Part 3
Lecture 1 Confounding

1

Who I am…

Pascal Tyrrell, PhD

Associate Professor

Department of Medical Imaging, Faculty of Medicine

Institute of Medical Science, Faculty of Medicine

Department of Statistical Sciences, Faculty of Arts and Science

COMPARISON OF MEAN FVC

Among twenty persons who recently joined an exercise gymnasium, ten had no experience carrying out any exercises whereas the other ten had some experience doing exercises at home.

The forced vital capacity (FVC) was measured on each of the twenty individuals. The mean FVC was compared between the two groups.

CREATION OF A TEMPORARY SAS DATASET

The SAS code in the next two slides creates temporary SAS datasets PAIRED and UNPAIRED.

Each of the DATA , INPUT, DATALINES and RUN ends with a semi-colon **(;).** When the computer is shut off the datasets no longer exists.

The INPUT statement is used to define four variables: ID, EXER, HGT and FVC. The character variable ID variable is followed by a dollar sign ($) in the INPUT statement.

The other three variables are numeric variables. In the second and third lines the character variable EXERCISE is created.

Example 1 COMPARING MEAN FVC IN TWO EXERCISE GROUPS DATASET CONSISTS OF 20 DIFFERENT PATIENTS

```
DATA UNPAIRED ; INPUT ID $  EXER  HGT  FVC  @@ ;
EXERCISE = "YES" ;
IF EXER=0 THEN EXERCISE=" NO"; DATALINES ;
 1 0 120 1.00   2 0 130 1.40   3 0 135 2.04
 4 0 145 2.00   5 0 140 2.70   6 0 150 2.00
 7 0 155 3.25   8 0 160 2.50   9 0 170 3.20
10 0 190 4.50
11 1 140 1.92  12 1 150 3.30  13 1 154 3.00
14 1 143 2.82  15 1 164 3.55  16 1 170 4.30
17 1 174 3.68  18 1 172 2.78  19 1 174 4.20
20 1 183 4.28
RUN ;
```

Example 2 COMPARING MEAN FVC IN TWO EXERCISE GROUPS THE DATA CONSISTS OF N=10 PAIRS OF PATIENTS

```
DATA PAIRED ; INPUT ID $  EXER  HGT  FVC @@ ;
EXERCISE = "YES" ;
IF EXER=0 THEN EXERCISE =" NO"; DATALINES ;
 1 0 120 1.00   2 0 130 1.40   3 0 135 2.04
 4 0 145 2.00   5 0 140 2.70   6 0 150 2.00
 7 0 155 3.25   8 0 160 2.50   9 0 170 3.20
10 0 190 4.50
 1 1 140 1.92   2 1 150 3.30   3 1 154 3.00
 4 1 143 2.82   5 1 164 3.55   6 1 170 4.30
 7 1 174 3.68   8 1 172 2.78   9 1 174 4.20
10 1 183 4.28
RUN ;
```

ID paired values go from 1 to 10

COMPARING PAIRED AND UNPAIRED MEANS WITH SIMPLIFIED DATASETS

```
DATA  PAIRS ;
INPUT FVC1 FVC2 @@ ;
DFVC = FVC1 - FVC2 ;
DATALINES ;
1.00  1.92   1.40  3.30   2.04  3.00  2.00  2.82
2.70  3.55   2.00  4.30   3.25  3.68  2.50  2.78
3.20  4.20   4.50  4.28
RUN ;

PROC MEANS DATA = PAIRS N MEAN VAR STDDEV ;
VAR FVC1 FVC2 DFVC ;
RUN ;
```

