Lecture 18
Topic 13: Analysis of Covariance (ANCOVA), Part I

Suppose that the response variable (Y) is linearly related to some other continuous variable (X) that the experimenter cannot control but can observe, *along with Y*:

Examples:

Initial weight of animals in a feeding trial.

Native soil fertility in a yield trial.

Overall level of DNA transcription in a gene expression study.

Maturity at harvest in a vegetable quality study.

In such studies:

X is a **covariable** (or **covariate** or **concomitant variable**)

ANCOVA uses X essentially as a *continuous blocking variable* to improve the precision of an experiment.

ANCOVA can provide great insight into the nature of treatment effects.

ANCOVA is a combination of **ANOVA** and **Linear Regression**

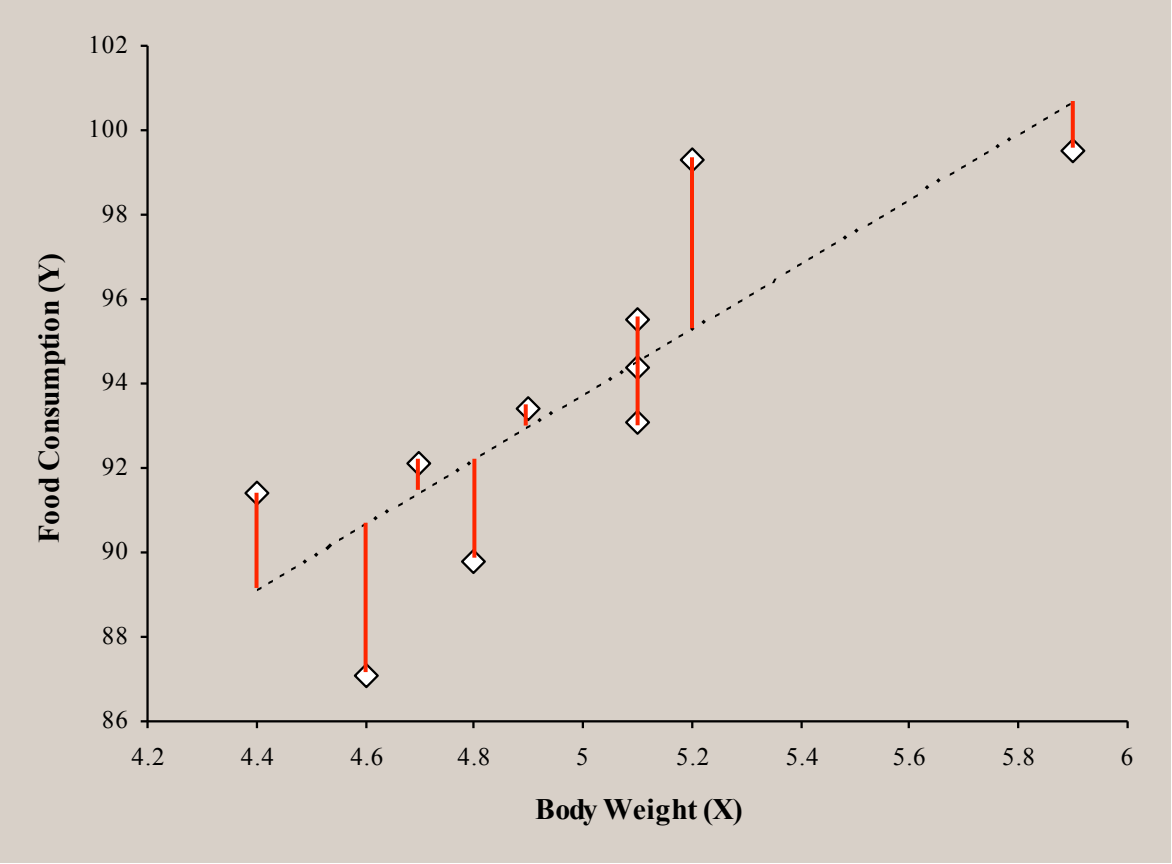
Let us regress (a review of regression concepts)

The equation of a straight line is $Y = a + bX$

a = the **intercept** b = the **slope**

Example: Body weight (X) vs. individual food consumption (Y) for 10 animals.

Body weight (X)	Food consumption (Y)
4.6	87.1
5.1	93.1
4.8	89.8
4.4	91.4
5.9	99.5
4.7	92.1
5.1	95.5
5.2	99.3
4.9	93.4
5.1	94.4



The Principle of Least Squares
The line of best fit is the one which minimizes the sum of squared deviations.

