
REGRESSION ANALYSIS: INTERACTIONS AND SUB-GROUPS AND OTHER TOPICS

Clinical Pharmacology Conference

Tatsuki Koyama, PhD
Department of Biostatistics
tatsuki.koyama@vumc.org

April 6, 2021

Disclaimer: The following example is fictitious, and many of the analytic strategies are for demonstrative purposes only, and do not reflect good analytic practice.

Example: CEASAR

Comparative Effectiveness Analysis of Surgery and Radiation (CEASAR) is an observational study that recruited men who were diagnosed with prostate cancer from 2011 to 2012.

- CEASAR enrolled more than 3,000 men.
- The primary outcome variable is based on a patient-reported quality-of-life, whose score ranges from 0 to 100.
- The majority of patients underwent surgery (radical prostatectomy), and other treatment options included radiation (EBRT) and active surveillance.
- For this example, the data have been altered and truncated ($n = 100$).
 - The main outcome variable is postQoL: post-treatment Quality of Life score.
 - The baseline score is preQoL.

preQoL	postQoL	Treatment	PSA	Age	Hypertension
Min. :16	Min. : 3	Radiation:37	Min. : 0.1	Min. :42	No :40
1st Qu.:36	1st Qu.:42	Surgery :63	1st Qu.: 8.2	1st Qu.:58	Yes:60
Median :57	Median :57		Median :11.1	Median :66	
Mean :56	Mean :54		Mean :10.7	Mean :64	
3rd Qu.:76	3rd Qu.:69		3rd Qu.:13.3	3rd Qu.:71	
Max. :95	Max. :86		Max. :21.4	Max. :79	

1 Interaction vs Subgroup

1.1 Simple case

Question: Is the baseline PSA associated with postQoL?

Linear Regression Model

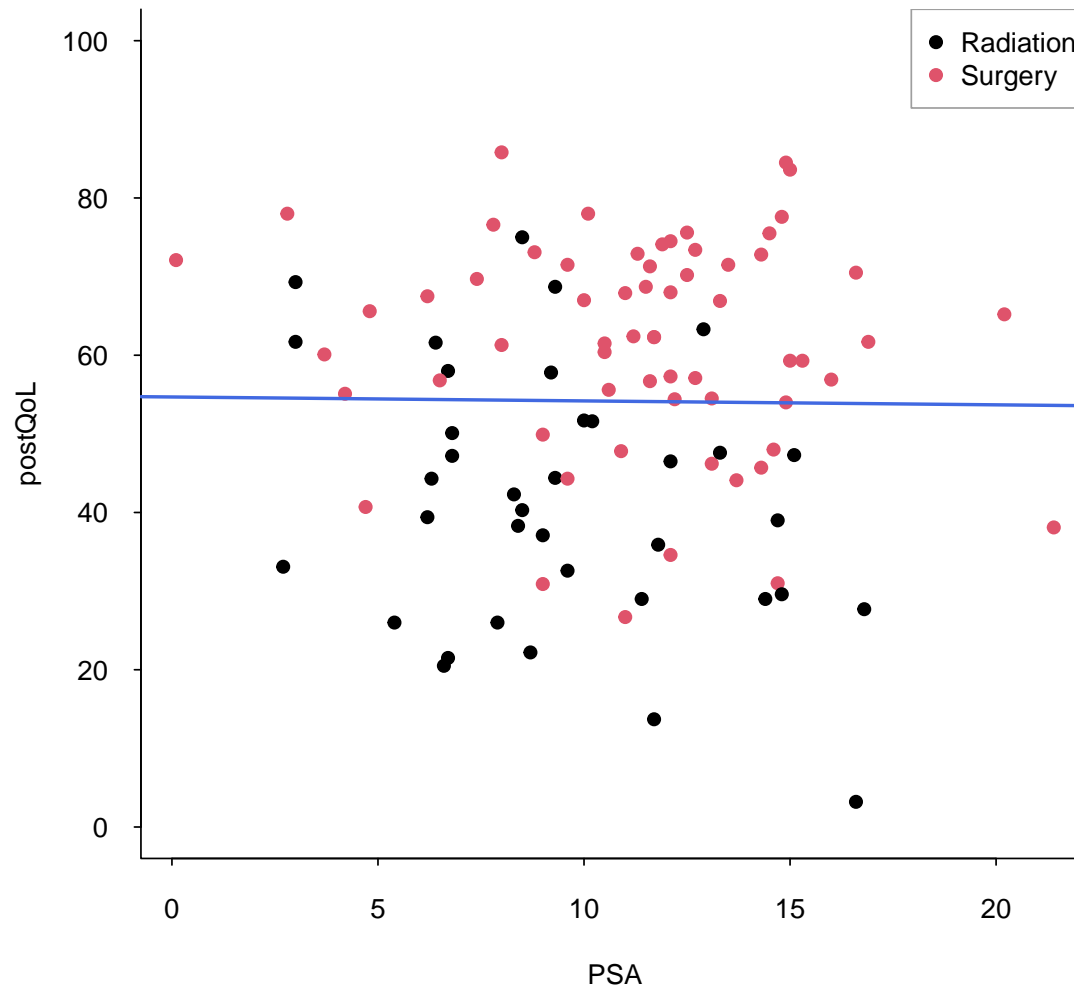
```
ols(formula = postQoL ~ PSA, data = ds)
```

