6. The general linear model (GLM)

Karl B Christensen
http://publicifsv.sund.ku.dk/~kach/SPSS

Contents

- Analysis of covariance (ANCOVA)
    - the general linear model
- Interaction
- Multiple regression

Data example: Framingham study

Data from the Framingham study

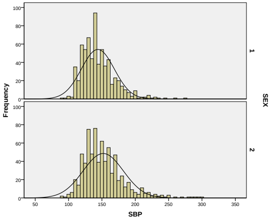
http://publicifsv.sund.ku.dk/~kach/SPSS/framingham.sav

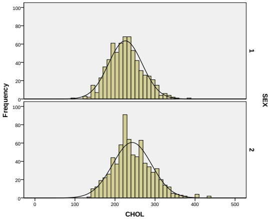
1406 persons. Variables

Variable	Explanation (Unit)
ID	subject id
SEX	gender (1 for males, 2 for females)
AGE	Age (years)
FRW	"Framingham relative weight" at baseline (%), range 52-222, 11 missing values
SBP	systolic blood pressure at baseline (mmHg), range 90-300
DBP	diastolic blood pressure at baseline (mmHg), range 50-160
CHOL	cholesterol at baseline (mg/100ml), range 96-430
CIG	cigarettes per day at baseline (n), 0-60, 1 missing value
CHD	coronary heart disease (0-10), 0: no CHD during follow-up, 1: CHD at baseline (prevalent cases), 2-10: CHD=x if CHD was diagnosed at follow-up no. x

Histograms for comparison of sex groups

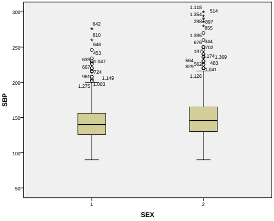
```
GET FILE='P:\framingham.sav'.
GRAPH
  /HISTOGRAM(NORMAL)=sbp
  /PANEL ROWVAR=sex ROWOP=CROSS.
GRAPH
  /HISTOGRAM(NORMAL)=chol
  /PANEL ROWVAR=sex ROWOP=CROSS.
```

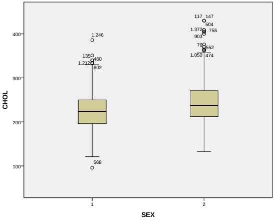




Box plots for comparison of sex groups

```
GET FILE='P:\framingham.sav'.
EXAMINE VARIABLES=sbp chol BY sex
  /PLOT=BOXPLOT
  /STATISTICS=NONE
  /NOTOTAL
  /PANEL ROWVAR=sexnr ROWOP=CROSS.
```





Box plots for comparison of sex groups

Notice the 'outliers' (indicated by their observation number) reflecting the skewed distribution. Can also use

```
GET FILE='P:\framingham.sav'.
EXAMINE VARIABLES=sbp chol BY sex
  /PLOT=HISTOGRAM
  /STATISTICS=NONE
  /NOTOTAL
  /PANEL ROWVAR=sexnr ROWOP=CROSS.
```

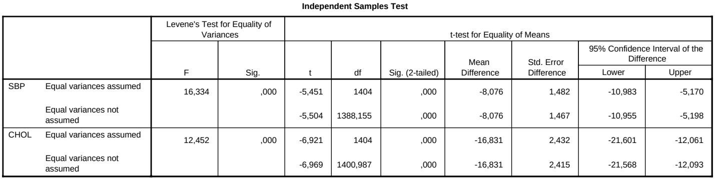
Group comparisons

- Note that for `EXAMINE VARIABLES` we can specify more than one
- Obvious sex difference for `sbp` as well as for `chol`

t-tests:

```
GET FILE='P:\framingham.sav'.
T-TEST GROUPS=sex(1 2)
  /MISSING=ANALYSIS
  /VARIABLES=sbp chol
  /CRITERIA=CI(.95).
```

Group comparisons



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
SBP	Equal variances assumed	16,334	,000	-5,451	1404	,000	-8,076	1,482	-10,983	-5,170
	Equal variances not assumed			-5,504	1388,155	,000	-8,076	1,467	-10,955	-5,198
CHOL	Equal variances assumed	12,452	,000	-6,921	1404	,000	-16,831	2,432	-21,601	-12,061
	Equal variances not assumed			-6,969	1400,987	,000	-16,831	2,415	-21,568	-12,093

