6. The general linear model

Use of SAS
March 2011

Contents

Example on lung capacity

32 patients for heart/lung transplantation

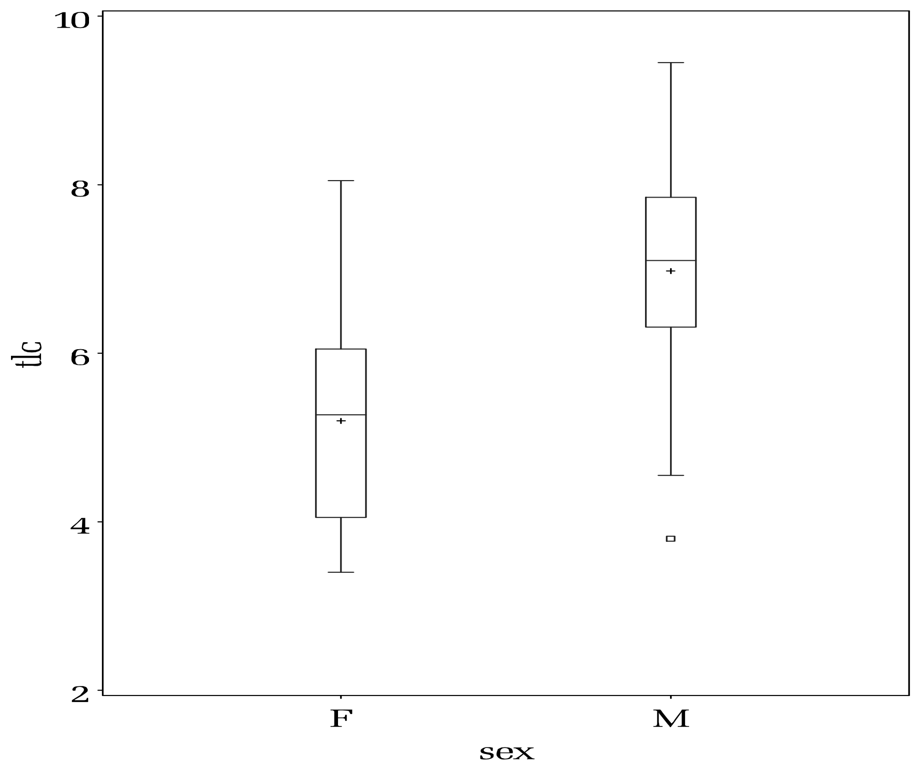
`TLC` (Total Lung Capacity) is determined from whole-body plethysmography

Are men and women different with respect to total lung capacity?

OBS	SEX	AGE	HEIGHT	TLC
1	F	35	149	3.40
2	F	11	138	3.41
3	M	12	148	3.80
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
29	F	20	162	8.05
30	M	25	180	8.10
31	M	22	173	8.70
32	M	25	171	9.45

Box plots for comparison of gender groups

```
proc boxplot data=tlc;
  plot tlc*sex / height=3 boxstyle=schematic;
run;
```



Marginal comparisons

```
proc ttest data=tlc;
  class sex;
  var tlc height;
run;
```

```
The TTEST Procedure
                                Statistics

                             Lower CL          Upper CL  Lower CL
Variable  sex            N       Mean     Mean     Mean   Std Dev  Std Dev

tlc       F             16      4.505   5.1981   5.8913    0.9609   1.3008
tlc       M             16     6.2106   6.9769   7.7431    1.0623    1.438
tlc       Diff (1-2)           -2.769   -1.779   -0.789    1.0957   1.3711
height    F             16     155.82   160.81    165.8    6.9203   9.3682
height    M             16     168.38   174.06   179.74    7.8755   10.661
height    Diff (1-2)            -20.5   -13.25   -6.004    8.0195   10.036
```

Statistics

Variable	sex	Upper CL Std Dev	Std Err	Minimum	Maximum
tlc	F	2.0133	0.3252	3.4	8.05
tlc	M	2.2256	0.3595	3.8	9.45
tlc	Diff (1-2)	1.8328	0.4848		
height	F	14.499	2.342	138	177
height	M	16.5	2.6653	148	189
height	Diff (1-2)	13.414	3.5481		

T-Tests

Variable	Method	Variances	DF	t Value	Pr > |t|
tlc	Pooled	Equal	30	-3.67	0.0009
tlc	Satterthwaite	Unequal	29.7	-3.67	0.0009
height	Pooled	Equal	30	-3.73	0.0008
height	Satterthwaite	Unequal	29.5	-3.73	0.0008

