

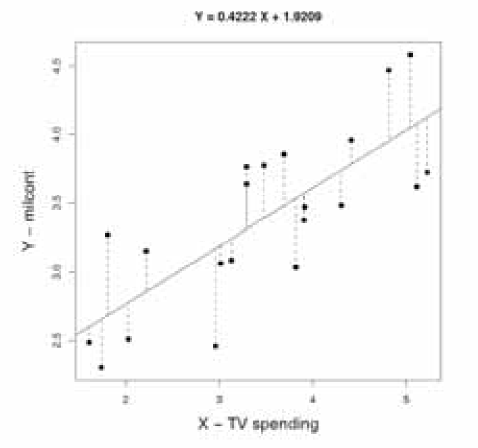
Heteroscedasticity

Company	TV spending (M$)	Milcont (Mil. weekly contacts)
MILLER.LITE	50.1	32.1
PEPSI	74.1	32.5
STROH'S	19.3	11.7
FEDERAL.EXPRESS	22.9	21.9
BURGER.KING	82.4	52.4
COCA-COLA	40.1	47.2
MC.DONALD'S	185.9	41.4
MCI	26.9	43.2
DIET.COLA	20.4	21.4
FORD	166.2	37.3
LEVI'S	123	87.4
BUD.LITE	45.6	20.8
ATT.BELL	154.9	97.9
CALVIN.KLEIN	5	12
WENDY'S	49.7	29.2
POLAROID	26.9	38
SHASTA	5.7	10
MEOW.MIX	7.6	12.3
OSCAR.MEYER	9.2	23.4
CREST	32.4	43.6
KIBBLES.N.BITS	6.1	26.4

Heteroscedasticity



Heteroscedasticity



17

Ridge regression

- the estimation of matrix $(\mathbf{X}'\mathbf{X})^{-1}$ can be critical (insufficient number of observations, multi-collinearity): ill-posed problem
- limit the width of the hypothesis space F (regularization theory)

$$\min_{\mathbf{w}} RR(\mathbf{w}, \mathcal{D}) = \min_{\mathbf{w}} \lambda \|\mathbf{w}\|^2 + \sum_{i=1}^{m} (y_i - \mathbf{w}'\mathbf{x}_i)^2$$
$$= \min_{\mathbf{w}} \lambda \|\mathbf{w}\|^2 + (\mathbf{y} - \mathbf{X}\mathbf{w})'(\mathbf{y} - \mathbf{X}\mathbf{w}).$$

18

Lasso regression

- instead of L_2 norm, use L_1 norm

$$\min_{\mathbf{w}} LR(\mathbf{w}, \mathcal{D}) = \min_{\mathbf{w}} \lambda\, |\mathbf{w}| + \sum_{i=1}^{m} (y_i - \mathbf{w}'\mathbf{x}_i)^2$$
$$= \min_{\mathbf{w}} \lambda\, |\mathbf{w}| + (\mathbf{y} - \mathbf{X}\mathbf{w})'(\mathbf{y} - \mathbf{X}\mathbf{w}).$$

Generalized linear models

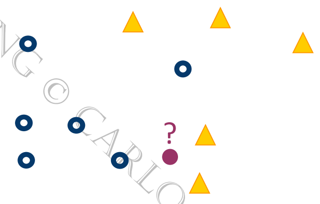
- Functions g_h represent any set of bases, such as polynomials, kernels and other groups of nonlinear functions

$$Y = \sum_h w_h g_h(X_1, X_2, \ldots, X_n) + b + \varepsilon$$

- Coefficients w_h and b can be determined through the minimization of the sum of squared errors. Function SSE in this formulation is more complex than for linear regression, solution of the minimization problem more difficult

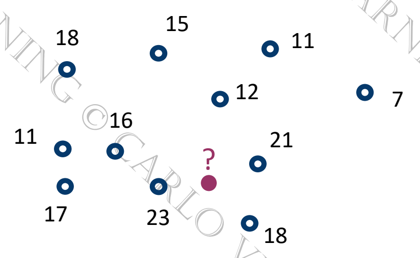
K-nearest neighbour (K-NN) for regression

