STA303/1002 - Methods of Data Analysis II

(Week 05 lecture note - Extra Topics)

Wei (Becky) Lin

Jan. 24/26, 2017

UNIVERSITY OF TORONTO

Review: Topics learned last week

- Identifiability problem
  - Imposing constraint will affect the summary output, i.e. the LSE of regression coefficient but not for the ANOVA output.
  - Imposing constraint won't affect the F test (near the end of summary output) for

$$H_0 : \alpha_i = \beta_j = (\alpha\beta)_{ij} = 0, \forall i, j$$

← No any main/interaction effect.

$$H_a : \exists i \text{ or/and } j, s.t. \alpha_i = \beta_j = (\alpha\beta)_{ij} \neq 0$$

- Understand and interpret interaction effect in Two-way ANOVA
  - Know how to obtain an interaction plot
  - Know how to read an interaction plot
  - Know how to test the interaction effect is significant or not.
- Data Example
  - Battery data example.

Learning Objective This Week

- ANCOVA: Analysis of Covariance
- Weighted Least Square (WLS)

ANCOVA - Analysis of Covariance

Introduction to ANCOVA

In ANOVA, it is assumed that source of variations are mainly due to one or more treatment factors. If confounding or nuisance variable are expected to be another major source of variations, they should be taken into account either in the design or the analysis level.

- What does ANCOVA stand for?
  - Answer: Analysis of Covariance.
- What is the main goal of ANCOVA?
  - The primary goal of ANCOVA is to make mean comparison between treatment groups, but this is done by making an adjustment based on a regression line with potential confounding covariates.

Introduction to ANCOVA

- What is ANCOVA?
  - The model that allows comparison of treatment effects at any fixed level of X is Analysis of Covariance model.
  - For simplicity, assume that there are one factor and p-covariates, then the analysis of covariance model is

non-factor variables

ANOVA adjusted by X

$$Y_{ik} = \mu + \alpha_i + \sum_{j=1}^{p} \beta_j x_{ijk} + \epsilon_{ik}, k = 1, \dots, n_i; i = 1, \dots, a$$

$\hookrightarrow Y_{ik} - \sum_{j=1}^{p} \beta_j X_{ijk} = \mu + \alpha_i + \varepsilon_{ik}$

  - Note that in R, *aov*() does the type I SS (sequential sums of squares). So how the order of covariates enter in the aov model matters. If use aov() in R, stick to the model form: $Y \sim X$+factor variable in order to obtain the corrrect treatment effect after adjusted by covariate X, otherwise we need to use III anova SS, *Anova*() in R package: *car*.
  - Type III SS produces the appropriate results in ANCOVA. If use *aov*(), need to make sure treatment factor is the last variable to enter the regression.

Introduction to ANCOVA

- Basic Approach
  - ANCOVA can be thought of as removing out an unwanted influence on the dependent variable (Y), then proceeding with an ANOVA.
  - **ANCOVA** is really "ANOVA with covariates" or, more simply, a combination of a linear regression and an ANOVA model.
  - Use when you have some categorical factors and some quantitative predictors. Continuous variables are referred to as covariates variables.
  - The idea is that additional covariates **are not** necessarily of primary interest, but still their inclusion in the model will help explain more of the response, and hence reduce the error variance.
  - In some situations, failure to include an important covariate can yield misleading results. (Refer to the following data examples 1 and 2)

Motivation: data example 1

- Want to study three potential treatments of an aggressive form of cancer
- Response variable is the number of months a patient lives after being placed on a treatment
- We start to analyze the data as a one-way ANOVA

Motivation: data example 1

```
library(ggplot2)
treatment=as.factor(c(1 ,1 ,1, 2, 2, 2, 3, 3, 3))
duration= c(1.2 ,1.9 ,3.4 ,4.0, 5.2, 5.8, 7.7, 8.3, 8.9) # stage of Cancer (yrs
lifetime=c(42, 45, 35, 32, 22 ,20 ,16, 13 ,11) # months
cancer=data.frame(treatment,duration,lifetime)
ggplot(cancer,aes(x=treatment,y=lifetime,col=treatment))+geom_point()
```

Motivation: data example 1

```
fit = lm(lifetime~treatment,data=cancer)
anova(fit)
```