		Model Likelihood		Discrimination	
		Ratio Test		Indexes	
Obs	100	LR chi2	0.01	R2	0.000
sigma	17.7190	d.f.	1	R2 adj	-0.010
d.f.	98	Pr(> chi2)	0.9170	g	0.207

Residuals

	Min	1Q	Median	3Q	Max
	-50.713	-12.444	2.981	14.470	31.484

	Coef	S.E.	t	Pr(> t)
Intercept	54.6911	5.1634	10.59	<0.0001
PSA	-0.0468	0.4543	-0.10	0.9181



Question: Is the baseline PSA associated with the postQoL differently in the Surgery and Radiation groups?

```
##### Surgery subgroup ##
(model1.Sur <- ols(postQoL ~ PSA, data = subset(ds, Treatment == "Surgery")))
```

Linear Regression Model

```
ols(formula = postQoL ~ PSA, data = subset(ds, Treatment == "Surgery"))
```

		Model Likelihood		Discrimination	
		Ratio Test		Indexes	
Obs	63	LR chi2	0.62	R2	0.010
sigma	13.6662	d.f.	1	R2 adj	-0.006
d.f.	61	Pr(> chi2)	0.4319	g	1.476

Residuals

	Min	1Q	Median	3Q	Max
	-35.1311	-6.5948	0.7103	10.5206	24.0136

	Coef	S.E.	t	Pr(> t)
Intercept	65.6238	5.3556	12.25	<0.0001
PSA	-0.3448	0.4447	-0.78	0.4412

```
##### Radiation subgroup ##
(model1.Rad <- ols(postQoL ~ PSA, data = subset(ds, Treatment == "Radiation")))
```

Linear Regression Model

```
ols(formula = postQoL ~ PSA, data = subset(ds, Treatment == "Radiation"))
```

		Model Likelihood		Discrimination	
		Ratio Test		Indexes	
Obs	37	LR chi2	3.77	R2	0.097
sigma	15.8094	d.f.	1	R2 adj	0.071
d.f.	35	Pr(> chi2)	0.0522	g	5.834

Residuals

Min	1Q	Median	3Q	Max
-28.268	-8.590	-1.473	11.344	32.282

	Coef	S.E.	t	Pr(> t)
Intercept	54.5228	7.2448	7.53	<0.0001
PSA	-1.3888	0.7167	-1.94	0.0608

```
##### Interaction model ##
(model1.Int <- ols(postQoL ~ PSA * Treatment, data = ds))
```

Linear Regression Model

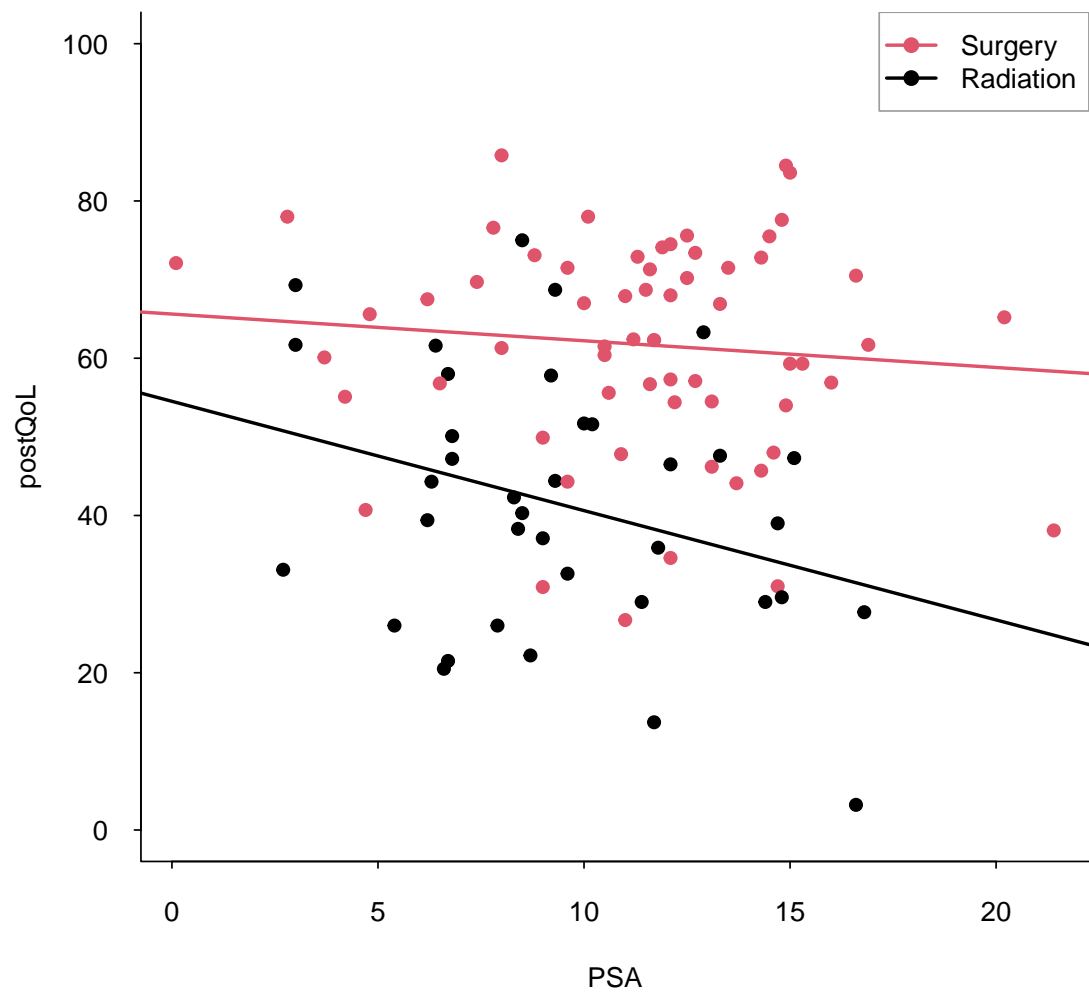
```
ols(formula = postQoL ~ PSA * Treatment, data = ds)
```