COMPARING PAIRED AND UNPAIRED MEANS WITH SIMPLIFIED DATASETS

The MEANS Procedure

Variable	N	Mean	Variance	Std Dev
FVC1	10	2.4590000	1.0274767	1.0136452
FVC2	10	3.3830000	0.6004011	0.7748555
DFVC	10	-0.9240000	0.5374267	0.7330939

```
PROC TTEST  DATA = PAIRS  ; PAIRED FVC1 * FVC2 ; RUN ;

PROC CORR   DATA = PAIRS  ; VAR    FVC1   FVC2 ; RUN ;

/*****************************************************************
* Var(X2 - X1) = Var(X2)  + Var(X1) - 2 x Std(X2) x  Std(X1) x  R    *
*    0.537     =  1.028   +  0.600  - 2 x 1.014   x  0.775   x  0.694 *
*****************************************************************/
```

The TTEST Procedure

Difference: FVC1 - FVC2

N	Mean	Std Dev	Std Err	Minimum	Maximum
10	-0.9240	0.7331	0.2318	-2.3000	0.2200

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
-0.9240	-1.4484	-0.3996	0.7331	0.5042	1.3383

DF	t Value	Pr > |t|
9	-3.99	0.0032

The CORR Procedure

2 Variables:	FVC1 FVC2

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
FVC1	10	2.45900	1.01365	24.59000	1.00000	4.50000
FVC2	10	3.38300	0.77486	33.83000	1.92000	4.30000

Pearson Correlation Coefficients, N = 10
Prob > |r| under H0: Rho=0

	FVC1	FVC2
FVC1	1.00000	0.69418 0.0259
FVC2	0.69418 0.0259	1.00000

The GLM Procedure

Dependent Variable: FVC2

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2.60389189	2.60389189	7.44	0.0259
Error	8	2.79971811	0.34996476		
Corrected Total	9	5.40361000			

R-Square	Coeff Var	Root MSE	FVC2 Mean
0.481880	17.48679	0.591578	3.383000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
FVC1	1	2.60389189	2.60389189	7.44	0.0259

Source	DF	Type III SS	Mean Square	F Value	Pr > F
FVC1	1	2.60389189	2.60389189	7.44	0.0259

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	2.078143467	0.51364757	4.05	0.0037
FVC1	0.530645194	0.19453821	2.73	0.0259

```
PROC GLM DATA = PAIRS ;
MODEL  FVC2 = FVC1 ;
RUN ;
```

FITTING A STRAIGHT LINE BETWEEN FVC1 AND FVC2

MY NOTE:
Slope of line FVC2 = B0 + B1 x FVC1 is 0.531

In the special case of simple regression:

$$r = \sqrt{R^2}$$
$$= \sqrt{0.48}$$
$$= 0.69$$

TWO PROCEDURES FOR COMPARING PAIRED MEANS IN ANALYSIS OF RELATIONSHIP BETWEEN EXERCISE AND FVC

```
PROC TTEST DATA = PAIRS ;        **<< Done ;
VAR DFVC ;
RUN ;

PROC GLM DATA = PAIRED ;
CLASS EXERCISE ID ;
MODEL FVC =  EXERCISE  ID / SS3 ;
LSMEANS EXERCISE / TDIFF PDIFF STDERR CL ;
RUN ;
```

The GLM Procedure

Dependent Variable: FVC

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	16.50136000	1.65013600	6.14	0.0058
Error	9	2.41842000	0.26871333		
Corrected Total	19	18.91978000			

R-Square	Coeff Var	Root MSE	FVC Mean
0.872175	17.74651	0.518376	2.921000

Source	DF	Type III SS	Mean Square	F Value	Pr > F
EXERCISE	1	4.26888000	4.26888000	15.89	0.0032
ID	9	12.23248000	1.35916444	5.06	0.0121

The GLM Procedure
Least Squares Means

EXERCISE	FVC LSMEAN	Standard Error	H0:LSMEAN=0	H0:LSMean1=LSMean2	
			Pr > |t|	t Value	Pr > |t|
NO	2.45900000	0.16392478	<.0001	-3.99	0.0032
YES	3.38300000	0.16392478	<.0001		

EXERCISE	FVC LSMEAN	95% Confidence Limits	
NO	2.459000	2.088176	2.829824
YES	3.383000	3.012176	3.753824