Calculating *a* and *b*

Equations for the intercept *a* and the slope *b* that minimize the SSE:

$$b = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} \equiv \frac{S(XY)}{SS(X)} \quad \text{and} \quad a = \overline{Y} - b\overline{X}$$

For the sample dataset above:

$$b = \frac{\left[(4.6 - 4.98)(87.1 - 93.56) + \ldots + (5.1 - 4.98)(94.4 - 93.56)\right]}{(4.6 - 4.98)^2 + \ldots + (5.1 - 4.98)^2} = 7.69$$

$$a = 93.56 - 7.69(4.98) = 55.26$$

The equation of the best fit line: $\boldsymbol{Y = 55.26 + 7.69X}$

S(XY) is called the **corrected sum of cross products**

$\frac{S(XY)}{n-1}$ is called the **sample covariance**

In R:

```
X <- c(4.6, 4.7, 5.1, 5.1, 4.8, 5.2, 4.4, 4.9, 5.9, 5.1)
Y <- c(87.1, 92.1, 93.1, 95.5, 89.8, 99.3, 91.4, 93.4, 99.5, 94.4)

regression <- lm(Y ~ X)
anova(regression)
summary(regression)
```

Output

```
Analysis of Variance Table
Response: Y
           Df   Sum Sq   Mean Sq  F value    Pr(>F)
X           1 90.83551 90.835510 16.23204 0.0037939 **
Residuals   8 44.76849  5.596061


Coefficients:
             Estimate Std. Error t value   Pr(>|t|)
(Intercept) 55.263281   9.534890  5.7959 0.00040706 ***
X            7.690104   1.908735  4.0289 0.00379385 **
Multiple R-squared:  0.6698586
```

The model accounts for 67% of the variation in the experiment.

$$Y = 55.26 + 7.69X$$

Analysis of adjusted *Y*'s

The SSE (**44.77**) represents the variation in food consumption (Y) that would have been observed if all the animals used in the experiment had had the same initial body weight (X):

X	Y	Adjusted Y $= Y - b(X - \bar{X})$
4.6	87.1	90.0222
5.1	93.1	92.1772
4.8	89.8	91.1842
4.4	91.4	95.8602
5.9	99.5	92.4252
4.7	92.1	94.2532
5.1	95.5	94.5772
5.2	99.3	97.6082
4.9	93.4	94.0152
5.1	94.4	93.4772
$\bar{X} = 4.98$	SS = 135.60	SS = 44.77

The results of a regression on the **adjusted Y's**

```
adjY <- Y - 7.69 * (X - mean(X))

adj_regression <- lm(adjY ~ X)
anova(adj_regression)
summary(adj_regression)
```

Output

```
Analysis of Variance Table

Response: adjY
          Df   Sum Sq   Mean Sq F value  Pr(>F)
X          1  0.00000 0.0000000       0 0.99996
Residuals  8 44.76849 5.5960612
Coefficients:
                 Estimate   Std. Error t value   Pr(>|t|)
(Intercept) 9.355948e+01 9.534890e+00 9.81233 9.7761e-06 ***
X           1.041667e-04 1.908735e+00 0.00005    0.99996
Multiple R-squared:  3.722857e-10
```

With all animals adjusted to the same initial weight:

1. Body weight (X) no longer explains any variation in the study (SSX = 0, slope ~ 0).
2. The SSE (44.77) is exactly the same as we saw before!

Adjusting each Y to a common X by the best-fit equation is equivalent, in terms of accounting for variation, to a linear regression.