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F
tlc	Folded F	15	15	1.22	0.7028
height	Folded F	15	15	1.30	0.6228

Obvious gender difference for `tlc` as well as `height`

Confounding when comparing groups

- occurs if the distribution of an important explanatory variable differ between the groups

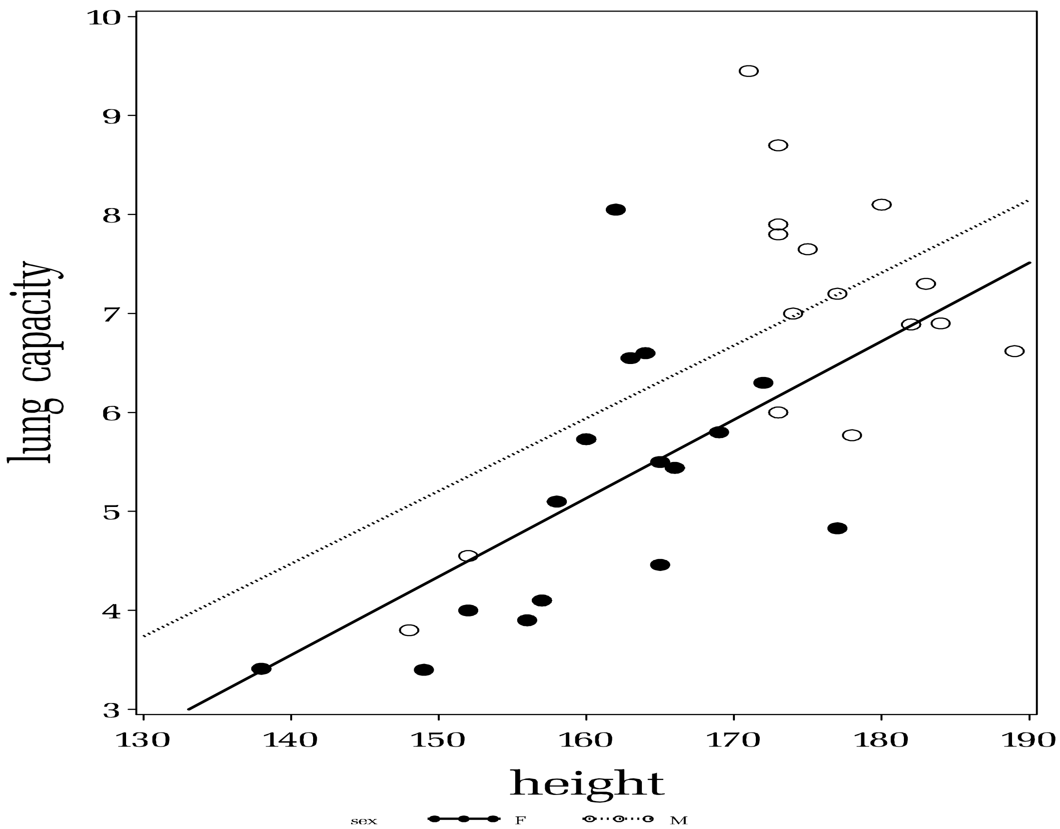
Can be avoided by performing a **regression analysis** with the relevant variables as covariates.

Example:

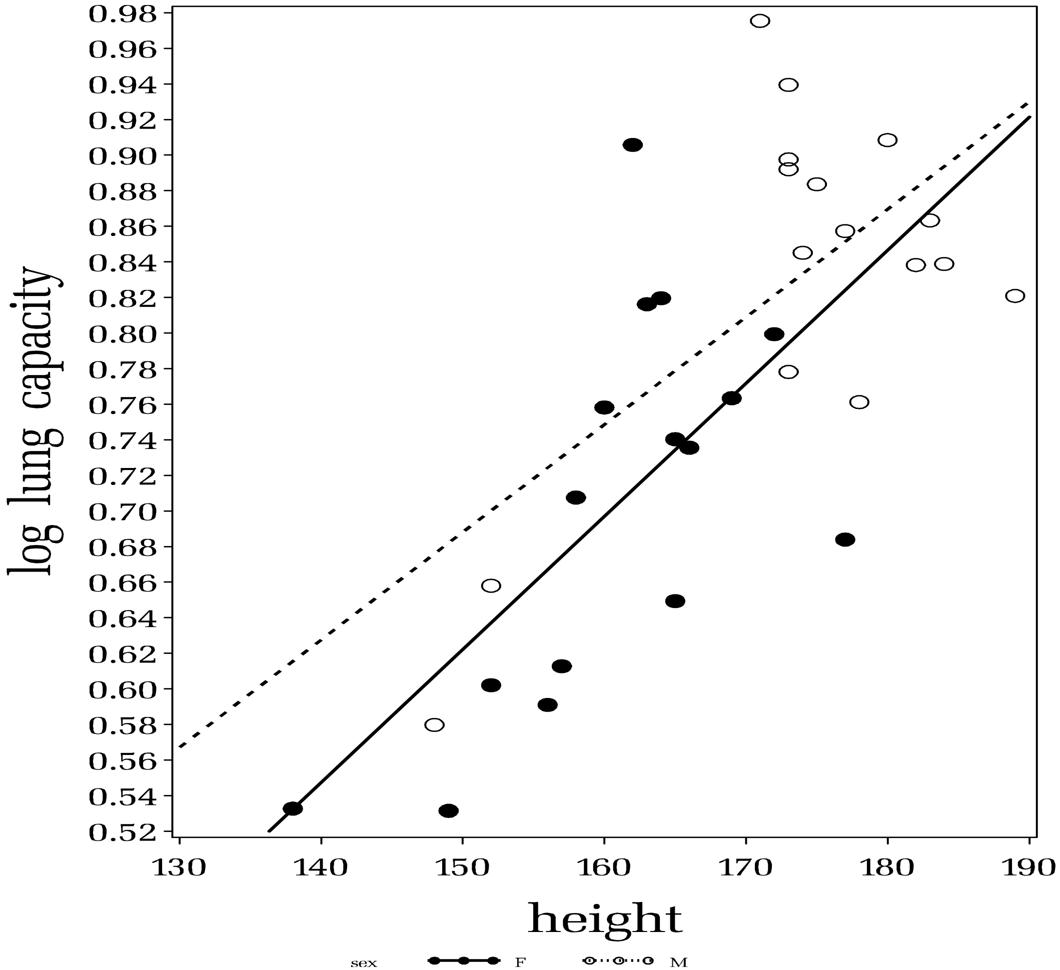
- Comparison of lung function between men and women
  — they are not of equal height

Relation between `tlc` and `height`:

```
proc gplot data=tlc;
  plot tlc*height=sex;
symbol1 v=dot i=rl c=BLACK l=1 w=2 h=2;
symbol2 v=circle i=rl c=BLACK l=33 w=2 h=2;
run;
```

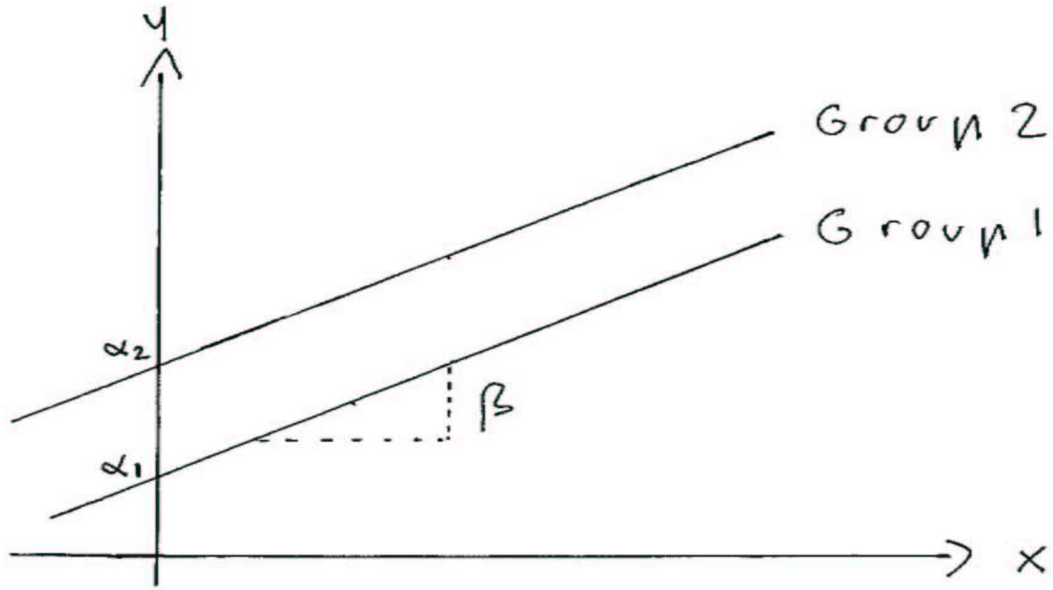


Relation between `tlc` (after transformation with base 10 logarithms) and **`height`**,