	Model Likelihood	Discrimination
	Ratio Test	Indexes
Obs	100	LR chi2 42.39
sigma	14.4844	d.f. 3
d.f.	96	Pr(> chi2) 0.0000
		R2 0.345
		R2 adj 0.325
		g 10.914

Residuals

Min	1Q	Median	3Q	Max
-35.1311	-6.7096	0.6741	11.0923	32.2823

	Coef	S.E.	t	Pr(> t)
Intercept	54.5228	6.6376	8.21	<0.0001
PSA	-1.3888	0.6567	-2.11	0.0370
Treatment=Surgery	11.1010	8.7337	1.27	0.2068
PSA * Treatment=Surgery	1.0440	0.8083	1.29	0.1996



Coefficients:

	Surgery Subgroup		Radiation Subgroup		Interaction	
	Coefficient	<i>p</i>	Coefficient	<i>p</i>	Coefficient	<i>p</i>
Intercept	65.62	0.000	54.52	0.000	54.52	0.000
PSA	-0.34	0.44	-1.39	0.061	-1.39	0.037
Surgery	-	-	-	-	11.10	0.21
PSA × Surgery	-	-	-	-	1.04	0.20

The interaction model:

$$Y = \beta_0 + \beta_p X_p + \beta_s X_s + \beta_{ps} X_p X_s$$

where

$$X_s = \begin{cases} 1 & \text{if Surgery} \\ 0 & \text{if Radiation,} \end{cases}$$

and X_p is PSA value.

Then for Radiation group, we have

$$Y = \beta_0 + \beta_p X_p$$

and for Surgery group, we have

$$\begin{aligned} Y &= \beta_0 + \beta_p X_p + \beta_s + \beta_{ps} X_p \\ &= (\beta_0 + \beta_s) + (\beta_p + \beta_{ps}) X_p \end{aligned}$$

Thus, in this case, coefficient estimates in the interaction model and those in each sub-group model have simple algebraic relationship. But degrees of freedom and the standard error estimates are different.

One additional and very important advantage of the interaction model is its ability to formally test for differences of PSA effect between Treatment groups.

Question: Is PSA effect the same in Surgery and Radiation groups? (Are the slopes different?)

$$H_0 : \beta_{ps} = 0$$

$$H_1 : \beta_{ps} \neq 0$$

```
confint(m1I)
```

```
                2.5 % 97.5 %  
(Intercept)    41.35 67.698  
PSA             -2.69 -0.085  
TreatmentSurgery -6.24 28.437  
PSA:TreatmentSurgery -0.56  2.649
```

1.2 With other covariates

What if we would like to examine the association between PSA and postQoL within each Treatment accounting for preQoL, age, and Hypertension.

```
## Surgery subgroup ##
```

```
## ( model1.Sur <- ols( postQoL ~ PSA, data=subset(ds, Treatment=='Surgery') ) )
(model2.Sur <- ols(postQoL ~ PSA + preQoL + Age + Hypertension, data = subset(ds, Treatment == "Surgery")))
```

Linear Regression Model

```
ols(formula = postQoL ~ PSA + preQoL + Age + Hypertension, data = subset(ds,
  Treatment == "Surgery"))
```

		Model Likelihood	Discrimination		
		Ratio Test	Indexes		
Obs	63	LR chi2	48.40	R2	0.536
sigma	9.5917	d.f.	4	R2 adj	0.504
d.f.	58	Pr(> chi2)	0.0000	g	11.365