Least Squares Means for Effect EXERCISE				
i	j	Difference Between Means	95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-0.924000	-1.448424	-0.399576

STUDENT T TEST COMPARISON OF UNPAIRED MEANS

The TTEST Procedure

Variable: FVC

EXERCISE	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
NO		10	2.4590	1.0136	0.3205	1.0000	4.5000
YES		10	3.3830	0.7749	0.2450	1.9200	4.3000
Diff (1-2)	Pooled		-0.9240	0.9022	0.4035		
Diff (1-2)	Satterthwaite		-0.9240		0.4035		

EXERCISE	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
NO		2.4590	1.7339	3.1841	1.0136	0.6972	1.8505
YES		3.3830	2.8287	3.9373	0.7749	0.5330	1.4146
Diff (1-2)	Pooled	-0.9240	-1.7717	-0.0763	0.9022	0.6817	1.3342
Diff (1-2)	Satterthwaite	-0.9240	-1.7759	-0.0721			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	18	-2.29	0.0343
Satterthwaite	Unequal	16.841	-2.29	0.0352

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	9	9	1.71	0.4358

```
PROC TTEST  DATA = UNPAIRED  ;
CLASS EXERCISE ;
VAR FVC ;
RUN ;
```

Is the variance similar in the two groups?

F Ratio = $(1.014^2)/(0.775^2)$

= 1.028/0.600

= 1.71

<<< Pooled variance method

TWO PROCEDURES FOR COMPARING UNPAIRED MEANS

```
PROC TTEST  DATA = UNPAIRED  CL = NONE ;   **<< Done;
CLASS EXERCISE ;
VAR FVC  ;
RUN ;

PROC  GLM  DATA = UNPAIRED ;
CLASS EXERCISE ;
MODEL FVC  =  EXERCISE  /  SS3 ;
LSMEANS EXERCISE / TDIFF PDIFF STDERR CL ;
RUN ;
```

TWO PROCEDURES FOR COMPARING UNPAIRED MEANS

The GLM Procedure

Dependent Variable: FVC

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4.26888000	4.26888000	5.24	0.0343
Error	18	14.65090000	0.81393889		
Corrected Total	19	18.91978000			

R-Square	Coeff Var	Root MSE	FVC Mean
0.225631	30.88619	0.902186	2.921000

Source	DF	Type III SS	Mean Square	F Value	Pr > F
EXERCISE	1	4.26888000	4.26888000	5.24	0.0343

The GLM Procedure
Least Squares Means

EXERCISE	FVC LSMEAN	Standard Error	H0:LSMEAN=0	H0:LSMean1=LSMean2	
			Pr > |t|	t Value	Pr > |t|
NO	2.45900000	0.28529614	<.0001	-2.29	0.0343
YES	3.38300000	0.28529614	<.0001		

EXERCISE	FVC LSMEAN	95% Confidence Limits	
NO	2.459000	1.859615	3.058385
YES	3.383000	2.783615	3.982385

Least Squares Means for Effect EXERCISE				
i	j	Difference Between Means	95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-0.924000	-1.771658	-0.076342

PREVIOUS DATASET ALSO HAD THE HEIGHT OF EACH OF THE TWENTY SUBJECTS

```
PROC SORT DATA = UNPAIRED ; BY EXER ; RUN ;
PROC TTEST DATA = UNPAIRED ;
CLASS EXERCISE ;
VAR HGT ; RUN ;

PROC GLM  DATA = UNPAIRED ;
BY EXER ;
MODEL FVC = HGT ; RUN ;

PROC GLM DATA = UNPAIRED ;
CLASS EXERCISE ;
MODEL FVC = EXERCISE  HGT  EXERCISE*HGT /SS3 ; RUN ;
```