Confounding when comparing groups

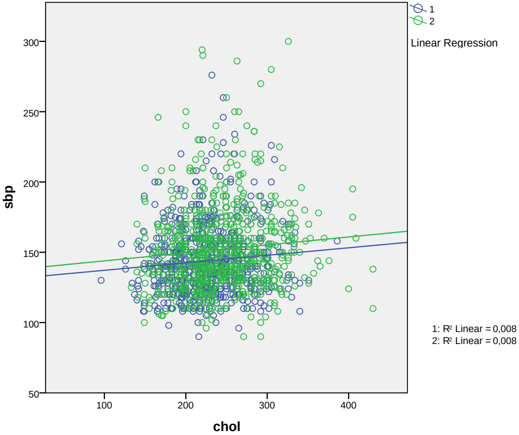
- Occurs if the distributions of some other relevant explanatory variables differ between the groups. Here "relevant" means things we would have liked to be the same (or at least very similar) for everybody, because we think of it as noise or distortion.
- Can be reduced by performing a regression analysis with the relevant variables as covariates.
- Confounding could be a problem in the current example ..

Relation between sbp and chol

```
GET FILE=’P:\framingham.sav’.
IGRAPH
/VIEWNAME=’Scatterplot’
/X1=VAR(chol) TYPE = SCALE
/Y=VAR(sbp) TYPE = SCALE
/COLOR = VAR(sex) TYPE = CATEGORICAL
/COORDINATE = VERTICAL
/FITLINE METHOD = REGRESSION LINEAR LINE = MEFFECT SPIKE=OFF
/CATORDER VAR(sex) (ASCENDING VALUES OMITEMPTY) /SCATTER COINCIDENT = NONE.
EXECUTE.
```

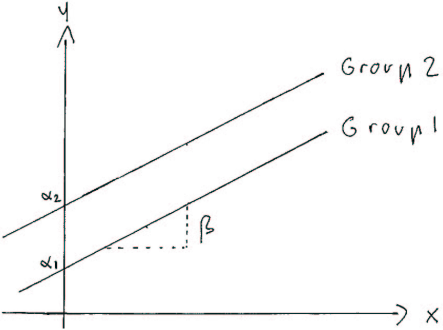
(or point-and-click: ’Set Markers by’, ’fit line at sub group’)

Relation between sbp and chol



Analysis of covariance (ANCOVA)

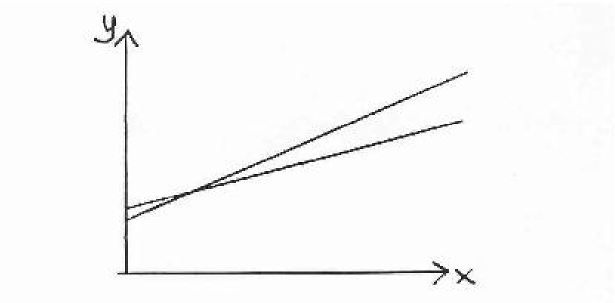
Comparison of parallel regression lines



Model: $y_{gi} = \alpha_g + \beta x_{gi} + \varepsilon_{gi}$ $\qquad g = 1, 2; i = 1, \cdots, n_g$

Here $\alpha_2 - \alpha_1$ is the expected difference in the response between the two groups *for fixed* value of the covariate, that is, when comparing any two subjects who have the same value of (match on) the covariate x ("adjusted for x").

But what if the lines are not parallel?
More general model: $y_{gi} = \alpha_g + \beta_g x_{gi} + \varepsilon_{gi}$



If $\beta_1 \neq \beta_2$ there is an *interaction* between `chol` and `sex`

Interaction

Interaction between a covariate and sex

- The effect of the covariate depends on sex
- The difference between men and women depends on the value of the covariate