Residuals

	Min	1Q	Median	3Q	Max
	-22.1258	-4.7737	-0.2537	5.6112	22.4852

	Coef	S.E.	t	Pr(> t)
Intercept	51.1801	13.2827	3.85	0.0003
PSA	-1.1277	0.3309	-3.41	0.0012
preQoL	0.4356	0.0639	6.82	<0.0001
Age	-0.0383	0.1577	-0.24	0.8090
Hypertension=Yes	-0.2003	2.6580	-0.08	0.9402

```
## Radiation subgroup ##

## ( model1.Rad <- ols( postQoL ~ PSA, data=subset(ds, Treatment=='Radiation') ) )
(model2.Rad <- ols(postQoL ~ PSA + preQoL + Age + Hypertension, data = subset(ds, Treatment == "Radiation")))
```

Linear Regression Model

```
ols(formula = postQoL ~ PSA + preQoL + Age + Hypertension, data = subset(ds,
  Treatment == "Radiation"))
```

		Model Likelihood		Discrimination	
		Ratio Test		Indexes	
Obs	37	LR chi2	20.67	R2	0.428
sigma	13.1585	d.f.	4	R2 adj	0.356
d.f.	32	Pr(> chi2)	0.0004	g	12.230

Residuals

	Min	1Q	Median	3Q	Max
	-22.712	-9.585	-1.557	8.701	24.310

	Coef	S.E.	t	Pr(> t)
Intercept	16.0469	22.7618	0.70	0.4859
PSA	-2.0112	0.6371	-3.16	0.0035
preQoL	0.5074	0.1183	4.29	0.0002
Age	0.2426	0.2733	0.89	0.3814
Hypertension=Yes	3.0653	4.9225	0.62	0.5379

```
## Interaction model ## ( model1.Int <- ols( postQoL ~ PSA * Treatment, data=ds) )
(model2.Int <- ols(postQoL ~ PSA * Treatment + preQoL + Age + Hypertension, data = ds))
```

Linear Regression Model

```
ols(formula = postQoL ~ PSA * Treatment + preQoL + Age + Hypertension,
     data = ds)
```

		Model Likelihood	Discrimination
		Ratio Test	Indexes
Obs	100	LR chi2	102.70
		R2	0.642
sigma	10.8848	d.f.	6
		R2 adj	0.619
d.f.	93	Pr(> chi2)	0.0000
		g	16.247

Residuals

	Min	1Q	Median	3Q	Max
	-22.3830	-7.4809	-0.6394	6.2940	25.1030

	Coef	S.E.	t	Pr(> t)
Intercept	30.6460	12.5304	2.45	0.0163
PSA	-2.0236	0.5055	-4.00	0.0001
Treatment=Surgery	10.7106	6.5753	1.63	0.1067
preQoL	0.4697	0.0573	8.20	<0.0001
Age	0.0742	0.1397	0.53	0.5967
Hypertension=Yes	1.1225	2.3926	0.47	0.6401
PSA * Treatment=Surgery	0.8945	0.6090	1.47	0.1453

Coefficients:

	Surgery Subgroup		Radiation Subgroup		Interaction	
	Coefficient	<i>p</i>	Coefficient	<i>p</i>	Coefficient	<i>p</i>
Intercept	51.18	0.000	16.05	0.49	30.65	0.016
PSA	-1.13	0.001	-2.01	0.003	-2.02	0.000
Surgery	-	-	-	-	10.71	0.11
PSA × Surgery	-	-	-	-	0.89	0.15
preQoL	0.44	0.000	0.51	0.000	0.47	0.000
Age	-0.04	0.81	0.24	0.38	0.07	0.60
Hypertension	-0.20	0.94	3.07	0.54	1.12	0.64

Again, a clear advantage of the interaction model is the ability to test for differences of PSA effect between treatments.

And now, there doesn't seem any simple algebraic relationship between these coefficients. It is because the interaction model does not estimate preQoL, Age, or Hypertension effect separately for Surgery and Radiation groups.

If we want to estimate these secondary effects separately for the two groups, we must have treatments interacting with every single covariate.

```
##### Big interaction model ##
(model3.Int <- ols(postQoL ~ Treatment * (PSA + preQoL + Age + Hypertension), data = ds))
```

Linear Regression Model

```
ols(formula = postQoL ~ Treatment * (PSA + preQoL + Age + Hypertension),
     data = ds)
```

	Model Likelihood	Discrimination
	Ratio Test	Indexes
Obs	100	LR chi2 104.00
sigma	10.9933	d.f. 9
d.f.	90	Pr(> chi2) 0.0000
		R2 0.647
		R2 adj 0.611
		g 16.292