μ_i: survival time of trmt i, $i=1,2,3$

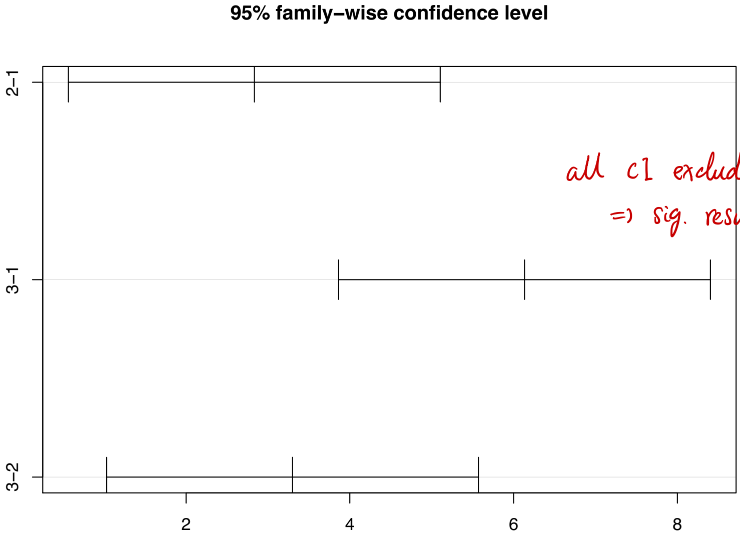
```
## Analysis of Variance Table
##
## Response: lifetime
##           Df Sum Sq Mean Sq F value   Pr(>F)
## treatment  2 56.536 28.2678  34.426 0.000515 ***
## Residuals  6  4.927  0.8211
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$\begin{cases} H_0: \mu_1=\mu_2=\mu_3 \\ H_a: \text{at least one is diff from others.} \end{cases}$

- Seems clear there is a significant treatment effect.

Motivation: data example 1

```r
mod = aov(lifetime~treatment,data=cancer)
plot(TukeyHSD(mod,"treatment",conf.level=0.95))
```



all CI excludes 0

=) sig. result (i.e. Rej H_0)

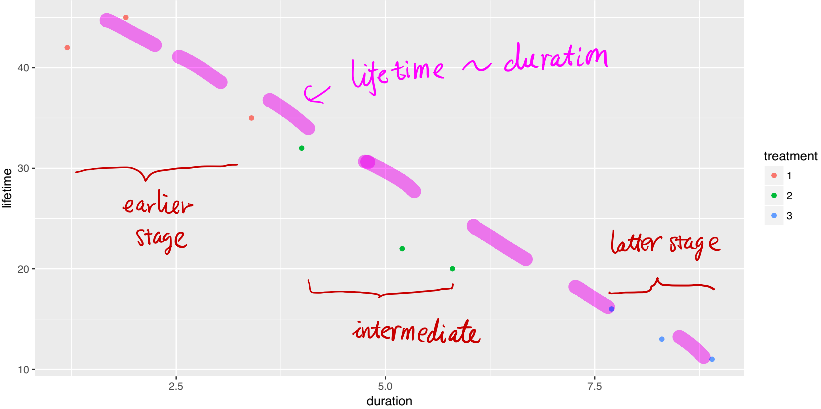
Motivation: data example 1

- From ANOVA output and multiple comparison, it seems clear there is a significant treatment effect.
- There is a significant treatment effect. It suggests that Treatment 1 is clearly the best (since people live longer).
- So we put a large group of people on Treatment 1 expecting them to live 40+ months, but unfortunately they do not live this long. What did we do wrong????

Motivation: data example 1

The oversignt

- The stage to which the cancer has progressed at the time that treatment begins. → duration covariates
  - This is important, because those at earlier stages of disease will naturally live longer on average.



Motivation: data example 1

Life time VS Duration

- From previous plot, clearly, there is a linear relationship between the duration of the cancer and the length of time someone has left to live.
- Furthermore, we notice from the plot that the group assigned to the first treatment were all in an earlier stage of the disease, those assigned to the second treatment were all in a middle stage, and those assigned to the third treatment were all in a later stage.

Reexamine the treatment effect

- After seeing the new plot, it is clear that we can't compare the lifetimes without considering the duration of the disease.
- We would suspect that the treatments are not all that different after looking at this new plot.

Motivation: data example 1

The following ANCOVA output leads to that conclusion:

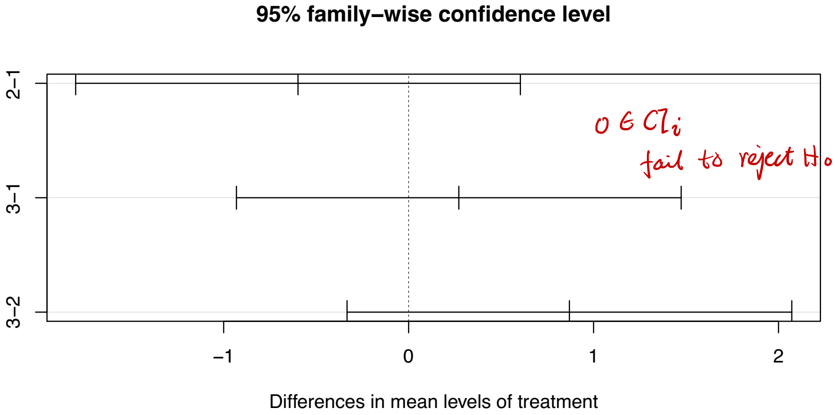
```
fit= lm(lifetime~duration+treatment)
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: lifetime
##           Df  Sum Sq Mean Sq  F value    Pr(>F)
## duration   1 1225.17 1225.17 199.1476 3.216e-05 ***
## treatment  2   23.63   11.81   1.9204    0.2405
## Residuals  5   30.76    6.15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Motivation: data example 1