Analysis of covariance

Comparison of **parallel** regression lines



Model:

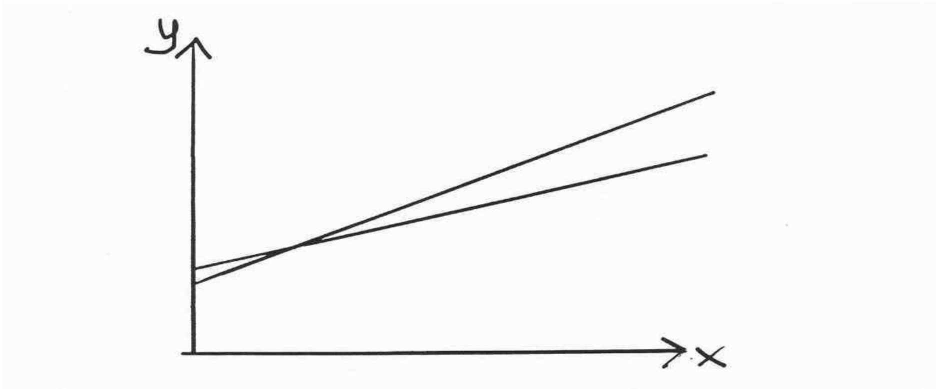
$y_{gi} = \alpha_g + \beta x_{gi} + \varepsilon_{gi} \qquad\qquad g = 1, 2; i = 1, \cdots, n_g$

Here $\alpha_2 - \alpha_1$ is the expected difference in the response between the two groups *for fixed* value of the covariate

We have adjusted for x.

But what if the lines are not at all parallel?

More **general model:** $y_{gi} = \alpha_g + \beta_g x_{gi} + \varepsilon_{gi}$



When $\beta_1 \neq \beta_2$, we say that there is **interaction** between `height` and `sex`

- The effect of height depends on gender
- The difference between men and women depends on height

In case of interaction: Do not interpret marginal effects.

Model with interaction

```
proc glm data=tlc;
  class sex;
  model ltlc=sex height sex*height / solution;
run;
```

```
The GLM Procedure
   Class Level Information
Class         Levels    Values
sex                2    F M

Number of observations    32

Dependent Variable: ltlc
                                  Sum of
Source                    DF     Squares   Mean Square  F Value  Pr > F

Model                      3   0.27230446   0.09076815    13.05  <.0001
Error                     28   0.19478293   0.00695653
Corrected Total           31   0.46708739


R-Square     Coeff Var      Root MSE     ltlc Mean

0.582984      10.85524      0.083406      0.768346
```

Source	DF	Type I SS	Mean Square	F Value	Pr > F
sex	1	0.13626303	0.13626303	19.59	0.0001
height	1	0.13451291	0.13451291	19.34	0.0001
height*sex	1	0.00152852	0.00152852	0.22	0.6429

Source	DF	Type III SS	Mean Square	F Value	Pr > F
sex	1	0.00210426	0.00210426	0.30	0.5867
height	1	0.13597107	0.13597107	19.55	0.0001
height*sex	1	0.00152852	0.00152852	0.22	0.6429

Parameter		Estimate	Standard Error	t Value	Pr > |t|
Intercept		-.2190181620 B	0.35221658	-0.62	0.5391
sex	F	-.2810587157 B	0.51102682	-0.55	0.5867
sex	M	0.0000000000 B	.	.	.
height		0.0060473650 B	0.00201996	2.99	0.0057
height*sex	F	0.0014344422 B	0.00306016	0.47	0.6429
height*sex	M	0.0000000000 B	.	.	.

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

Where are the two lines in the output?