Model with interaction

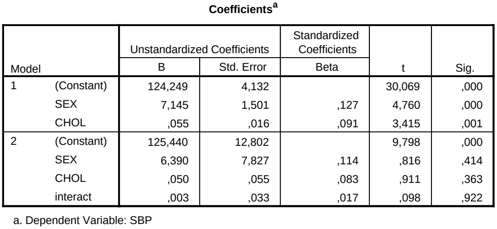
```
* Create interaction term and run a regression to see if interaction is significant.
COMPUTE interact = sex*chol.
EXECUTE.
REGRESSION
/STATISTICS COEFF ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT sbp
/METHOD=ENTER sex chol
/METHOD=ENTER interact.
```

Two models:

$$\mathtt{sbp}_i = \beta_0 + \beta_1 \cdot \mathtt{sex} + \beta_2 \cdot \mathtt{chol} + \epsilon_i$$

and a model with interaction added.

The interaction is not significant (p=0.922)



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	124,249	4,132		30,069	,000
	SEX	7,145	1,501	,127	4,760	,000
	CHOL	,055	,016	,091	3,415	,001
2	(Constant)	125,440	12,802		9,798	,000
	SEX	6,390	7,827	,114	,816	,414
	CHOL	,050	,055	,083	,911	,363
	interact	,003	,033	,017	,098	,922

a. Dependent Variable: SBP

The regression parameters

Model without interaction (two parallel lines):

$$\texttt{sbp}_i = 124.249 + 7.145 \cdot \texttt{sex} + 0.055 \cdot \texttt{chol} + \epsilon_i$$

Model with interaction:

$$\texttt{sbp}_i = 125.440 + 6.390 \cdot \texttt{sex} + 0.050 \cdot \texttt{chol} + 0.003 \cdot (\texttt{sex} \cdot \texttt{chol}) + \epsilon_i$$

Where are the two lines in the output?

Line for males (the reference group):

$$\begin{aligned}\texttt{sbp}_i &= 125.440 + 6.390 \cdot 1 + 0.050 \cdot \texttt{chol} + 0.003(1 \cdot \texttt{chol}) + \epsilon_i \\ &= (125.440 + 6.390) + (0.050 + 0.003) \cdot \texttt{chol} + \epsilon_i \\ &= 131.830 + 0.053 \cdot \texttt{chol} + \epsilon_i\end{aligned}$$

Line for females:

$$\begin{aligned}\texttt{sbp}_i &= 125.440 + 6.390 \cdot 2 + 0.050 \cdot \texttt{chol} + 0.003 \cdot (2 \cdot \texttt{chol}) + \epsilon_i \\ &= 138.220 + 0.056 \cdot \texttt{chol} + \epsilon_i\end{aligned}$$

slopes are almost equal: 0.053 and 0.056. The difference between them is 0.003

Same model, new procedure

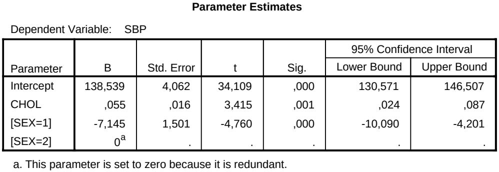
Analyze-General Linear Model-Univariate

```
UNIANOVA SBP BY SEX WITH CHOL
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /PRINT=PARAMETER
  /PLOT=RESIDUALS
  /CRITERIA=ALPHA(.05)
  /DESIGN=CHOL SEX.
```

http://publicifsv.sund.ku.dk/~kach/SPSS/F6_gif1.gif

Note: ask for residual plots

Same model, new procedure



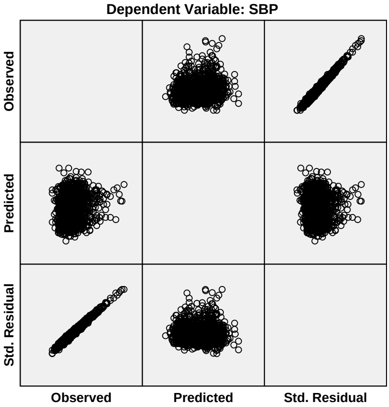
Parameter Estimates

Dependent Variable: SBP

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	138,539	4,062	34,109	,000	130,571	146,507
CHOL	,055	,016	3,415	,001	,024	,087
[SEX=1]	-7,145	1,501	-4,760	,000	-10,090	-4,201
[SEX=2]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