Classification: majority voting among *K* nearest



K-nearest neighbour (K-NN) for regression

Regression: average of *K* nearest

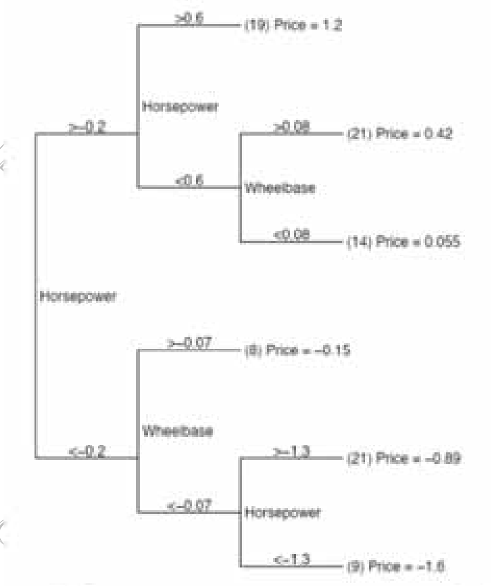


Regression trees

- replace Gini, entropy or misclassification with

$$S = \sum_{c \in \text{leaves}(T)} \sum_{i \in C} (y_i - m_c)^2$$

- where the prediction is $m_c = \frac{1}{n_c} \sum_{i \in C} y_i$

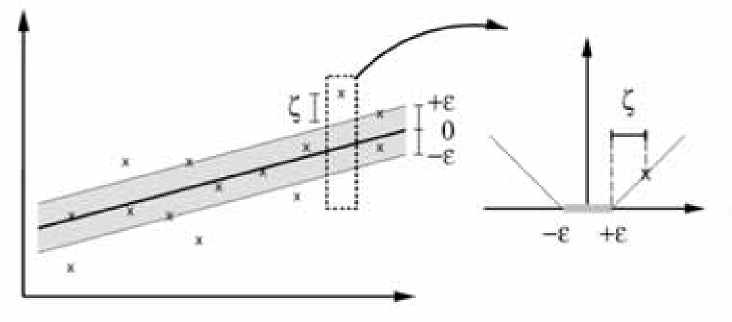


23

Support vector regression

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{\ell} (\xi_i + \xi_i^*) \\ \text{subject to} & \begin{cases} y_i - \langle w, x_i \rangle - b & \leq \varepsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i & \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* & \geq 0 \end{cases} \end{array}$$

$$|\xi|_\varepsilon := \begin{cases} 0 & \text{if } |\xi| \leq \varepsilon \\ |\xi| - \varepsilon & \text{otherwise} \end{cases}$$

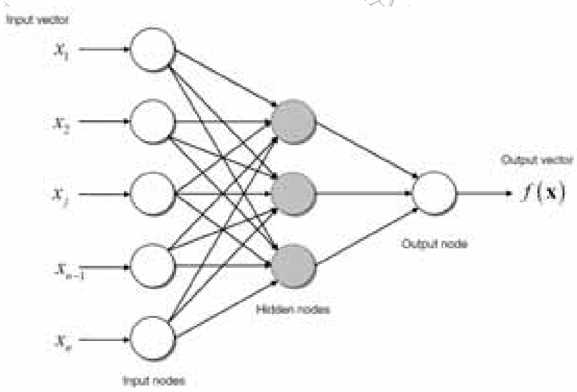


24

Neural networks for regression

- linear activation function for the output node

$$g(\mathbf{w}'\mathbf{x} - \vartheta) = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n - \vartheta = \mathbf{w}'\mathbf{x} - \vartheta$$

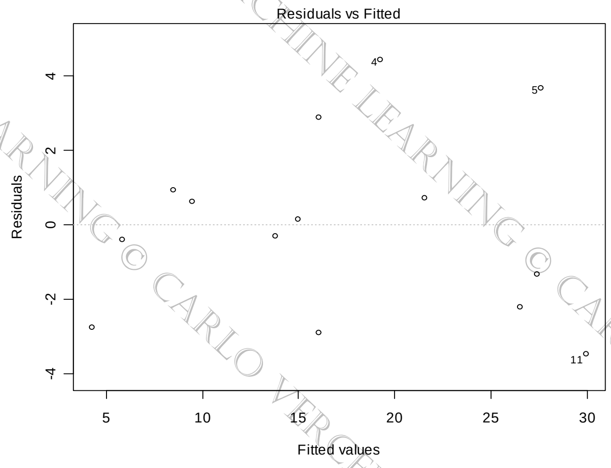


Normality and independence of the residuals

- goodness-of-fit test (chi-squared, Kolmogorov-Smirnov)
- graphical analysis
  - scatterplot residuals vs. fitted
  - scatterplot standardized residuals vs. fitted
  - qq-plot of the residuals