Line for males (the reference group):

$$\text{log10(Lung capacity)} = -0.219 + 0.00605 \times \text{height}$$

Line for females:

$$\begin{aligned}\text{log10(Lung capacity)} &= -0.219 + (-0.281) + (0.00605 + 0.00143) \times \text{height} \\ &= -0.500 + 0.00748 \times \text{height}\end{aligned}$$

Same model, new parametrisation

```
proc glm data=tlc;
class sex;
  model ltlc=sex sex*height / noint solution; run;

...
Source                      DF     Type I SS   Mean Square  F Value   Pr > F

sex                          2   19.02765491    9.51382745  1367.61   <.0001
height*sex                   2    0.13604143    0.06802071     9.78   0.0006

Source                      DF   Type III SS   Mean Square  F Value   Pr > F

sex                          2    0.01537968    0.00768984     1.11   0.3451
height*sex                   2    0.13604143    0.06802071     9.78   0.0006

                                           Standard
Parameter              Estimate               Error    t Value    Pr > |t|

sex          F      -.5000768777         0.37025922      -1.35      0.1876
sex          M      -.2190181620         0.35221658      -0.62      0.5391
height*sex F        0.0074818072         0.00229877       3.25      0.0030
height*sex M        0.0060473650         0.00201996       2.99      0.0057
```

Same model, 2 different parametrisations

```
proc glm data=tlc; class sex;
  model ltlc=sex height sex*height / solution;
run;
```

- One level for the reference group (`sex='M'` and `height=0`)
- A difference between genders (at `height=0`)
- An effect of `height` (slope) for the reference group
- A difference in slopes for the genders

```
proc glm data=tlc; class sex;
  model ltlc=sex sex*height / noint solution;
run;
```

- A level for each group (`sex`) (at `height=0`)
- An effect of `height` (slope) for each group (`sex`)

Here:

No indication of interaction, we omit the term

```
proc glm data=tlc;
  class sex;
  model ltlc=sex height / solution clparm;
run;
```

```
The GLM Procedure
 Dependent Variable: ltlc

                                    Sum of
Source                    DF       Squares   Mean Square  F Value  Pr > F

Model                      2    0.27077594    0.13538797    20.00  <.0001
Error                     29    0.19631145    0.00676936
Corrected Total           31    0.46708739

R-Square     Coeff Var      Root MSE     ltlc Mean

0.579712      10.70821      0.082276      0.768346
```

Source	DF	Type I SS	Mean Square	F Value	Pr > F
sex	1	0.13626303	0.13626303	20.13	0.0001
height	1	0.13451291	0.13451291	19.87	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
sex	1	0.00968023	0.00968023	1.43	0.2415
height	1	0.13451291	0.13451291	19.87	0.0001

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		-.3278068826	B	0.26135206	-1.25	0.2198
sex	F	-.0421012632	B	0.03520676	-1.20	0.2415
sex	M	0.0000000000	B	.	.	.
height		0.0066723630		0.00149683	4.46	0.0001

Parameter		95% Confidence Limits	
Intercept		-.8623318537	0.2067180884
sex	F	-.1141071749	0.0299046484
sex	M	.	.
height		0.0036110089	0.0097337172

Note: The effect of gender has disappeared!!

In **this** example we have seen

- The observed difference in lung capacity between men and women can be explained by height difference

  However, there *may* still be a gender difference (women vs. men), estimated as $-0.0421 \pm 2 \times 0.0352 = (-0.1141, 0.0299)$, corresponding to the interval (0.77, 1.07) for ratios.

  If we would rather see it as men vs. women, we invert the figures to get the confidence interval (0.93,1.30) for ratios, i.e. there may be a 30% increased lung function for men.

It **may also occur**, that

- Apparently identical groups (e.g. blood pressure for men and women) may show up differences when we correct for inhomogeneities between groups
  (e.g. obesity)

We may conclude: It is **important** to remember all relevant covariates.

General statistical tool: **Multiple regression / General linear model**

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 (covariates) is written:

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

Interpretation of regression coefficients β

Model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$

Example Y: blood pressure X_1: age X_2: weight

Consider two subjects:

A has covariate values $(35, 75)$; B has covariate values $(36, 75)$

Expected difference in blood pressure $(B - A)$

$\beta_0 + \beta_1 \cdot 36 + \beta_2 \cdot 75 - [\beta_0 + \beta_1 \cdot 35 + \beta_2 \cdot 75] = \beta_1$