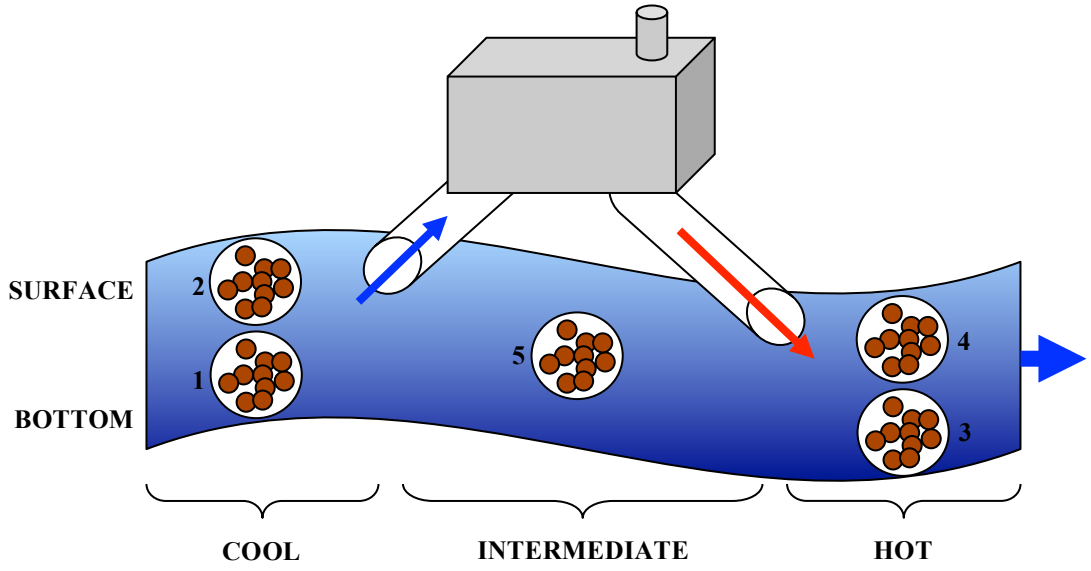
Oysters!

The objectives of a pilot experiment to study oyster growth:

1. To determine if exposure to artificially-heated water affects growth
2. To determine if position in the water column (surface vs. bottom) affects growth

Twenty bags of ten oysters each were placed across 5 locations near a riverside power-generation plant (i.e. 4 bags / location):

TRT1: cool-bottom **TRT2**: cool-surface **TRT3**: hot-bottom
TRT4: hot-surface **TRT5**: control (mid-depth, mid-temperature)



The bags were weighed at the beginning and the end of the experiment.

The data:

Trtmt	Rep	Initial	Final
1	1	27.2	32.6
1	2	32	36.6
1	3	33	37.7
…	…	…	…
5	2	19.6	23.4
5	3	25.1	30.3
5	4	18.1	21.8

The code:

```
# I. Simple overall regression
oyster_reg_mod<-lm(Final ~ Initial, oyster_dat)
anova(oyster_reg_mod)
summary(oyster_reg_mod)
```

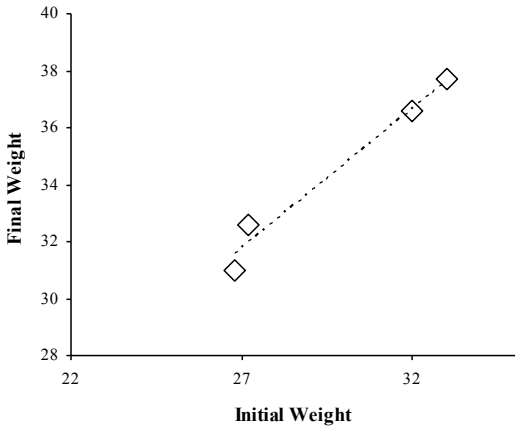
```
Analysis of Variance Table

Response: Final
          Df    Sum Sq   Mean Sq   F value     Pr(>F)
Initial    1 342.35782 342.35782 377.79308 1.5761e-13 ***
Residuals 18  16.31168   0.90620


Coefficients:
             Estimate Std. Error  t value   Pr(>|t|)
(Intercept) 3.7646865  1.4094093  2.67111   0.015577 *
Initial     1.0512544  0.0540855 19.43690 1.5761e-13 ***
Multiple R-squared:  0.9545217
```

```
# II. Using a loop to perform regressions at each treatment level
Trtmt_levels<-c(1:5)
for (i in Trtmt_levels) {
  with(subset(oyster_dat, Trtmt == Trtmt_levels[i]), {
    print(Trtmt_levels[i])
    print(summary(lm(Final ~ Initial)))
  })
}
```