The following ANCOVA output leads to that conclusion:

```
plot(TukeyHSD(aov(lifetime~duration+treatment),"treatment"))
```



Conclusion from data example 1

Conclusion

- Stage of disease is the contributing factor toward lifetime – it really didn't have anything to do with the choice of treatment.
- It just happened that everyone on treatment 1 was in an earlier stage of the disease and so that made it look like there was a treatment effect.
- In fact, if we have to recommend a treatment, we might prefer Treatment 3 (although there is no sig. difference among any of the treatments).

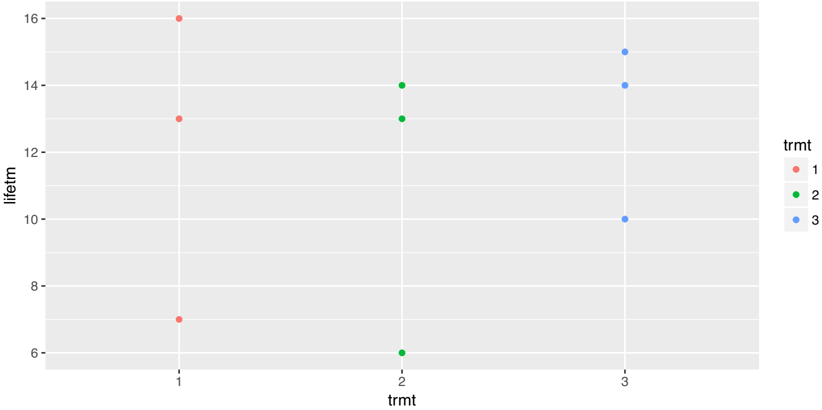
Motivation: data example 2

- In example 1, we see that failure to include an important covariate can yield significant treatment effect conclusion which is misleading.
- However, it is also possible to have a difference in means, but not be able to see it unless you first adjust for a covariate.
- Consider the same type of setting as where we want to test the effect of treatment on cancer, but with different data.

```
trmt=as.factor(c(1 ,1 ,1, 2, 2, 2, 3, 3, 3))
durationtm= c(1.2 ,1.9 ,3.4 ,5.6, 5.2, 5.8, 7.7, 8.3, 8.9)
lifetm=c(16, 13, 7, 14, 13 ,6 ,15, 14 ,10)
cancer2=data.frame(trmt,lifetm,durationtm)
```

Motivation: data example 2

```r
library(ggplot2)
ggplot(cancer2, aes(x=trmt,y=lifetm,col=trmt))+geom_point()
```



on avg, the survival time at each trmt roughly the same.

Motivation: data example 2

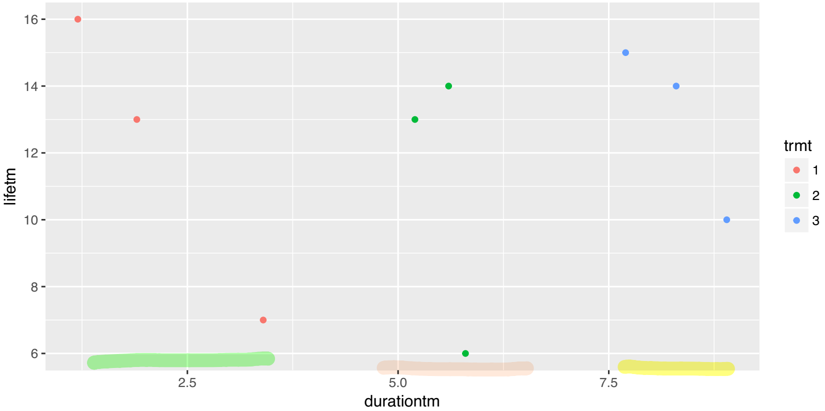
```
fit2=lm(lifetm~trmt,data=cancer2)
anova(fit2)
```

```
## Analysis of Variance Table
##
## Response: lifetm
##           Df Sum Sq Mean Sq F value Pr(>F)
## trmt       2      6   3.000  0.1915 0.8306
## Residuals  6     94  15.667
```

- No significant differences between the treatments, right?
- WRONG! The covariate (stage of disease) is omitted in the model:

Motivation: data example 2

```r
library(ggplot2)
ggplot(cancer2, aes(x=durationtm,y=lifetm,col=trmt))+geom_point()
```



- Again all taking Treatment 1 were in the early stages of the disease, all on Treatment 2 in the middle stages, and all on Treatment 3 in the latter stages.
- Is there a treatment effect? Surely Treatment 3 is better.