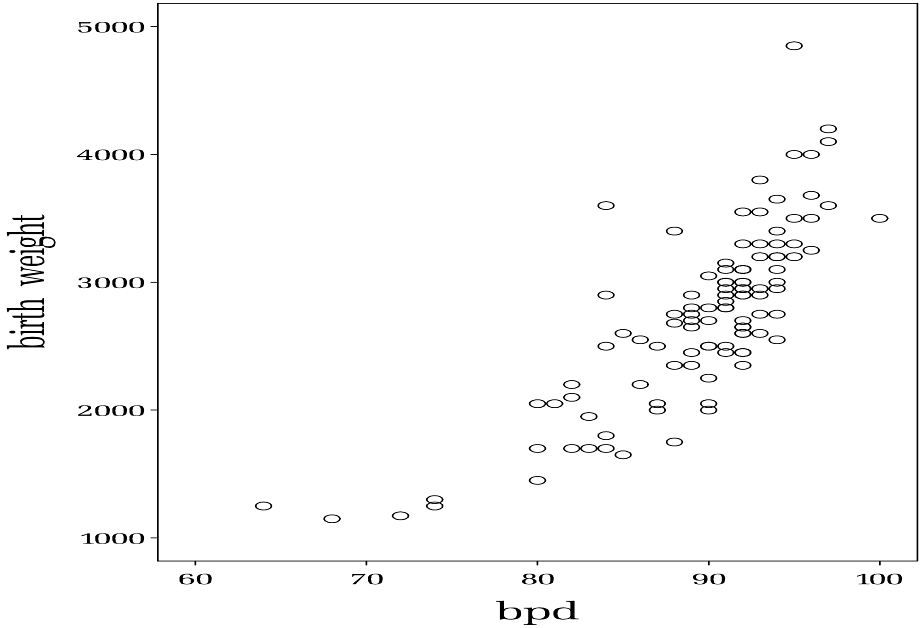
β_1: is the increase in blood pressure when X_1 is increased one unit *and the other predictors are kept fixed*

Note, that the result does not depend on the level of X_1 (here 35). No matter where we start, the effect of a one unit increase is the same. The effect is linear.

Note also, that the result does not depend on the level of X_2 (here 75). The effect of a one unit increase in X_1 is the same for all values of X_2. This can be changed by including an interaction term.

Ultra sound scanning, immediately before birth (Secher et al.)

OBS	WEIGHT	BPD	AD
1	2350	88	92
2	2450	91	98
.	.	.	.
.	.	.	.
106	1173	72	73
107	2900	92	104



```
proc reg data=secher;
model lweight=lbpd lad / clb;
run;
```

```
Dependent Variable: lweight
                              Sum of          Mean
Source               DF      Squares        Square  F Value  Pr > F

Model                 2     14.95054       7.47527   314.93  <.0001
Error               104      2.46861       0.02374
Corrected Total     106     17.41915

Root MSE             0.15407    R-Square     0.8583
Dependent Mean      11.36775    Adj R-Sq     0.8556
Coeff Var            1.35530

                        Parameter Estimates

                     Parameter       Standard
Variable     DF       Estimate          Error    t Value    Pr > |t|

Intercept     1       -8.45636        0.95457      -8.86      <.0001
lbpd          1        1.55194        0.22945       6.76      <.0001
lad           1        1.46666        0.14669      10.00      <.0001

 Variable     DF       95% Confidence Limits

Intercept     1      -10.34931       -6.56341
lbpd          1        1.09694        2.00695
lad           1        1.17577        1.75756
```

Interpretation of regression parameters

β_j: The effect of the j'th explanatory variable, **corrected** for the effect of the other explanatory variables –
i.e. when these are **kept fixed**

E.g: The effect of $\log_{10}(\texttt{bpd})$ corrected for the effect of $\log_{10}(\texttt{ad})$ is found to be $\hat{\beta}_1 = 1.552$