Same model, new procedure



Model: Intercept + CHOL + SEX

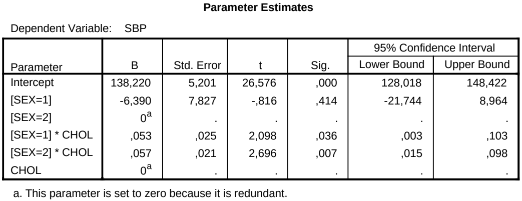
Same model, new procedure

Add interaction in `UNIANOVA`

```
UNIANOVA SBP BY SEX WITH CHOL
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /SAVE=PRED RESID
  /PRINT=PARAMETER
  /PLOT=RESIDUALS
  /CRITERIA=ALPHA(.05)
  /DESIGN=SEX CHOL*SEX CHOL.
```

Same model, new procedure

Add interaction in `UNIANOVA`



Parameter Estimates

Dependent Variable: SBP

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	138,220	5,201	26,576	,000	128,018	148,422
[SEX=1]	-6,390	7,827	-,816	,414	-21,744	8,964
[SEX=2]	0[a]	.	.	.	.	.
[SEX=1] * CHOL	,053	,025	2,098	,036	,003	,103
[SEX=2] * CHOL	,057	,021	2,696	,007	,015	,098
CHOL	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

Two different parameterizations

- (extrapolated) level at covariate=0 for reference group
- (extrapolated) difference between groups at covariate=0
- An effect of the covariate (the slope) for the reference group
- The difference between the slopes for the two groups

Another

- The (extrapolated) level at covariate=0 for each group
- The effect of the covariate (the slope) for each group

Confounding?

In this example it seems that

1. difference between a randomly chosen man and a randomly chosen woman is

   $$8.076$$

2. difference between a randomly chosen man and a randomly chosen woman with the same value of `CHOL` is

   $$7.145$$

Exercise: Another look the Juul data

1. Create a new data set including only individuals above 25 years, and make a new variable with log-transformed SIGF1.
2. Plot relationship between age and log-transformed SIGF-I.
3. Make separate regression lines for men and women.
4. Do a regression analysis to explore if slopes are equal in men and women.
5. Give an estimate for the difference in slopes, with 95% confidence interval.
6. Can we interpret this estimate on the original scale ?

Model checking

Use

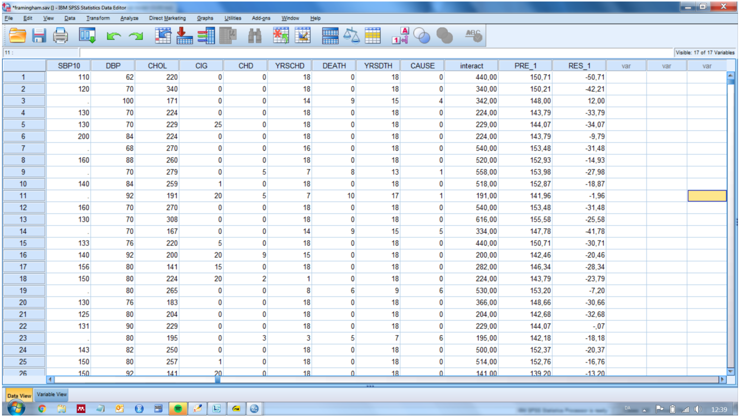
```
/SAVE=PRED RESID
```

in the syntax

```
UNIANOVA SBP BY SEX WITH CHOL
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /PRINT=PARAMETER
  /PLOT=RESIDUALS
  /CRITERIA=ALPHA(.05)
  /DESIGN=CHOL SEX.
```

In addition to the plots for from last time residuals should be plotted against sex in order to check variance homogeneity