The TTEST Procedure

Variable: HGT

EXERCISE	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
NO		10	149.5	20.4736	6.4743	120.0	190.0
YES		10	162.4	14.7136	4.6528	140.0	183.0
Diff (1-2)	Pooled		-12.9000	17.8277	7.9728		
Diff (1-2)	Satterthwaite		-12.9000		7.9728		

EXERCISE	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
NO		149.5	134.9	164.1	20.4736	14.0824	37.3767
YES		162.4	151.9	172.9	14.7136	10.1205	26.8612
Diff (1-2)	Pooled	-12.9000	-29.6502	3.8502	17.8277	13.4709	26.3641
Diff (1-2)	Satterthwaite	-12.9000	-29.7731	3.9731			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	18	-1.62	0.1231
Satterthwaite	Unequal	16.339	-1.62	0.1248

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	9	9	1.94	0.3392

Is mean height different between those who exercise and those who don't?

Not significantly yet with a difference of 12.9 cm…

The GLM Procedure

Dependent Variable: FVC

EXER=0

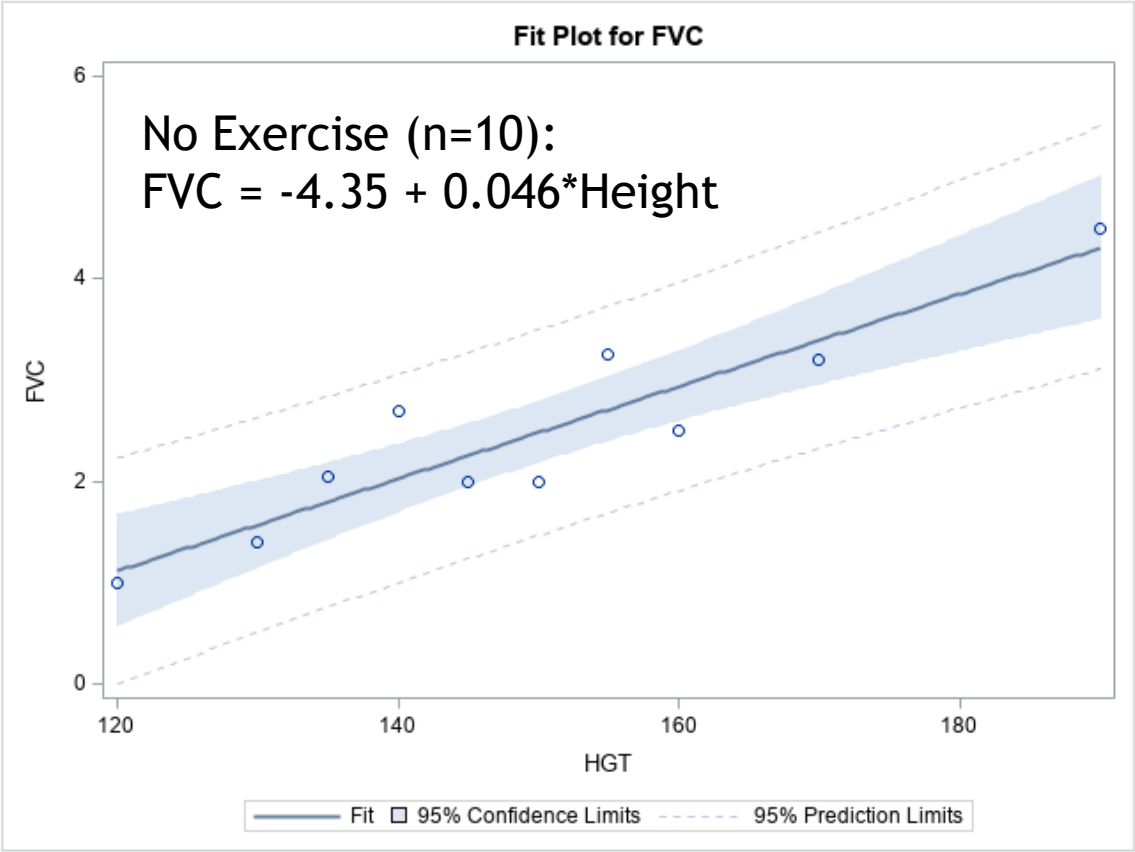
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	7.83700014	7.83700014	44.46	0.0002
Error	8	1.41028986	0.17628623		
Corrected Total	9	9.24729000			