Residuals

Min	1Q	Median	3Q	Max
-22.7119	-7.5050	-0.7648	5.8757	24.3096

	Coef	S.E.	t	Pr(> t)
Intercept	16.0469	19.0163	0.84	0.4010
Treatment=Surgery	35.1332	24.3594	1.44	0.1527
PSA	-2.0112	0.5323	-3.78	0.0003
preQoL	0.5074	0.0989	5.13	<0.0001
Age	0.2426	0.2284	1.06	0.2909
Hypertension=Yes	3.0653	4.1125	0.75	0.4580
Treatment=Surgery * PSA	0.8836	0.6536	1.35	0.1798
Treatment=Surgery * preQoL	-0.0718	0.1230	-0.58	0.5609
Treatment=Surgery * Age	-0.2809	0.2913	-0.96	0.3374
Treatment=Surgery * Hypertension=Yes	-3.2656	5.1180	-0.64	0.5250

Coefficients:

	Surgery Subgroup		Radiation Subgroup		Big Interaction	
	Coefficient	<i>p</i>	Coefficient	<i>p</i>	Coefficient	<i>p</i>
Intercept	51.18	0.000	16.05	0.49	16.05	0.40
PSA	-1.13	0.001	-2.01	0.003	-2.01	0.000
Surgery	-	-	-	-	35.13	0.15
PSA × Surgery	-	-	-	-	0.88	0.18
preQoL	0.44	0.000	0.51	0.000	0.51	0.000
preQoL × Surgery	-	-	-	-	-0.07	0.56
Age	-0.04	0.81	0.24	0.38	0.24	0.29
Age × Surgery	-	-	-	-	-0.28	0.34
Hypertension	-0.20	0.94	3.07	0.54	3.07	0.46
Hypertension × Surgery	-	-	-	-	-3.27	0.53

For this example, this means we must estimate 10 coefficients. With a sample size of 100, perhaps, it is too much. But that's exactly what we are doing with these **subgroup** analyses.

Number of coefficients:

Surgery Subgroup	5
Radiation Subgroup	5
Total	10
Interaction Model (PSA and Treatment)	7
Big interaction Model	10

2 Baseline Adjustment vs Difference

Suppose we would like to compare the two treatments on *postQoL*. We know that *postQoL* is correlated with *preQoL*, so we will take that information into account.

Baseline preQoL

	N	Min	Q1	Med	Q3	Max	Mean	SD	SE
Radiation	37	19	36	48	67	86	51	20	3.2
Surgery	63	16	39	63	81	95	59	23	2.9
Combined	100	16	36	57	76	95	56	22	2.2

6 month postQoL

	N	Min	Q1	Med	Q3	Max	Mean	SD	SE
Radiation	37	3.2	29	40	52	75	41	16	2.7
Surgery	63	26.7	55	62	72	86	62	14	1.7
Combined	100	3.2	42	57	69	86	54	18	1.8

Change postQoL - preQoL

	N	Min	Q1	Med	Q3	Max	Mean	SD	SE
Radiation	37	-56	-17	-6.6	5.0	25	-10.0	19	3.1
Surgery	63	-33	-11	0.3	12.7	55	2.2	18	2.2
Combined	100	-56	-14	-2.9	9.2	55	-2.3	19	1.9

One approach is to compute difference, $postQoL - preQoL$, to define **QoL change**.

```
## QoL Change
(ba <- ols((postQoL - preQoL) ~ Treatment, data = ds))

Linear Regression Model

ols(formula = (postQoL - preQoL) ~ Treatment, data = ds)

              Model Likelihood      Discrimination
              Ratio Test              Indexes
Obs          100  LR chi2      10.32  R2          0.098
sigma18.0939  d.f.           1      R2 adj     0.089
d.f.          98  Pr(> chi2) 0.0013  g          5.761

Residuals

      Min       1Q   Median       3Q      Max
-46.0081 -11.9429  -0.8255  12.7158  52.5571

              Coef      S.E.   t      Pr(>|t|)
Intercept      -9.9919  2.9746  -3.36  0.0011
Treatment=Surgery 12.2347  3.7477   3.26  0.0015
```

Mean change in QoL (6 month – baseline) is higher for Surgery group by 12.23. Also, the mean change in Radiation group is –9.99.

Another approach is to regress *postQoL* on *Treatment* while accounting for *preQoL*.

```
## preQoL adjustment
(b0 <- ols(postQoL ~ Treatment + preQoL, data = ds))
```

Linear Regression Model

```
ols(formula = postQoL ~ Treatment + preQoL, data = ds)
```

	Model Likelihood	Discrimination
	Ratio Test	Indexes
Obs	100	LR chi2 78.28
sigma	12.0420	d.f. 2
d.f.	97	Pr(> chi2) 0.0000
		R2 0.543
		R2 adj 0.533
		g 14.990

Residuals

Min	1Q	Median	3Q	Max
-35.4737	-7.3794	-0.7145	8.1922	30.3027

	Coef	S.E.	t	Pr(> t)
Intercept	21.5708	3.4549	6.24	<0.0001
Treatment=Surgery	17.1698	2.5332	6.78	<0.0001
preQoL	0.3861	0.0551	7.01	<0.0001

“On average, *postQoL* is higher for Surgery group by 17.17 while adjusting for *preQoL*.” (Please remember this number, 17.17.)

Let's compare the regression models of the two approaches.