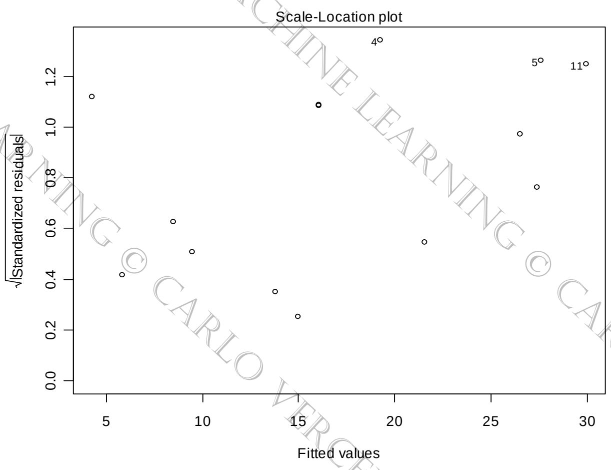
minimo	quartile inf.	mediana	quartile sup.	massimo
-3.46271	-1.98842	-0.07337	0.87332	4.42181

Scatterplot residuals vs. fitted



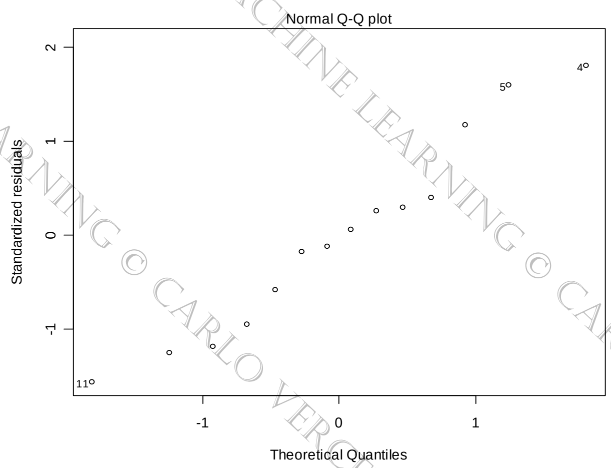
27

Scatterplot standardized residuals vs. fitted



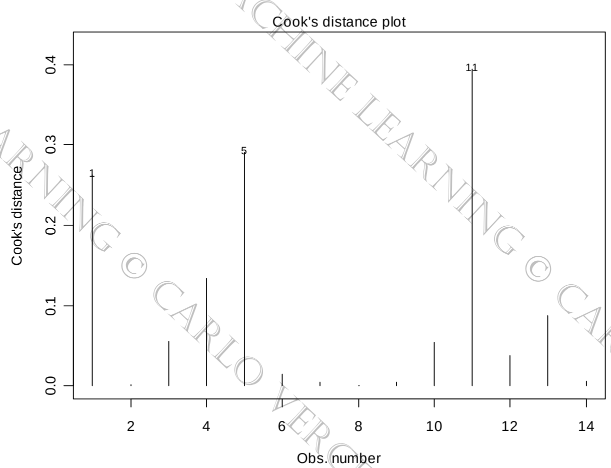
28

qq-plot of the residuals



29

Cook distance



30

Significance of the coefficients

- regression coefficients $\hat{\mathbf{w}}$ represent an estimate of **w**
- covariance matrix of the estimator $$\mathbf{V} = \mathrm{cov}(\hat{\mathbf{w}}) = \bar{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}.$$
- confidence intervals $$w_j \pm t_{\alpha/2}\sqrt{v_{jj}}.$$
- for simple regression $$w \pm t_{\alpha/2}\frac{\bar{\sigma}}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{\mu}_x)^2}},$$ $$b \pm t_{\alpha/2}\frac{\bar{\sigma}}{\sqrt{n}}\sqrt{1 + \frac{n\bar{\mu}_x}{\sum_{i=1}^{m}(x_i - \bar{\mu}_x)^2}}.$$