R-Square	Coeff Var	Root MSE	FVC Mean
0.847492	17.07461	0.419865	2.459000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
HGT	1	7.83700014	7.83700014	44.46	0.0002

Source	DF	Type III SS	Mean Square	F Value	Pr > F
HGT	1	7.83700014	7.83700014	44.46	0.0002

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-4.354990060	1.03055318	-4.23	0.0029
HGT	0.045578529	0.00683588	6.67	0.0002



MYNOTE :: 7.84/9.25 = 0.85

The GLM Procedure

Dependent Variable: FVC

EXER=1

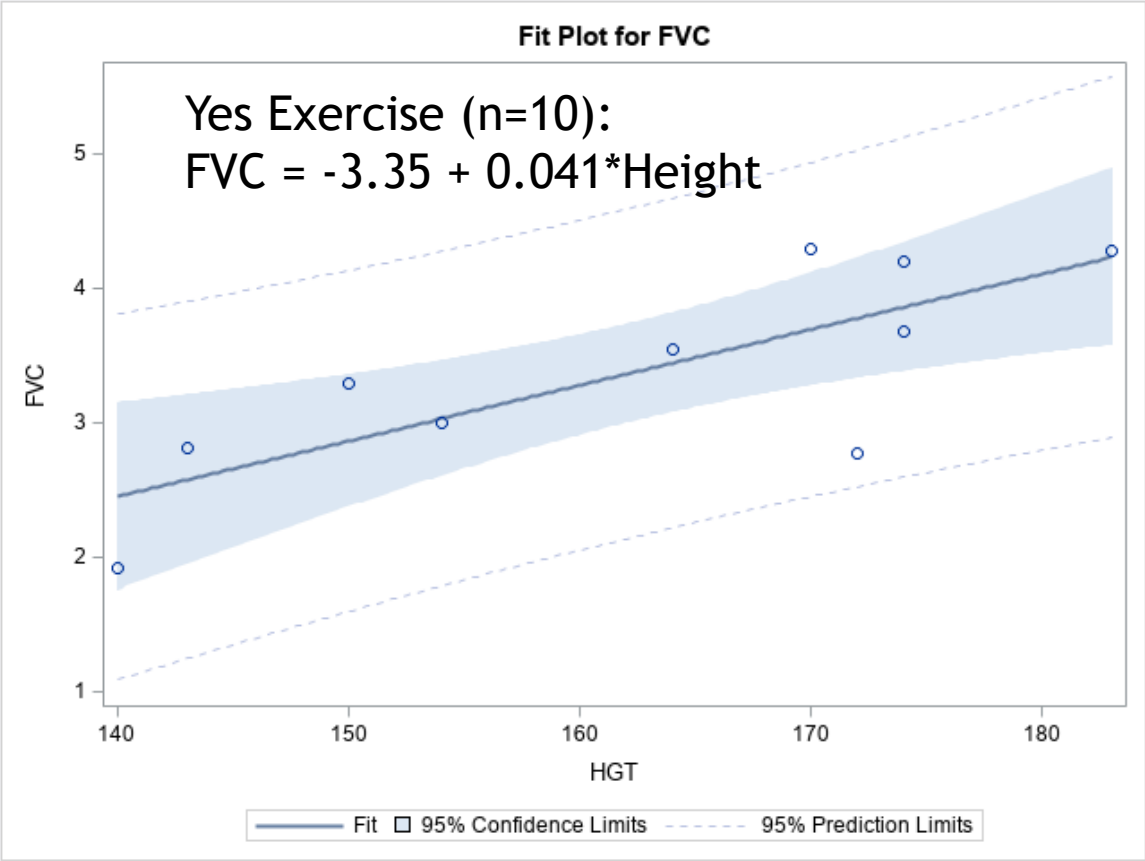
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3.34977466	3.34977466	13.05	0.0069
Error	8	2.05383534	0.25672942		
Corrected Total	9	5.40361000			