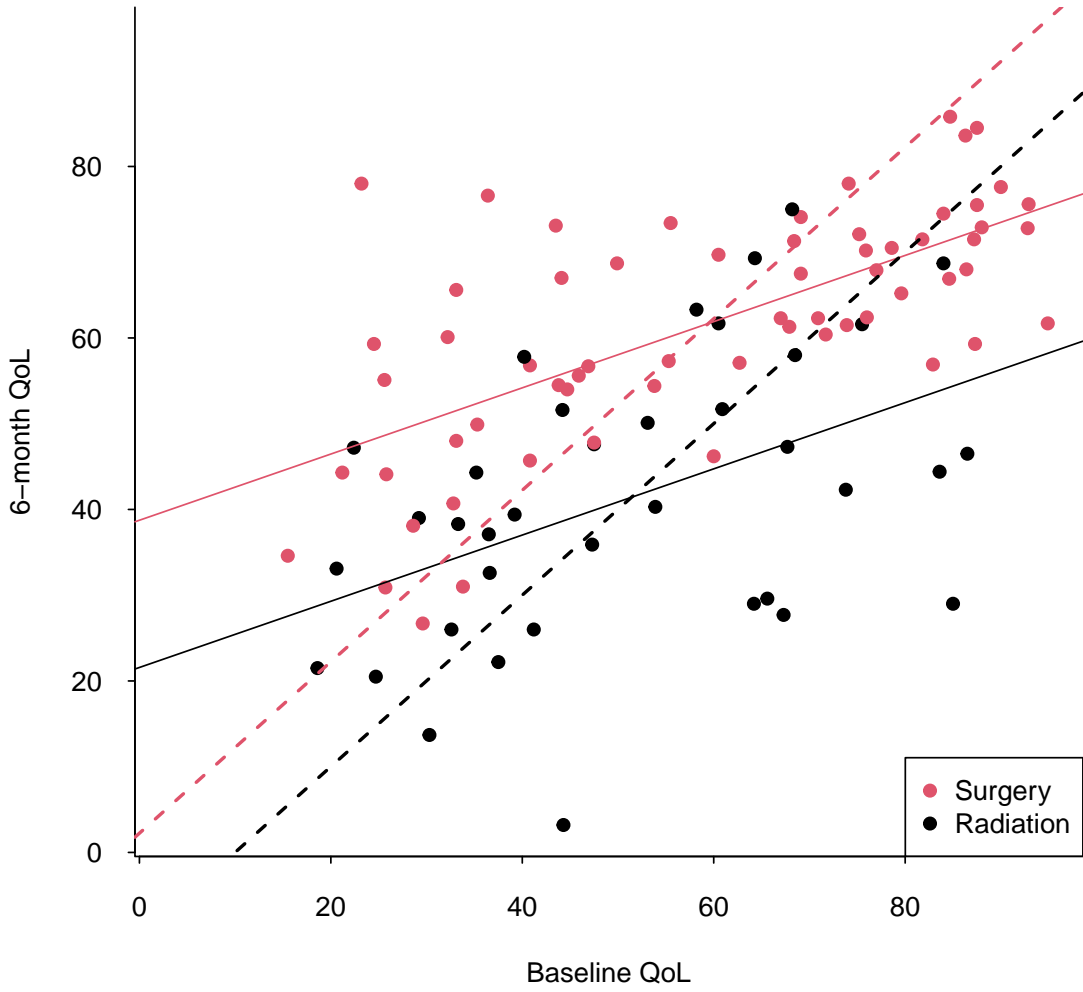
Approach 1: (Take difference)

$$\begin{aligned}Y_{post} - Y_{pre} &= \beta_0 + \beta_s X_s \\Y_{post} &= \beta_0 + \beta_s X_s + 1 \cdot Y_{pre}\end{aligned}$$

Approach 2: (Regress Y_{post} on Y_{pre})

$$Y_{post} = \beta'_0 + \beta'_s X_s + \beta'_y Y_{pre}$$

Comparing these equations, we notice that approach 1 forces the coefficient on Y_{pre} to be 1, while approach 2 allows us to estimate the coefficient using the data.



3 Analyzing Difference with Baseline as a Covariate

I have seen regression models where the response is the difference **and** the baseline value is included as a covariate. The question may be: Does difference from baseline depend on the baseline values?

```
(m00 <- ols((postQoL - preQoL) ~ Treatment + preQoL, data = ds))
```

Linear Regression Model

```
ols(formula = (postQoL - preQoL) ~ Treatment + preQoL, data = ds)
```

		Model Likelihood		Discrimination	
		Ratio Test		Indexes	
Obs	100	LR chi2	92.78	R2	0.605
sigma	2.0420	d.f.	2	R2 adj	0.596
d.f.	97	Pr(> chi2)	0.0000	g	16.927

Residuals

	Min	1Q	Median	3Q	Max
	-35.4737	-7.3794	-0.7145	8.1922	30.3027

	Coef	S.E.	t	Pr(> t)
Intercept	21.5708	3.4549	6.24	<0.0001
Treatment=Surgery	17.1698	2.5332	6.78	<0.0001
preQoL	-0.6139	0.0551	-11.15	<0.0001

You might want to say, “On average, Surgery group’s $postQoL - preQoL$ is 17.17 higher while adjusting for $preQoL$.”