Motivation: data example 2

- Note that duration of the cancer by itself appears insignificant using type I SS
- We must realize that the duration of the cancer at time of treatment IS important and MUST be included in the model – or we get mistaken results. We must adjust for it before we can see the differences in treatments.

```
mod2=lm(lifetm~durationtm+trmt,data=cancer2)
anova(mod2)
```

```
## Analysis of Variance Table
##
## Response: lifetm
##            Df Sum Sq Mean Sq F value  Pr(>F)
## durationtm  1  0.770   0.770  0.1396 0.72402
## trmt        2 71.641  35.820  6.4917 0.04076 *
## Residuals   5 27.589   5.518
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Motivation: data example 2

- Note that these two examples were just for illustration – in reality, one should redesign this experiment and collect more data

Conclusion:

- The following output indicates that Treatment 3 is significantly better than the other two treatments.
- This time the potentially deadly mistake would be to assume based on a one-way ANOVA that the treatments were equivalent and use the cheapest one.

```
library(lsmeans)
leastsquare = lsmeans(mod2, "trmt",adjust="tukey")
leastsquare
```

$\tilde{Y}_{ik} = Y_{ik} - \sum_{j=1}^{p} \hat{\beta}_j X_{ijk}$

```
##  trmt    lsmean        SE df   lower.CL  upper.CL
##  1    -1.927184 4.237379  5 -12.819714  8.965345
##  2    11.879612 1.379701  5   8.332978 15.426245
##  3    26.047573 3.997993  5  15.770404 36.324741
##
## Confidence level used: 0.95
```

lsmean: the mean of each trmt is based on $\tilde{Y}_{ik}$ (adjusted mean)

One-way ANCOVA model

one factor

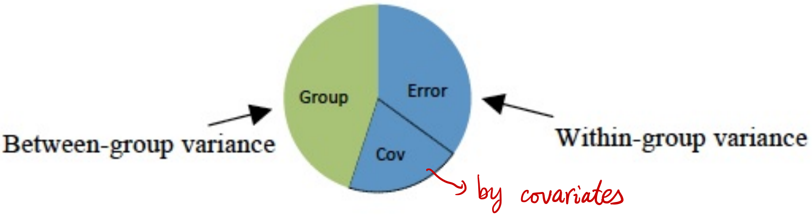
one non-factor variable

$$Y_{ij} = \underbrace{\mu + \alpha}_{}{}_i + \beta\underbrace{(X_{ij} - \bar{X}_{i.})}_{} + \epsilon_{ij}$$

- Y_{ij} is the j-th observation on response variable in the i-th group.
- X_{ij} is the j-th observation on the covariate in the i-th group.
- $i = 1, \ldots, a$: indicate levels (groups) of factor
- $j = 1, \ldots, n_i = r$: observations for level i has r replicates (balanced design)
- $\epsilon_{ij} \sim_{iid} N(0, \sigma^2)$
- Note that the difference $\alpha_i - \alpha_j$ does NOT dependent on the value of x.

Again to put constraints: $\sum_i \alpha_i = 0$ or $\alpha_1 = 0$ (R default)

Basic Ideas behind ANCOVA



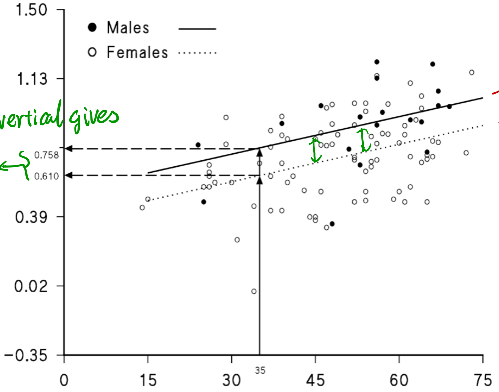
- To reduce the unexplained variation (MSE: residual mean square) in the data and increase the precision of the estimates of model parameters.
- To adjusted treatment means to a common value of the covariate x.
- Used in case that we are not able to use randomization of treatments as a way of controlling for a potential confounding variable.
- **Reference:** D.L. Clason, D.J. Mundfrom, *Adjusted Means in Analysis of Covariance: Are They Meaningful?*, Multiple Linear Regression Viewpoints, 2012, Vol. 38(1)

Assumptions of ANCOVA

1. Linearity of regression: Y and X must be linear.
2. Homogeneity of variance: equal variances for different treatment classes and observations.
3. Independence of error terms.
4. Normality of error terms. (assumptions 2-4:$\epsilon_{ik} \sim_{iid} N(0, \sigma^2)$)
5. Homogeneity of regression slopes: regression lines should be parallel among groups, i.e. the slope of the regression line is the same for all treatment combination subgroups (no interaction between covariate and factor).

We require the homogeneity of regression assumption, since it is important that treatments do not affect the value of the covariates,X, otherwise, comparison of treatment effects at a common X-value would not be meaningful.

Homogeneity of regression slopes



difference in vertical gives adjusted effect

slope for each sub groups are same for Y~X.
⇒ Same Reg slope assumption is met

- Ideally the covariate will not be in any way related to the treatment factors.
- $Y_{ij} \sim N(\mu + \alpha_i + \beta(X_{ij} - \bar{X}_{i.}), \sigma^2)$ and being independent.
  - For each i, we have a simple linear regression in which the slopes are the same, but the intercepts may differ

One-way ANCOVA: Least Squares Estimates