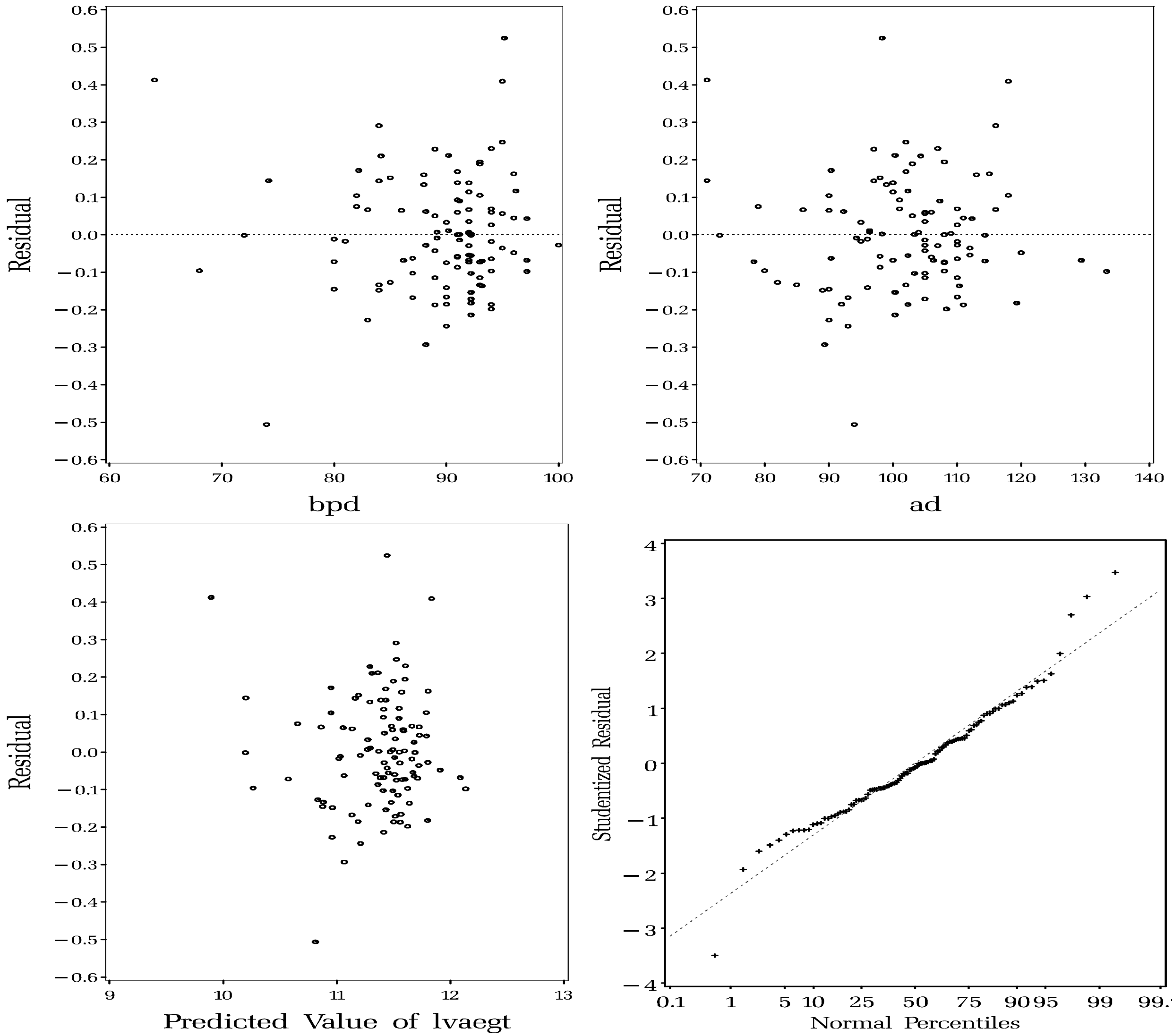
but in the marginal model **without** correction for $\log(\texttt{ad})$, we get: $\hat{\beta}_1^*$=3.332

The difference can be very **important**!

Group variables

Group variables can be directly handled in PROC GLM by choosing the group variable as a CLASS variable.

Residual plots



Ex. O'Neill et.al. (1983):

Lung function for 25 patients with cystic fibrosis.

Table 12.11 Data for 25 patients with cystic fibrosis (O'Neill *et al.*, 1983)