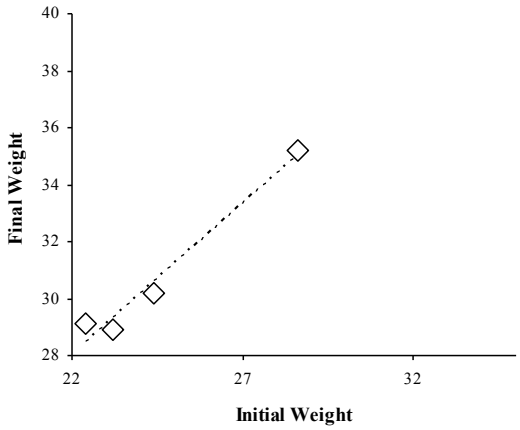
Parameter estimates within each treatment group:

```
Coefficients:
                 Estimate Std. Error  t value   Pr(>|t|)
Slope(Trt1)     0.9826468  0.1094183  8.98064   0.012173 *
Slope(Trt2)     1.5013550  0.3923086  3.82697   0.061997 .
Slope(Trt3)     1.0560666  0.1280121  8.24974   0.014377 *
Slope(Trt4)    1.05692503 0.06842649 15.44614  0.0041652 **
Slope(Trt5)    1.22388605 0.02530981 48.35619 0.00042738 ***
```

TRT1: $Y = 5.24 + 0.98X$



TRT3: $Y = 4.82 + 1.06X$




```
# III. The one-way ANOVA
oyster_anova_mod<-lm(Final ~ Trtmt, oyster_dat)
anova(oyster_anova_mod)
```

The ANOVA

"Are there differences in final weight across locations?"

```
Response: Final
          Df   Sum Sq   Mean Sq F value   Pr(>F)
Trtmt      4 198.4070 49.601750 4.64255 0.012239 *
Residuals 15 160.2625 10.684167
```

This model explains roughly 55% of the observed variation.

The ANCOVA

"Are there differences in final weights across locations, adjusting for differences in initial weights?"

```
# IV. The ANCOVA
#library(car)
oyster_ancova_mod<-lm(Final ~ Trtmt + Initial, oyster_dat)
anova(oyster_ancova_mod)
Anova(oyster_ancova_mod, type = 2)
```

The ANCOVA (Type I SS)

```
Analysis of Variance Table

Response: Final
           Df     Sum Sq    Mean Sq   F value     Pr(>F)
Trtmt       4 198.407000  49.601750 164.46503 1.3398e-11 ***
Initial     1 156.040177 156.040177 517.38400 1.8674e-12 ***
Residuals  14   4.222323   0.301595
```

The ANCOVA (Type II SS)

```
Anova Table (Type II tests)

Response: Final
              Sum Sq Df  F value     Pr(>F)
Trtmt      12.089359  4  10.0212 0.00048186 ***
Initial   156.040177  1 517.3840 1.8674e-12 ***
Residuals   4.222323 14
```

The Type I SS for TRT (198.4) is the **unadjusted treatment SS**.

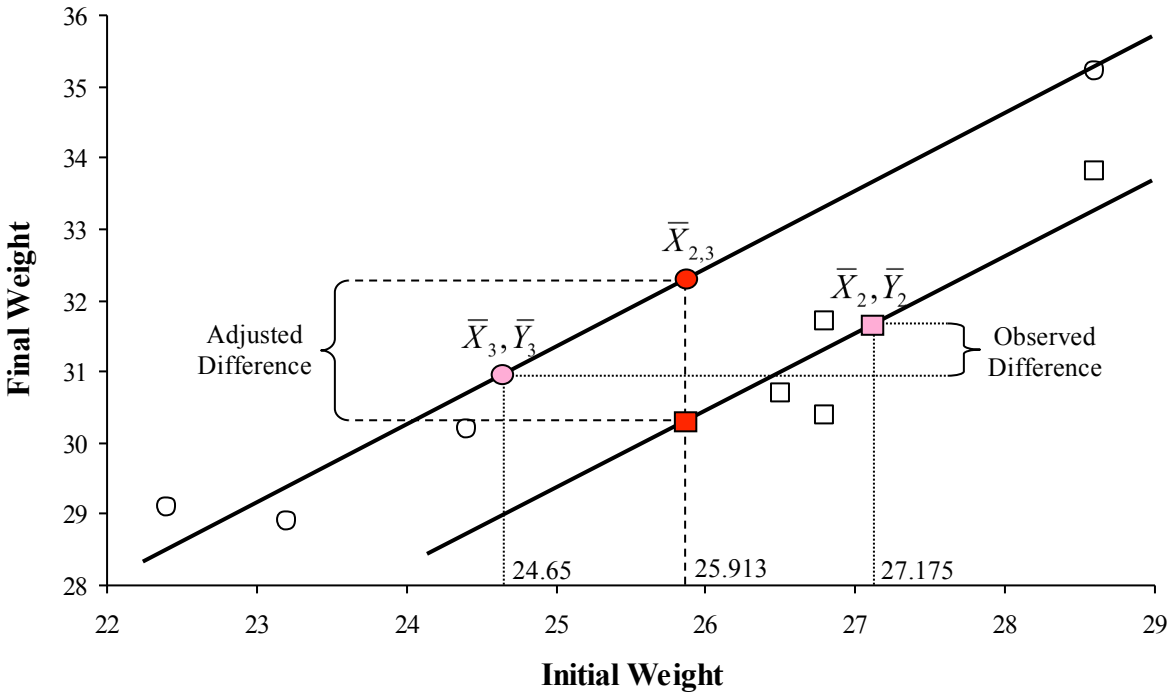
The Type II SS for TRT (12.1) is the **adjusted treatment SS** and allows us to test the treatment effects, adjusting for all other factors included in the model.