one-factor + p-covariate.

Note that (1) is easily extended to two-way ANOVA + p-covariates!

$$Y_{ik} = \mu + \alpha_i + \sum_{j=1}^{p} \beta_j x_{ijk} + \epsilon_{ik}, \tag{1}$$

where

- i for the treatment group, $i = 1, \ldots, I$,
- k for the replicates in each treatment group, $k = 1, \ldots, n_i$, assume balance, so $n_i = r$
- Y_{ik} is the k-th observation in i-th group.
- j for the j-th covariate X in.

Again, the least square estimates of μ, α and β may be derived via minimizing the sum of squared errors. As in the one-way ANOVA model, the corresponding normal equations are not linearly independent. To obtain the LSE easier, we rewrite model (1) as

$$Y_{ik} = \tau_i + \sum_{j=1}^{p} \beta_j (\underbrace{x_{ijk} - \bar{x}_{ij.}}_{\text{centre each level.}}) + \epsilon_{ik}, \tag{2}$$

where $\tau_i = \mu + \alpha_i + \sum_{j=1}^{p} \beta_j \bar{x}_{ij.}$

One-way ANCOVA: Least Squares Estimates

Model (2) in matrix form

$$y = A\tau + X\beta + \epsilon$$

Where $A_{n\times I}$ design matrix, and $X_{n\times p}$ design matrix associated with $(x_{ijk} - \bar{x}_{ij.})$. Since X was centered, $A'X = 0$. The LSE for τ, β is

$$\hat{\tau}_i = \bar{Y}_{i.}, \quad \hat{\beta} = (X'X)^{-1}X'Y$$

This leads to

$$\hat{\mu} + \hat{\alpha}_i = \bar{Y}_{i.} - \sum_{j=1}^{p} \beta_j \bar{x}_{ij.},\ i = 1, \ldots, I \tag{3}$$

It is clear that (3) has I+1 parameters but there are only I equations. There is no unique solution for $\hat{\mu}, \hat{\alpha}_i$. Impose a constraint on $\alpha_1 = 0$ (default in R):

$- \sum_{j=1}^{p} \hat{\beta}_j \bar{x}_{ij}$

$$\hat{\mu} = \bar{Y}_{1.} - \sum_{j=1}^{p} \beta_j \bar{x}_{1j.}, \quad \hat{\alpha}_i = \bar{Y}_{i.} - (\bar{Y}_{1.} - \sum_{j=1}^{p} \beta_j \bar{x}_{1j.})$$

Analysis of Covariance : no treatment effect testing

For one-way ANCOVA model (1), we want to test the null hypothesis:

$$H_0 : \underbrace{\alpha_i = 0, \forall i}_{\alpha_1=\alpha_2=\cdots=\alpha_I=0}, \quad H_a : \exists i, \alpha_i \neq 0$$

- Test statistics:

$$F = \frac{SS(T|X)/(I-1)}{SSE_F/(n-p-I)} \sim F_{I-1,n-p-I} \text{ under } H_0$$

- Decision: Reject H_0 if $F > F_{\alpha,I-1,n-p-I}$

SSE from full model (Y ~ X + T)

$$SSE_F = \sum_i \sum_k (y_{ik} - \hat{\mu} - \hat{\alpha}_i - \sum_{j=1}^{p} \hat{\beta}_j x_{ijk})^2 \sim \sigma^2 \chi^2_{n-p-I}$$

Reduced model (Y ~ X)

$$SSE_R^0(\beta) = \sum_i \sum_k (y_{ik} - \hat{\mu}^0 - \sum_{j=1}^{p} \hat{\beta}_j^0 \bar{x}_{ijk})^2$$

$$SS(T|X) = SSE_R^0(\beta) - SSE_F \sim \sigma^2 \chi^2_{I-1}$$

Analysis of Covariance : no covariate effect testing

For one-way ANCOVA model (1), we want to test the null hypothesis:

$$H_0 : \beta_j = 0, j = 1, \ldots, p \qquad H_a : \exists j, \beta_j \neq 0$$

- Test statistic:

$$F = \frac{SS(\beta|T)/p}{SSE_F/(n-p-I)} \sim F_{p,n-p-I} \text{ under } H_0$$

- Decision: reject H_0 if $F > F_{\alpha,p,n-p-I}$

Under no covariate effect null hypothesize, the model becomes the one-way ANOVA,

Reduced model: $Y \sim T$

$$SSE_R^0(T) = \sum_i \sum_k (Y_{ik} - \bar{Y}_{i.})^2 \sim \sigma^2 \chi^2_{n-I} \text{ under } H_0$$

Thus the sume of squares due to X: $Y \sim T$ $\quad$ $Y \sim X + T$

$$SS(\beta|T) = SSE_R^0(T) - SSE_F \sim \sigma^2 \chi^2_p$$

ANCOVA table with p Linear Covariates

Source	d.f.	SS	MS	F value
Covariate	p	$SS(\beta\|T)$	$SS(\beta\|T)/p$	$MS(\beta\|T)/MSE$
Treatment	I-1	SS_{trt}	$SS_{trt}/(I-1)$	MST/MSE
Error	n-I-p	SSE	SSE/(n-p-I)	
Total	n-1			

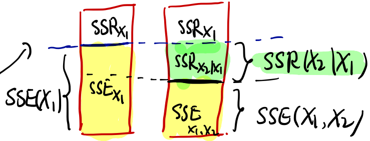
, n = total sample size.

Type I SS: $Y \sim x_1 + x_2 + x_3$

X_1 $SSR(X_1)$

x_2 $SSR(X_2|X_1)$ = Extra S.S. given X_1 is in model

X_3 $SSR(X_3|X_1, X_2)$ = Extra S.S. given X_1, X_2 already in model.



Type II/III SS: $Y \sim X_1 . X_2\ X_3$

X_1 $SSR(X_1|X_2, X_3)$

X_2 $SSR(X_2|X_1, X_3)$

X_3 $SSR(X_3|X_1, X_2)$

NOT sequential, it's all given the other Xs are already in the model.

ANCOVA data Example: Oyster

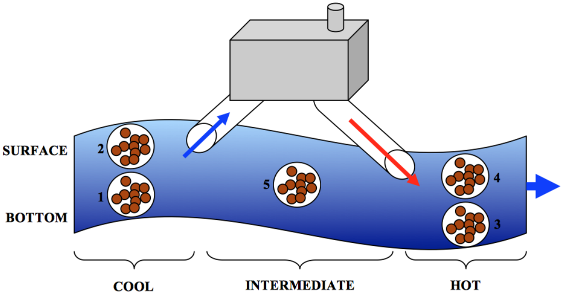
The objectives of a pilot experiment to study oyster growth:

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):

Treatment

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.

Oyster