Sub	Age	Sex	Height	Weight	BMP	FEV_1	RV	FRC	TLC	PEmax
1	7	0	109	13.1	68	32	258	183	137	95
2	7	1	112	12.9	65	19	449	245	134	85
3	8	0	124	14.1	64	22	441	268	147	100
4	8	1	125	16.2	67	41	234	146	124	85
5	8	0	127	21.5	93	52	202	131	104	95
6	9	0	130	17.5	68	44	308	155	118	80
7	11	1	139	30.7	89	28	305	179	119	65
8	12	1	150	28.4	69	18	369	198	103	110
9	12	0	146	25.1	67	24	312	194	128	70
10	13	1	155	31.5	68	23	413	225	136	95
11	13	0	156	39.9	89	39	206	142	95	110
12	14	1	153	42.1	90	26	253	191	121	90
13	14	0	160	45.6	93	45	174	139	108	100
14	15	1	158	51.2	93	45	158	124	90	80
15	16	1	160	35.9	66	31	302	133	101	134
16	17	1	153	34.8	70	29	204	118	120	134
17	17	0	174	44.7	70	49	187	104	103	165
18	17	1	176	60.1	92	29	188	129	130	120
19	17	0	171	42.6	69	38	172	130	103	130
20	19	1	156	37.2	72	21	216	119	81	85
21	19	0	174	54.6	86	37	184	118	101	85
22	20	0	178	64.0	86	34	225	148	135	160
23	23	0	180	73.8	97	57	171	108	98	165
24	23	0	175	51.1	71	33	224	131	113	95
25	23	0	179	71.5	95	52	225	127	101	195

Which explanatory variables have a *marginal* effect on the outcome PE_{max}?

Table 12.12 Results of separately regressing PEmax on each explanatory variable

Explanatory variable	Regression coefficient	Standard error	t	P
Age	4.055	1.088	3.73	0.0011
Sex	−19.045	13.176	−1.45	0.16
Height	0.932	0.260	3.59	0.0016
Weight	1.187	0.301	3.94	0.0006
BMP	0.639	0.565	1.13	0.27
FEV_1	1.354	0.555	2.44	0.023
RV	−0.123	0.077	−1.59	0.12
FRC	−0.319	0.145	−2.20	0.038
TLC	−0.358	0.404	−0.89	0.38

Some effects may be caused by confounding.

Model with all covariates

```
proc reg data=pemax;
  model pemax=age sex height weight bmp fev1 rv frc tlc;
run;
```

```
The REG Procedure
Dependent Variable: pemax
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	176.05821	225.89116	0.78	0.4479
age	1	-2.54196	4.80170	-0.53	0.6043
sex	1	-3.73678	15.45982	-0.24	0.8123
height	1	-0.44625	0.90335	-0.49	0.6285
weight	1	2.99282	2.00796	1.49	0.1568
bmp	1	-1.74494	1.15524	-1.51	0.1517
fev1	1	1.08070	1.08095	1.00	0.3333
rv	1	0.19697	0.19621	1.00	0.3314
frc	1	-0.30843	0.49239	-0.63	0.5405
tlc	1	0.18860	0.49974	0.38	0.7112

Correlated covariates

Univariate analysis showed strong effects

Multiple analysis showed no effects

How can that be?

When we include many correlated covariates in the model, the power to detect effects will decrease. For instance, there will be limited information in the data about the effect of `height` for fixed level of `weight`, because when `height` is increased `weight` tends to increase also. Highly correlated covariates should be avoided.

It may be possible to regain power by excluding insignificant covariates.

Automatic model selection

- **Forward selection:** Start with no covariates. In every step, add the most significant variable

```
proc reg data=pemax;
  model pemax=age sex height weight bmp fev1 rv frc tlc
              / selection=forward;
run;
```

Final model: `weight bmp fev1`

- **Backward elimination**
  Start with all covariates. At each step, omit the least significant

```
proc reg data=pemax;
  model pemax=age sex height weight bmp fev1 rv frc tlc
              / selection=backward;
run;
```

Final model: `weight bmp fev1`

But:

If `weight` had been transformed with the logarithm from the start, we would have had the final model `age fev1`

Selection procedures

- `backward`
- `forward`
- ...

A 'best' method has not been identified, but backward elimination is generally recommended over forward selection.

WARNING: The output from the selected model does not take the model selection uncertainty into account. The output (regression coefficients and p-values) is identical to what would have been obtained had we fitted the final model with out doing any model selection. The importance of selected covariates is over-estimated.

Exercise: General linear models

We take another look at Juul's data.

1. Get the data into SAS using a libname statement.
2. Create a new data set including only individuals above 25 years.
3. Use PROC GPLOT to plot the relationship between age and $\sqrt{\text{SIGF-I}}$. Make separate regression lines for men and women.
4. Do a regression analysis to explore whether the slopes (age - $\sqrt{\text{SIGF-I}}$) are the same in men and women. Give an estimate for the difference in slopes, with 95% confidence interval.
5. Expand the regression model by including height. Delete in-significant covariates.