Type II SS produces the appropriate results in ANCOVA.
(not all levels of Trtmt are present in all levels of continuous X)

The power of the test for Trtmt effects increases when the covariate is included because most of the error in the simple ANOVA is due to variation among INITIAL values.

The take-home visualization of ANCOVA

The data for **Treatments 2 (white squares) and 3 (white circles)** from the oyster example.



Comparing unadjusted (observed) means:

$$\bar{Y}_3 = 30.85 < 31.65 = \bar{Y}_2$$

Comparing adjusted means:

$$\bar{Y}_3 = 32.05 > 30.12 = \bar{Y}_2$$

Least squares adjusted means

For valid comparisons, treatment means should be adjusted to what their values *would have been if all oysters had had the same initial weight*.

```
Final_means <- aggregate(oyster_dat$Final, list(oyster_dat$Trtmt), mean)
oyster.lsm <- lsmeans(oyster_ancova_mod, "Trtmt")
```

TRT	Initial Means	Unadjusted Means	Adjusted LS Means	Calculation [$\bar{Y}_{adj_i} = \bar{Y}_i - \beta(\bar{X}_i - \bar{X})$]
1	29.750	34.475	30.153	34.475 - 1.08318 (29.75 - 25.76)
2	27.175	31.650	30.117	31.650 - 1.08318 (27.18 - 25.76)
3	24.650	30.850	32.052	30.850 - 1.08318 (24.65 - 25.76)
4	26.425	32.225	31.504	32.225 - 1.08318 (26.43 - 25.76)
5	20.800	25.025	30.398	25.025 - 1.08318 (20.80 - 25.76)

The coefficient β = 1.08318 represents a "best" single slope value that describes the relationship between X and Y, accounting for all other classification variables:

```
> summary(oyster_ancova_mod)

Call:
lm(formula = Final ~ Trtmt + Initial, data = oyster_dat)

Residuals:
       Min          1Q      Median          3Q         Max
-0.8438076 -0.3154120 -0.2170735  0.4863336  0.8871085

Coefficients:
             Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)  2.25040039  1.44307538  1.55945 0.14120460
Trtmt2      -0.03581197  0.40722674 -0.08794 0.93116903
Trtmt3       1.89921708  0.45801799  4.14660 0.00098809 ***
Trtmt4       1.35157290  0.41936648  3.22289 0.00613476 **
Trtmt5       0.24445938  0.57658196  0.42398 0.67802248
Initial      1.08317982  0.04762051 22.74608 1.8674e-12 ***
```

This "best" slope can be used to create a new adjusted response variable:

$$Z = Y - \beta(X - \bar{X})$$

Contrasts

The adjusted means can be analyzed further with orthogonal contrasts:

```
#Comparing LSMeans, using the "lsmeans" package (function contrast())
oyster.lsm <- lsmeans(oyster_ancova_mod, "Trtmt")

#Contrasts
contrast(oyster.lsm, list("control vs. trtmt"=c(-1,-1,-1,-1,4),
                          "bottom vs. surface"=c(-1,1,-1,1,0),
                          "cool vs. hot"=c(-1,-1,1,1,0),
                          "depth*temp"=c(1,-1,-1,1,0)))
```

The output:

contrast	estimate	SE	df	t.ratio	p.value
control.vs..trtmt	-2.2371404940	1.7037332311	14	-1.313	0.2103
bottom.vs..surface	-0.5834561450	0.5504960078	14	-1.060	0.3071
cool.vs..hot	3.2866019399	0.6157932555	14	5.337	0.0001
depth.temp	-0.5118322117	0.5869457568	14	-0.872	0.3979

If the covariable is **not** included in the model, these exact same contrasts produce completely different results:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Trtmt	4	198.4070	49.60175	4.64255	0.012239	*
Trtmt: Cont v. Trt	1	169.3620	169.36200	15.85168	0.001204	**
Trtmt: Bot vs. Surf	1	2.1025	2.10250	0.19679	0.663659	
Trtmt: Cool vs. Hot	1	9.3025	9.30250	0.87068	0.365545	
Trtmt: Depth*Temp	1	17.6400	17.64000	1.65104	0.218305	
Residuals	15	160.2625	10.68417			

Why not use (Final – Initial) as the response variable?