R-Square	Coeff Var	Root MSE	FVC Mean
0.619914	14.97738	0.506685	3.383000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
HGT	1	3.34977466	3.34977466	13.05	0.0069

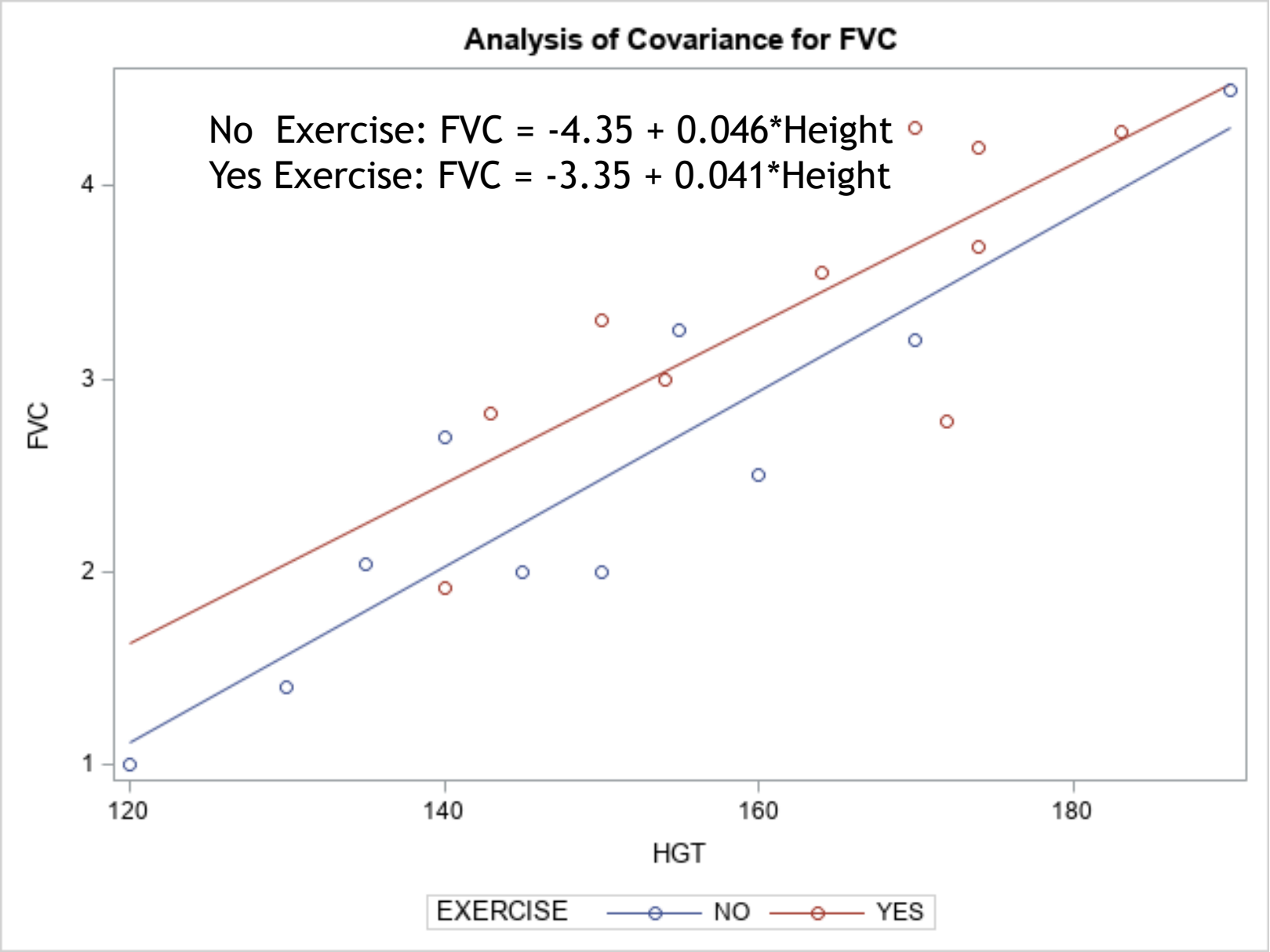
Source	DF	Type III SS	Mean Square	F Value	Pr > F
HGT	1	3.34977466	3.34977466	13.05	0.0069