Significance of the coefficients

- hypothesis test

$$\text{(null hypothesis) } H_0 : w = 0, \quad \text{(alternative hypothesis) } H_a : w \neq 0$$

predictor	value	standard error	t-value	Pr > |t|
(intercept)	1.36411	1.48944	0.916	0.3780
volume	0.29033	0.02423	11.980	< 0.0001

Analysis of variance

- df: *n* degrees of freedom
- sum of sq:

$$\mathrm{RSS}_{reg} = \sum_{i=1}^{m} (\hat{y}_i - \bar{\mu}_y)^2 = 933.18.$$

- df: *m-n-1* degrees of freedom

- df: *m*-1 degrees of freedom
- sum of sq:

$$\mathrm{RSS}_{tot} = \sum_{i=1}^{m} (y_i - \bar{\mu}_y)^2 = 1011.20.$$

predictor	df	sum of sq.	mean sq.	F-value	$\Pr > F$
volume	1	933.18	933.18	143.53	< 0.0001
residuals	12	78.02	6.50		
total	13	1011.20	77.79		

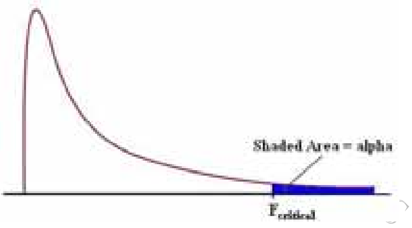
Analysis of variance

- the aim of regression models is to explain through predictive variables most part of variance inherent in dependent variable, leaving aside pure random fluctuation – residuals

- If this goal is achieved, one expects sample variance of residuals significantly smaller than sample variance of response variable

- if the residuals have normal distribution, the following ratio follows an *F* distribution with *n* e *m-n-1* degrees of freedom

$$F = \frac{\mathrm{RSS}_{\mathrm{reg}} / n}{\mathrm{SSE} / (m - n - 1)}$$



Determination coefficient

- determination coefficient

$$R^2 = \frac{\text{RSS}_{reg}}{\text{RSS}_{tot}} = \frac{\sum_{i=1}^{m}(\hat{y}_i - \bar{\mu}_y)^2}{\sum_{i=1}^{m}(y_i - \bar{\mu}_y)^2},$$

- in the example

$$R^2 = 0.9228.$$

- adjusted coefficient

$$R^2_{adj} = 1 - (1 - R^2)\frac{m-1}{m-n-1}.$$

- in the example

$$R^2_{adj} = 0.9164.$$

35

Determination and linear correlation coefficient

statistics	value
residual standard error	2.5500
multiple R-squared	0.9228
adjusted R-squared	0.9164
regression coeff.	0.9606

36

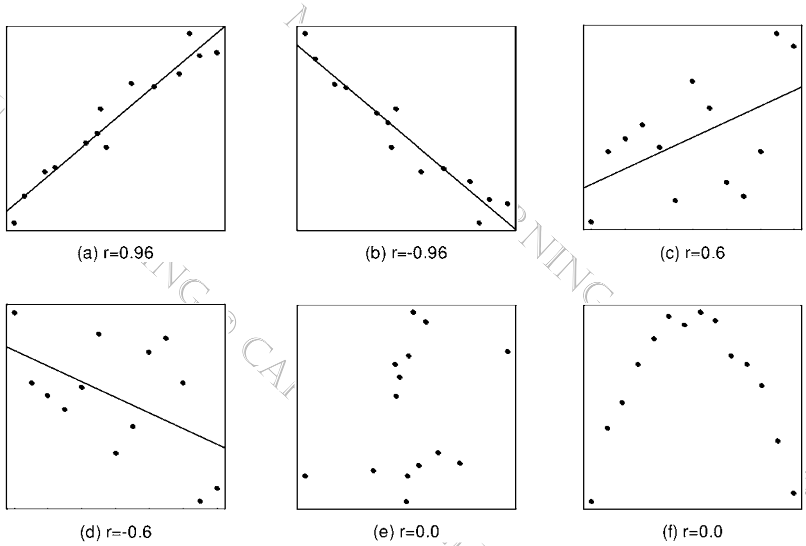
Linear correlation coefficient

$$r = \operatorname{corr}(\mathbf{y}, \mathbf{x}_i) = \frac{\sigma_{xy}}{\sqrt{\sigma_{xx}\sigma_{yy}}}$$
$$= \frac{\sum_{i=1}^{m}(x_i - \bar{\mu}_x)(y_i - \bar{\mu}_y)}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{\mu}_x)^2 \sum_{i=1}^{m}(y_i - \bar{\mu}_y)^2}}.$$