Model checking



	SBP10	DBP	CHOL	CIG	CHD	YRSCHD	DEATH	YRSDTH	CAUSE	interact	PRE_1	RES_1	var	var	var
1	110	62	220	0	0	18	0	18	0	440,00	150,71	-50,71			
2	120	70	340	0	0	18	0	18	0	340,00	150,21	-42,21			
3	.	100	171	0	0	14	9	15	4	342,00	148,00	12,00			
4	130	70	224	0	0	18	0	18	0	224,00	143,79	-33,79			
5	130	70	229	25	0	18	0	18	0	229,00	144,07	-34,07			
6	200	84	224	0	0	18	0	18	0	224,00	143,79	-9,79			
7	.	68	270	0	0	16	0	18	0	540,00	153,48	-31,48			
8	160	88	260	0	0	18	0	18	0	520,00	152,93	-14,93			
9	.	70	279	0	5	7	8	13	1	558,00	153,98	-27,98			
10	140	84	259	1	0	18	0	18	0	518,00	152,87	-18,87			
11	.	92	191	20	5	7	10	17	1	191,00	141,96	-1,96			
12	160	70	270	0	0	18	0	18	0	540,00	153,48	-31,48			
13	130	70	308	0	0	18	0	18	0	616,00	155,58	-25,58			
14	.	70	167	0	0	14	9	15	5	334,00	147,78	-41,78			
15	133	76	220	5	0	18	0	18	0	440,00	150,71	-30,71			
16	140	92	200	20	9	15	0	18	0	200,00	142,46	-20,46			
17	156	80	141	15	0	18	0	18	0	282,00	146,34	-28,34			
18	150	80	224	20	2	1	0	18	0	224,00	143,79	-23,79			
19	.	80	265	0	0	8	6	9	6	530,00	153,20	-7,20			
20	130	76	183	0	0	18	0	18	0	366,00	148,66	-30,66			
21	125	80	204	0	0	18	0	18	0	204,00	142,68	-32,68			
22	131	90	229	0	0	18	0	18	0	229,00	144,07	-,07			
23	.	80	195	0	3	3	5	7	6	195,00	142,18	-18,18			
24	143	82	250	0	0	18	0	18	0	500,00	152,37	-20,37			
25	150	80	257	1	0	18	0	18	0	514,00	152,76	-16,76			
26	150	92	141	20	0	18	0	18	0	141,00	139,20	-13,20			

Model checking

Use

```
/SAVE=PRED RESID
```

in the syntax

```
UNIANOVA SBP BY SEX WITH CHOL
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /PRINT=PARAMETER
  /PLOT=RESIDUALS
  /CRITERIA=ALPHA(.05)
  /DESIGN=CHOL SEX.
```

In addition to the plots from last time residuals should be plotted against sex in order to check variance homogeneity

Model checking

Residuals should be plotted against:

1. the explanatory variable x_i – to check linearity
2. the fitted values $\hat{y}_i$ – to check variance homogeneity (and normality)
3. '*normal scores*' i.e. probability plot or histogram – to check normality

First two should give impression random scatter, while the probability plot ought to show a straight line.

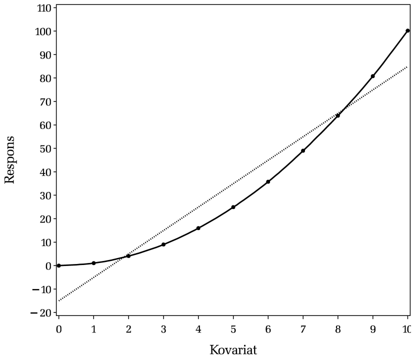
In addition to these plots from last time residuals should be plotted against sex in order to check variance homogeneity

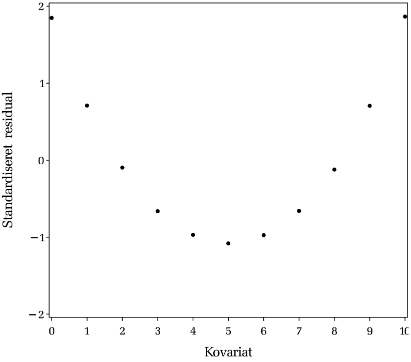
Model checking

Residuals plotted against the explanatory variable x_i – to check linearity

```
GRAPH
  /SCATTERPLOT(BIVAR)=chol WITH RES_1
  /MISSING=LISTWISE.
```