```r
oyster=read.table("/Users/Wei/TA/Teaching/oyster.txt",sep="",header=T)
str(oyster)
```

```
'data.frame':   20 obs. of  4 variables:
 $ Trt    : int  1 1 1 1 2 2 2 2 3 3 ...
 $ Rep    : int  1 2 3 4 1 2 3 4 1 2 ...
 $ Initial: num  27.2 32 33 26.8 28.6 26.8 26.5 26.8 28.6 22.4 ...
 $ Final  : num  32.6 36.6 37.7 31 33.8 31.7 30.7 30.4 35.2 29.1 ...
```

X: Initial weight

Y: final weight.

```r
oyster$Trt=as.factor(oyster$Trt)
head(oyster)
```

```
  Trt Rep Initial Final
1   1   1    27.2  32.6
2   1   2    32.0  36.6
3   1   3    33.0  37.7
4   1   4    26.8  31.0
5   2   1    28.6  33.8
6   2   2    26.8  31.7
```

Oyster

Find the slope of X for each treatment level

```
Trtlevels<-c(1:5)
for (i in Trtlevels) {
  with(subset(oyster, Trt== Trtlevels[i]),
       { cat("Trt=",Trtlevels[i], "slope:\n")
         print(summary(lm(Final ~ Initial))$coef[2,],digits=4)
         cat("\n")})
  }
```

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	***

We see slope are different, but is this difference significant or not. we need to test it. (slide 37)

Oyster: testing same regression slope

Final ~ Initial + Trt + Initial * Trt .

```
test = lm(Final~Initial*Trt, data=oyster)
anova(test)
```

```
Analysis of Variance Table

Response: Final
            Df Sum Sq Mean Sq   F value    Pr(>F)
Initial      1 342.36  342.36 1208.0336 9.211e-12 ***
Trt          4  12.09    3.02   10.6645  0.001247 **
Initial:Trt  4   1.39    0.35    1.2247  0.360175
Residuals   10   2.83    0.28
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For each group