But we have seen this number, 17.17, before. It turns out this approach is closely related to our favorite approach, “Regress $postQoL$ and use $preQoL$ as covariate,” only incorrect.

Coefficients:

Response	$postQoL$				$postQoL - preQoL$			
	Coefficient	SE	t	p	Coefficient	SE	t	p
Intercept	21.57	3.45	6.24	0.000	21.57	3.45	6.24	0.000
Surgery	17.17	2.53	6.78	0.000	17.17	2.53	6.78	0.000
$preQoL$	0.39	0.06	7.01	0.000	-0.61	0.06	-11.15	0.000

Let's compare the regression models:

Regress Y_{post} on Y_{pre}

$$Y_{post} = \beta_0 + \beta_s X_s + \beta_y Y_{pre}$$

Regress Difference on Y_{pre}

$$Y_{post} - Y_{pre} = \beta'_0 + \beta'_s X_s + \beta'_y Y_{pre}$$

$$Y_{post} = \beta'_0 + \beta'_s X_s + (\beta'_y + 1) Y_{pre}$$

Therefore, $\beta_0 = \beta'_0$, $\beta_s = \beta'_s$, and $\beta_y = \beta'_y + 1$.

Is there a problem?

- If the question is regarding β_s , then probably yes, because interpretation is confusing.
- If the question is regarding β_y , then definitely yes,

$$H_0 : \beta'_y = 0$$

$$H_1 : \beta'_y \neq 0$$

does not test what it seems to test. When there is no association between Y_{pre} and Y_{post} , $\beta'_y = -1$, and the above null hypothesis is false.

```
set.seed(324)
y0 <- rnorm(200)
y6 <- rnorm(200)

(regX <- ols((y6 - y0) ~ y0))
```