Look for $\cup$ or $\cap$ forms

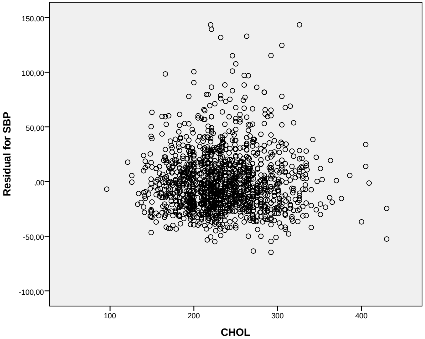




Model checking

Residuals plotted against the explanatory variable x_i – to check linearity

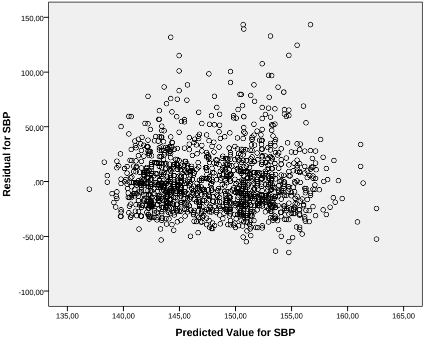
```
GRAPH
  /SCATTERPLOT(BIVAR)=chol WITH RES_1
  /MISSING=LISTWISE.
```



Model checking

Residuals plotted against the fitted values $\hat{y}_i$ – to check variance homogeneity (and normality)

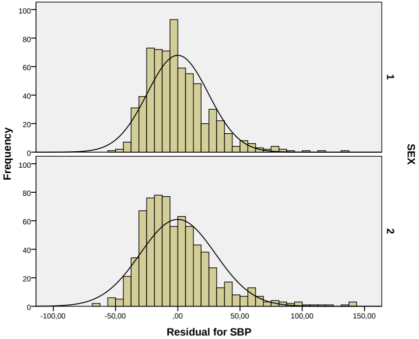
```
GRAPH
  /SCATTERPLOT(BIVAR)=PRE_1 WITH RES_1
  /MISSING=LISTWISE.
```



Model checking

Residuals plotted against '*normal scores*' i.e. probability plot or histogram – to check normality

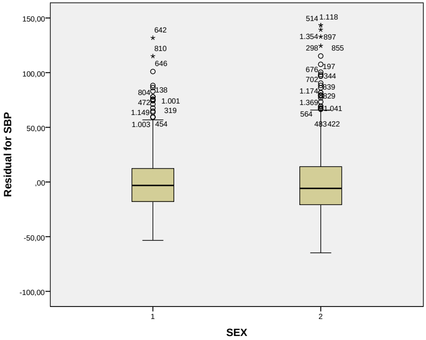
```
GRAPH
  /HISTOGRAM(NORMAL)=RES_1
  /PANEL ROWVAR=sex ROWOP=CROSS.
```



Model checking

Residuals plotted against sex in order to check variance homogeneity

```
EXAMINE VARIABLES=RES_1 BY sex
  /PLOT=BOXPLOT
  /STATISTICS=NONE
  /NOTOTAL
  /PANEL ROWVAR=sexnr ROWOP=CROSS.
```



Exercise: Another look the framingham data

1. Create a new data set with a new variable with log-transformed SBP.
2. Plot relationship between `chol` and log-transformed SBP.
3. Make separate regression lines for men and women.
4. Do a regression analysis to explore if slopes are equal in men and women.
5. Evaluate model fit

Multiple regression. General linear model (GLM).

Data: n sets of observations, made on the same 'unit':

unit	$x_1 .... x_p$	y
1	$x_{11} .... x_{1p}$	y_1
2	$x_{21} .... x_{2p}$	y_2
3	$x_{31} .... x_{3p}$	y_3
.	. . . . . .	.
n	$x_{n1} .... x_{np}$	y_n

The *linear regression model* with p explanatory variables[1] is written:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

[1] often called 'covariates'

Interpretation of regression coefficients

Model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$. Consider two subjects:
A has covariate values $(X_1, X_2, \ldots, X_p)$
B has covariate values $(X_1 + 1, X_2, \ldots, X_p)$
Expected difference in the response $(B - A)$

$$[\beta_0 + \beta_1(X_1 + 1) + \beta_2 X_{i2} + ...] - [\beta_0 + \beta_1 X_1 + \beta_2 X_{i2} + ...] = \beta_1$$

This means that β_1 is the effect of one unit's difference in X_1 *for fixed levels of the other variables* $(X_2, \ldots, X_p)$