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-3.350715459	1.87104009	-1.79	0.1111
HGT	0.041463765	0.01147886	3.61	0.0069



MYNOTE :: 3.35/5.40 = 0.62

Let’s put them together!



The GLM Procedure

Dependent Variable: FVC

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	15.45565480	5.15188493	23.80	<.0001
Error	16	3.46412520	0.21650783		
Corrected Total	19	18.91978000			

R-Square	Coeff Var	Root MSE	FVC Mean
0.816905	15.92961	0.465304	2.921000

Source	DF	Type III SS	Mean Square	F Value	Pr > F
EXERCISE	1	0.05129883	0.05129883	0.24	0.6330
HGT	1	9.73427630	9.73427630	44.96	<.0001
HGT*EXERCISE	1	0.02175368	0.02175368	0.10	0.7554

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	-3.350715459	B	1.71823307	-1.95	0.0689
EXERCISE NO	-1.004274601	B	2.06317244	-0.49	0.6330
EXERCISE YES	0.000000000	B	.	.	.
HGT	0.041463765	B	0.01054139	3.93	0.0012
HGT*EXERCISE NO	0.004114764	B	0.01298121	0.32	0.7554
HGT*EXERCISE YES	0.000000000	B	.	.	.

TABLE 3 ARE THESE LINES PARALLEL (NO INTERACTION) ??

```
PROC GLM DATA = UNPAIRED ;
CLASS EXERCISE ;
MODEL FVC = EXERCISE HGT EXERCISE*HGT /
SOLUTION SS3 ;
RUN ;
```

<<< Interaction term is not significant

Let's have a look at the equation for this interaction model…

```
PREDICTED FVC = FVCp      HEIGHT = HGT

EXER = 0  Exercise No
EXER = 1  Exercise Yes

FVCp =    B0          + B1 x EXER
     +    B2 x HGT    + B3 x EXER x HGT

  = -3.3507           +   1.0043 x EXER
  +  0.0415 x HGT     +  -0.0041 x EXER x HGT
```

PREDICTED FVC = FVCp HEIGHT = HGT

EXER = 0 Exercise No EXER = 1 Exercise Yes

EXER=0 FVCp = -3.3507 + 1.0043 x 0

+ 0.0415 x HGT - 0.0041 x 0 x HGT

= -3.3507 + 0.0415 x HGT

EXER=1 FVCp = -3.3507 - 1.0043 x 1

+ 0.0415 x HGT - 0.0041 x 1 x HGT

= -4.3550 + 0.0374 x HGT

MY NOTE: 0.0456 - 0.0415 = 0.0041 . Interaction term B3 is difference between slopes !!!