- **if r > 0, then *X* and *Y* are concordant.**
- **if r < 0, then *X* and *Y* are discordant.**
- **finally if r is close to 0, there is non linear relationship between *X* e *Y*.**

37

Linear correlation coefficient



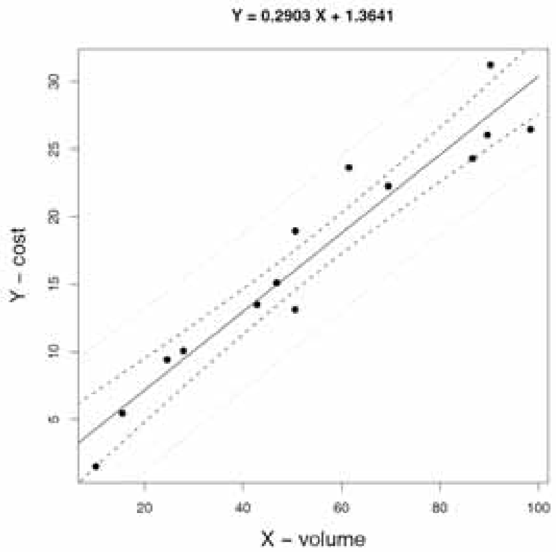
38

Multi-collinearity of the independent variables

- variance inflation factor: values greater than 5 point to the existence of multi-collinearity

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2}$$

Confidence and prediction limits



Confidence and prediction limits

- prediction associated to a new observation $\hat{y} = \hat{\mathbf{w}}\mathbf{x}$
- its variance is $\mathrm{Var}(\hat{y}) = \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}\bar{\sigma}^2,$
- for simple regression the confidence limit for E[*Y*] is

$$\hat{y} \pm t_{\alpha/2}\bar{\sigma}\sqrt{\frac{1}{m} + \frac{x - \bar{\mu}_x}{\sum_{i=1}^{m}(x_i - \bar{\mu}_x)^2}},$$

- whereas the prediction interval for *Y* is

$$\hat{y} \pm t_{\alpha/2}\bar{\sigma}\sqrt{1 + \frac{1}{m} + \frac{x - \bar{\mu}_x}{\sum_{i=1}^{m}(x_i - \bar{\mu}_x)^2}}.$$

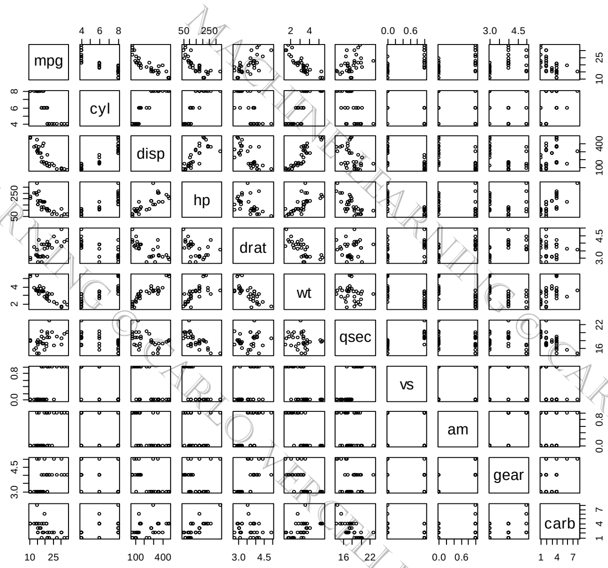
Example: mtcars

model	dependent variable	independent variable
A	mpg	cyl, disp, hp, drat, wt, qsec, vs, am, gear, carb
B	mpg	(backward) wt, qsec, am
C	mpg	hp, wt, (hp/wt)
D	log(mpg)	log(hp), log(wt)

model	res. std error	mult. *R*-sq.	adj. *R*-sq	*F*-statistic	*p*-value
A	2.6500	0.8690	0.8066	13.93	< 0.0001
B	2.4590	0.8497	0.8336	52.75	< 0.0001
C	2.1530	0.8848	0.8724	71.66	< 0.0001
D	1.1112	0.8829	0.8748	109.30	< 0.0001