ANOVA (no covariable)

Dependent Variable: Final Weight $R^2 = 0.55$

```
Response: Final
          Df Sum Sq Mean Sq F value  Pr(>F)
Trtmt      4 198.41  49.602  4.6425 0.01224 *
Residuals 15 160.26  10.684
```

Dependent Variable: Difference (Final – Initial) $R^2 = 0.70$

```
Response: Diff
          Df  Sum Sq Mean Sq F value    Pr(>F)
Trtmt      4 11.9830 2.99575  8.7382 0.0007537 ***
Residuals 15  5.1425 0.34283
```

ANCOVA (Initial Weight as covariable)

Dependent Variable: Final Weight $R^2 = 0.99$

```
Anova Table (Type II tests)

Response: Final
           Sum Sq Df F value    Pr(>F)
Trtmt      12.089  4  10.021 0.0004819 ***
Initial   156.040  1 517.384 1.867e-12 ***
Residuals   4.222 14
```

Dependent Variable: Difference (Final – Initial) $R^2 = 0.75$

```
Anova Table (Type II tests)

Response: Diff
           Sum Sq Df F value    Pr(>F)
Trtmt     12.0894  4  10.021 0.0004819 ***
Initial    0.9202  1   3.051 0.1025771
Residuals  4.2223 14
```

Comparison between ANCOVA and ANOVA of ratios

In a study of the effect of stress on the presence of enzyme A in the liver, researchers measured the total activity of enzyme A from liver homogenates of 10 control and 10 shocked animals and the total amount of N as an indicator of total enzyme activity in the liver. A/N = the activity of enzyme A per unit protein.

Control animals			Shocked animals		
N	**A**	**A/N**	**N**	**A**	**A/N**
84	76	90.4	122	108	88.5
28	38	133.9	98	158	161.2
166	72	43.4	115	58	50.0
98	64	65.3	86	65	75.5
105	53	50.0	69	40	58.0
84	28	32.8	86	65	75.5
72	31	43.0	102	82	80.3
80	28	34.3	112	94	84.1
84	28	32.7	98	65	66.3
105	49	46.1	74	76	102.7

ANOVA of the variable (A/N) **$R^2 = 0.16$**

```
Response: A/N
          Df  Sum Sq Mean Sq F value  Pr(>F)
Group      1  3535.4  3535.4  3.5123 0.07724 NS
Residuals 18 18118.5  1006.6
```

ANCOVA of the variable A, using N as a covariable **$R^2 = 0.43$**

```
Response: A
           Sum Sq Df F value  Pr(>F)
Group      5108.8  1  7.9807 0.01167 *
N          2162.6  1  3.3783 0.08360 .
Residuals 10882.4 17
```

The use of ANOVA to analyze ratios *Z = Y/X* is not correct.

Both *X* and *Y*, being random variables, exhibit random variation

Variation in Y affects *Z* in a linear way, but variation in X affects *Z* in a hyperbolic way

The error in Z depends not only on the error in *X* but also on the absolute value of *X*.

ANCOVA model

The linear model for ANOVA of a CRD:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

The linear model for linear regression:

$$Y_i = \mu + \beta(X_i - \overline{X}_{\cdot}) + \varepsilon_i$$

ANCOVA is a combination of ANOVA and regression:

$$Y_{ij} = \mu + \tau_i + \beta(X_{ij} - \overline{X}_{\cdot\cdot}) + \varepsilon_{ij}$$

A simple rearrangement:

$$Y_{ij} - \beta(X_{ij} - \overline{X}_{\cdot\cdot}) = \mu + \tau_i + \varepsilon_{ij}$$

An ANCOVA on the unadjusted values of Y is equivalent to a regular ANOVA on the adjusted values of Y

Assumptions of the ANCOVA model

OLD

1. The residuals are normally and independently distributed with zero mean and a common variance.

NEW

2. The X's are fixed, measured without error, and independent of treatments.

3. The regression of Y on X is linear and independent of treatments.