$Y = \beta_0 + \beta_i X + \varepsilon$

$H_0: \beta_1 = \beta_2 = \cdots = \beta_5$

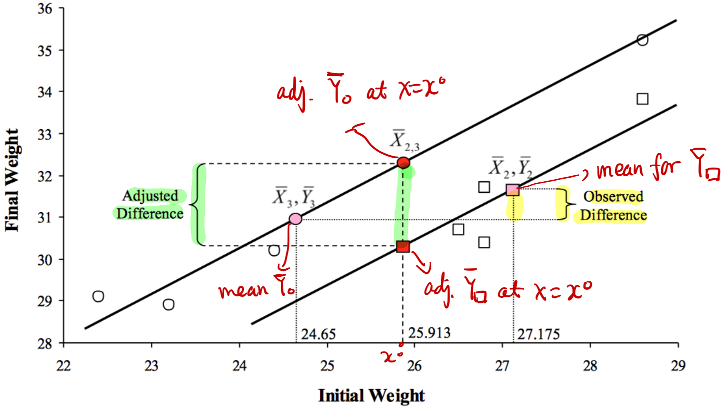
H_a: slopes are diff

No evidence to reject same slope assumpt.

So it's ok to assume the assumpt. is met.

Oyster: visualization of ANCOVA

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



- Comparing unadjusted(observed) means: $\bar{Y}_2 = 31.65 > \bar{Y}_3 = 30.85$
- Comparing adjusted(by initial weight) means: $\bar{Y}_2 = 30.12 < \bar{Y}_3 = 32.05$

Oyster: LS-square Mean - adjusted mean

```
unadjmeans <- aggregate(oyster$Final, list(oyster$Trt), mean)
unadjmeans
```

```
  Group.1      x
1       1 34.475
2       2 31.650
3       3 30.850
4       4 32.225
5       5 25.025
```

Raw mean for each level

```
library(lsmeans)
oyaoc2 <-lm(Final ~  Initial+ Trt, oyster)
oylsm <- lsmeans(oyaoc2, "Trt")
oylsm
```

```
 Trt   lsmean        SE df lower.CL upper.CL
 1   30.15311 0.3339174 14 29.43693 30.86929
 2   30.11730 0.2827350 14 29.51089 30.72371
 3   32.05233 0.2796295 14 31.45258 32.65208
 4   31.50469 0.2764082 14 30.91185 32.09752
 5   30.39757 0.3621988 14 29.62073 31.17441

Confidence level used: 0.95
```

adjusted mean for each level

Oyster: LS-square Mean - adjusted mean

TRT	Initial Means "X"	Unadjusted Means unadj $\bar{Y}_i$	Adjusted LS Means Adj. $\bar{Y}_i$	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 $\hat{\beta} = 1.08318$ represents a "best" single slope value that describes the relationship between X and Y, accounting for all other classification variables:

```
summary(lm(Final~Initial+Trt,data=oyster))$coef[,1:2]
```

```
               Estimate Std. Error
(Intercept)  2.25040039 1.44307538
Initial      1.08317982 0.04762051
Trt2        -0.03581197 0.40722674
Trt3         1.89921708 0.45801799
Trt4         1.35157290 0.41936648
Trt5         0.24445938 0.57658196
```

Oyster : ANCOVA

```
library(car)
oyaoc1 <-lm(Final ~ Trt + Initial, oyster)
anova(oyaoc1) # Type I SS
```

```
Analysis of Variance Table

Response: Final
          Df  Sum Sq Mean Sq F value    Pr(>F)
Trt        4 198.407  49.602  164.47 1.340e-11 ***
Initial    1 156.040 156.040  517.38 1.867e-12 ***
Residuals 14   4.222   0.302
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

← Wrong, undjusted by X. since use type I ss and Trt is first one to enter model.

```
Anova(oyaoc1, type = 2)
```

or type=3

```
Anova Table (Type II tests)

Response: Final
           Sum Sq Df F value    Pr(>F)
Trt        12.089  4  10.021 0.0004819 ***
Initial   156.040  1 517.384 1.867e-12 ***
Residuals   4.222 14
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(*) ← from oyac1

```
oyaoc2 <-lm(Final ~  Initial+ Trt, oyster)
anova(oyaoc2) # compare the Trt result with Type 2
```