Linear Regression Model

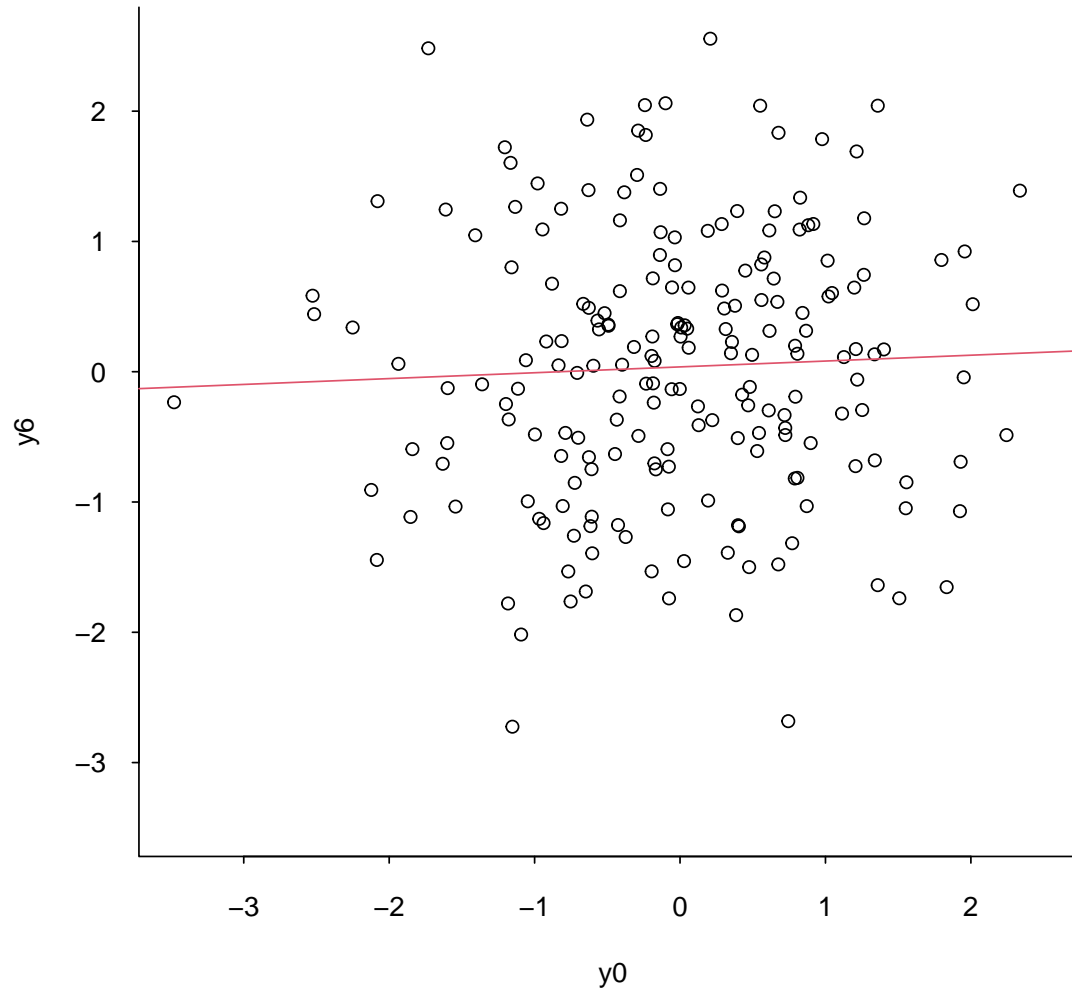
```
ols(formula = (y6 - y0) ~ y0)
```

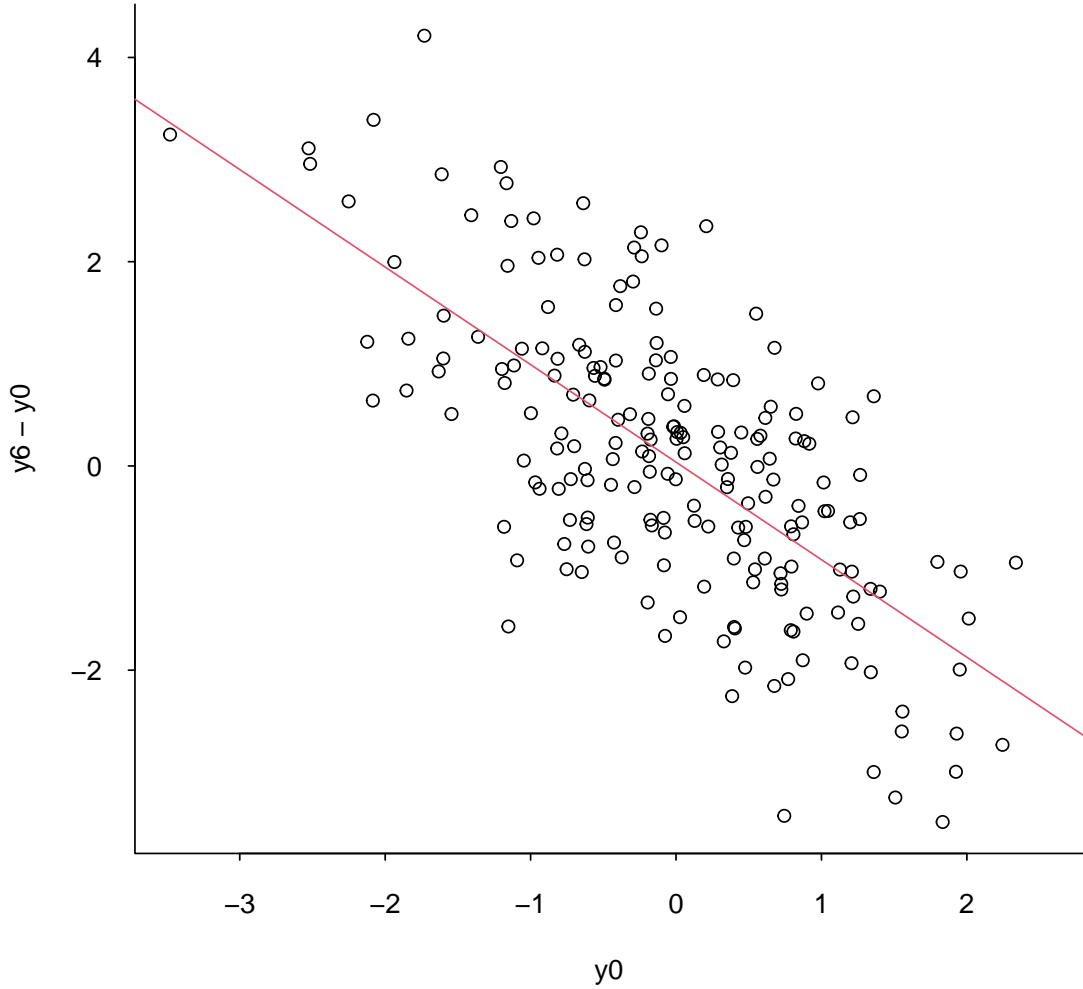
	Model Likelihood	Discrimination			
	Ratio Test	Indexes			
Obs	200	LR chi2	127.37	R2	0.471
sigma	1.0218	d.f.	1	R2 adj	0.468
d.f.	198	Pr(> chi2)	0.0000	g	1.083

Residuals

	Min	1Q	Median	3Q	Max
	-2.7525	-0.6859	0.0517	0.6644	2.5232

	Coef	S.E.	t	Pr(> t)
Intercept	0.0370	0.0723	0.51	0.6093
y0	-0.9551	0.0719	-13.28	<0.0001





The take-home messages

- Interaction models are *always* better than the subgroup models.
- Baseline adjustment is *almost always* better than taking the difference.
- Baseline adjustment on top of taking the difference is *never* a good idea.