Now back to the equation without the interaction term…

Dependent Variable: FVC

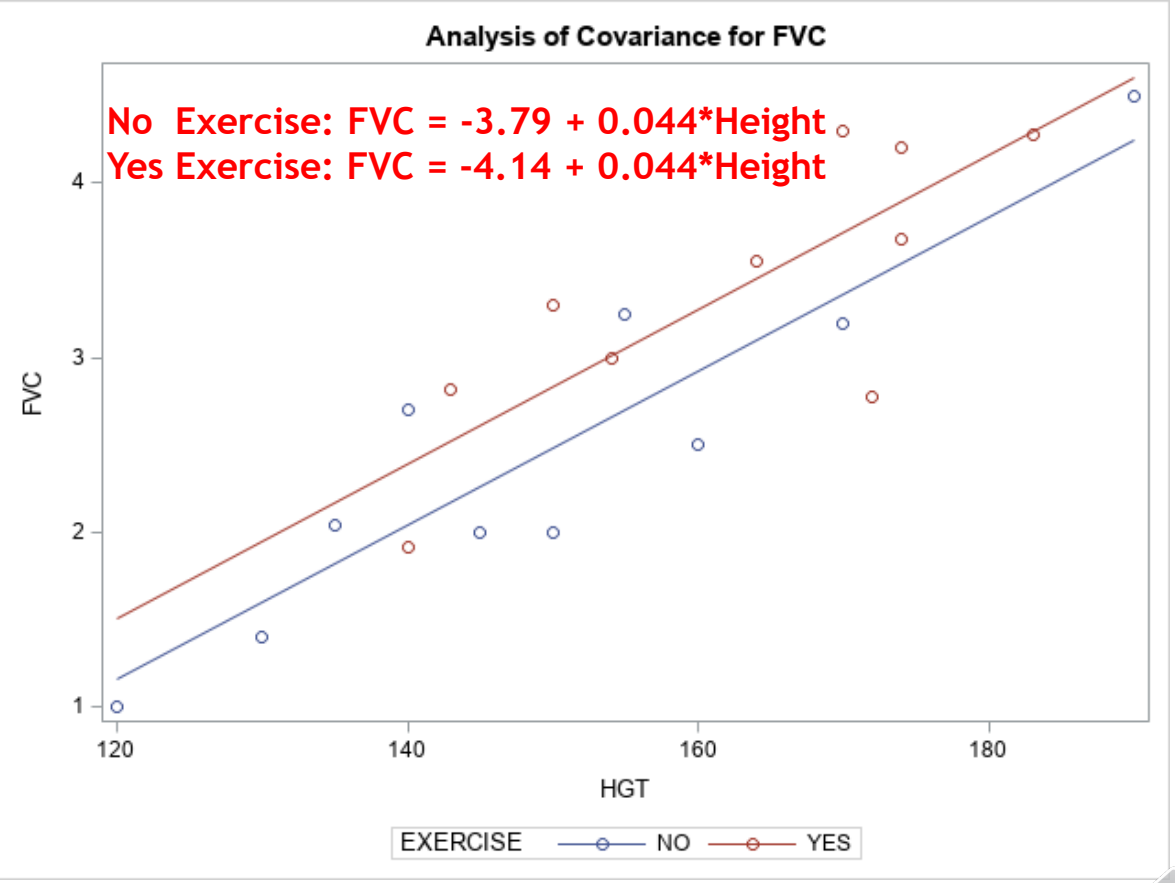
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	15.43390111	7.71695056	37.63	<.0001
Error	17	3.48587889	0.20505170		
Corrected Total	19	18.91978000			

R-Square	Coeff Var	Root MSE	FVC Mean
0.815755	15.50244	0.452826	2.921000

Source	DF	Type III SS	Mean Square	F Value	Pr > F
EXERCISE	1	0.54737618	0.54737618	2.67	0.1207
HGT	1	11.16502111	11.16502111	54.45	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	-3.791367530	B	0.98275543	-3.86	0.0013
EXERCISE NO	-0.354114895	B	0.21673694	-1.63	0.1207
EXERCISE YES	0.000000000	B	.	.	.
HGT	0.044177140		0.00598687	7.38	<.0001

```
PROC GLM DATA = UNPAIRED ;
CLASS EXERCISE ;
MODEL FVC = EXERCISE HGT / SOLUTION SS3 ;
RUN ;
```



25

End of Lecture 1

Next up in Part 3 Lecture 2: Interaction