- mpg : miles per (US) gallon;
- cyl : number of cylinders;
- disp : displacement (cu. in.);
- hp : gross horsepower;
- drat : rear axle ratio;
- wt : weight (lb/1000);
- qsec : 1/4 mile time;
- vs : engine shape (V/S);
- am : type of transmission (0 = automatic, 1 = manual);
- gear : number of forward gears;
- carb : number of carburetors.

Example: mtcars



Example: mtcars

predictor	value	std. error	t-value	Pr > |t|
(intercept)	12.30337	18.71788	0.657	0.5181
cyl	-0.11144	1.04502	-0.107	0.9161
disp	0.01334	0.01786	0.747	0.4635
hp	-0.02148	0.02177	-0.987	0.3350
drat	0.78711	1.63537	0.481	0.6353
wt	-3.71530	1.89441	-1.961	0.0633
qsec	0.82104	0.73084	1.123	0.2739
vs	0.31776	2.10451	0.151	0.8814
am	2.52023	2.05665	1.225	0.2340
gear	0.65541	1.49326	0.439	0.6652
carb	-0.19942	0.82875	-0.241	0.8122

predictor	df	sum of sq.	mean sq.	F-value	Pr > F
cyl	1	817.71	817.71	116.4245	< 0.0001
disp	1	37.59	37.59	5.3526	0.030911
hp	1	9.37	9.37	1.3342	0.261031
drat	1	16.47	16.47	2.3446	0.140644
wt	1	77.48	77.48	11.0309	0.003244
qsec	1	3.95	3.95	0.5623	0.461656
vs	1	0.13	0.13	0.0185	0.893173
am	1	14.47	14.47	2.0608	0.165858
gear	1	0.97	0.97	0.1384	0.713653
carb	1	0.41	0.41	0.0579	0.812179
residuals	21	147.49	7.02		
total	31	1126.04	985.57		

Example: mtcars

predictor	value	std. error	t-value	Pr > |t|
(intercept)	9.6178	6.9596	1.382	0.1780
wt	-3.9165	0.7112	-5.507	< 0.0001
qsec	1.2259	0.2887	4.247	0.0002
am	2.9358	1.4109	2.081	0.0467

predictor	df	sum of sq.	mean sq.	F-value	Pr > F
wt	1	847.73	847.73	140.2143	< 0.0001
qsec	1	82.86	82.86	13.7048	< 0.0001
am	1	26.18	26.18	4.3298	0.0467
residuals	28	169.29	6.05		
total	31	1126.06	962.82		

Example: mtcars

predictor	value	std. error	t-value	Pr > |t|
(intercept)	49.8084	3.60516	13.816	< 0.0001
hp	-0.12010	0.02470	-4.863	< 0.0001
wt	-8.21662	1.26971	-6.471	< 0.0001
hp/wt	0.02785	0.00742	3.753	< 0.0001

predictor	df	sum of sq.	mean sq.	F-value	Pr > F
hp	1	678.37	678.37	146.380	< 0.0001
wt	1	252.63	252.63	54.512	< 0.0001
hp:wt	1	65.29	65.29	14.088	< 0.0001
residuals	28	129.76	4.63		
total	31	1126.05	1000.92		

Example: mtcars

predictor	value	std. error	t-value	Pr > |t|
(intercept)	4.83469	0.22440	21.545	< 0.0001
log(hp)	-0.25532	0.05840	-4.372	< 0.0001
log(wt)	-0.56228	0.08742	-6.432	< 0.0001

predictor	df	sum of sq.	mean sq.	F-value	Pr > F
log(hp)	1	1.96733	1.96733	177.178	< 0.0001
log(wt)	1	0.45939	0.45939	41.373	< 0.0001
residuals	29	0.32201	0.01110		
total	31	2.74873	2.43782		