(*) = (**) = same result.

```
Analysis of Variance Table

Response: Final
          Df Sum Sq Mean Sq  F value    Pr(>F)
Initial    1 342.36  342.36 1135.159 8.389e-15 ***
Trt        4  12.09    3.02   10.021 0.0004819 ***
Residuals 14   4.22    0.30
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(**) ← from oyac2

Take a break, and see you on Thursday

Weighted Least Squares

Review: Matrix Differenciation

$\rightarrow R^n \rightarrow R^m$

$$\underset{m\times 1}{Y} = f(X), \text{ where } x_{n\times 1}$$

Then the $m \times n$ matrix of first-order partial derivatives of the transformation from x to Y:

$$\boxed{\frac{\partial \mathbf{Y}}{\partial \mathbf{X}}} = \begin{bmatrix} \frac{\partial Y_1}{\partial \mathbf{X}} \\ \frac{\partial Y_2}{\partial \mathbf{X}} \\ \vdots \\ \frac{\partial Y_n}{\partial \mathbf{X}} \end{bmatrix} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}_{m\times n}$$

$$MD01: \underset{m\times 1}{Y} = \underset{m\times n}{A} \; \underset{n\times 1}{X} \Rightarrow \frac{\partial Y}{\partial X} = A$$

Proof:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = AX$$

$$\Rightarrow Y_i = \Sigma_{k=1}^{n} a_{ik} X_k$$

$$\Rightarrow \frac{\partial Y_i}{\partial X_j} = a_{ij} \quad \forall i = 1, 2, \cdots m \text{ and } j = 1, 2, \cdots n$$

$$\Rightarrow \frac{\partial Y_i}{\partial X} = [a_{i1}, a_{i2}, \cdots a_{in}] \quad \Leftarrow \frac{\partial Y_i}{\partial X} = \left(\frac{\partial Y_i}{\partial X_1}, \frac{\partial Y_i}{\partial X_2}, \cdots \frac{\partial Y_i}{\partial X_n} \right)$$

$$\Rightarrow \boxed{\frac{\partial Y}{\partial X} = A}$$

Q.E.D.

if define $\frac{\partial Y_i}{\partial X} = \begin{pmatrix} \partial Y_i / \partial X_1 \\ \partial Y_i / \partial X_2 \\ \vdots \\ \partial Y_i / \partial X_n \end{pmatrix}$

need to transpose all MD01–MD04

Review: Matrix Differenciation

1 2

$$MD02: \underset{1\times 1}{\alpha} = \underset{1\times m}{Y^T} \underset{m\times n}{A} \underset{n\times 1}{X} \Rightarrow \boxed{\frac{\partial \alpha}{\partial X} = Y^T A,} \text{ and } \boxed{\frac{\partial \alpha}{\partial Y} = X^T A^T}$$

Proof:

1 $W^T = Y^T A \Rightarrow \alpha = W^T X$

$$\Rightarrow \frac{\partial \alpha}{\partial X} = W^T \quad \text{by MD01}$$

$$= Y^T A$$

2 $\alpha_{1\times 1} \Rightarrow \alpha = \alpha^T = X^T A^T Y$

$$\Rightarrow \frac{\partial \alpha}{\partial Y} = X^T A^T$$

Q.E.D.

Review: Matrix Differenciation

For the special case in which the scalar α is given by the quadratic form

$$MD03: \underset{1\times 1}{\alpha} = \underset{1\times n}{X^T} \; \underset{n\times n}{A} \; \underset{n\times 1}{X} \Rightarrow \frac{\partial \alpha}{\partial X} = X^T(A + A^T)$$

Proof:

by def'n $\alpha = (x_1, x_2, \cdots x_n) \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$

$$= \sum_{j=1}^{n} \sum_{i=1}^{n} a_{ij} x_i x_j$$

$$\Rightarrow \frac{\partial \alpha}{\partial x_k} = \sum_{j=1}^{n} a_{kj} x_j + \sum_{i=1}^{n} a_{ik} x_i$$

$$= (x_1, x_2, \cdots x_n) \begin{pmatrix} a_{k1} \\ \vdots \\ a_{kn} \end{pmatrix} + (x_1, \cdots x_n) \begin{pmatrix} a_{1k} \\ a_{2k} \\ \vdots \\ a_{nk} \end{pmatrix}$$

$$= X^T A_{\text{row } k}^T + X^T A_{\text{column } k}, \quad \text{for } k = 1, 2, \cdots n$$

$$\Rightarrow \boxed{\frac{\partial \alpha}{\partial X} = X^T A^T + X^T A = X^T (A^T + A)}$$

Review: Matrix Differenciation

For the special case in which the scalar α is given by the quadratic form, and the A is a symmetric matrix, $\Rightarrow A = A^T$

$$MD04: \underset{1\times 1}{\alpha} = \underset{1\times n}{X^T} \underset{n\times n}{A} \underset{n\times 1}{X} \Rightarrow \frac{\partial \alpha}{\partial X} = 2X^T A$$

Proof:

From MD03 we have

$$\frac{\partial \alpha}{\partial X} = X^T(A + A^T)$$

$$= X^T(A + A) \quad \text{since } A = A^T$$

$$= 2X^T A$$

Q.E.D.

Review: Gradient

Gradient of a differentiable real function $f(x) : R^K \rightarrow R$ argument is defined uniquely in terms of partial derivatives

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_k} \end{pmatrix} \in R^k$$

while the second-order gradient of the twice differentiable real function with respect to its vector argument is traditionally called the Hessian;

$$\nabla^2 f(x) = \begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots \frac{\partial^2 f(x)}{\partial x_1 \partial x_k} & \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots \frac{\partial^2 f(x)}{\partial x_2 \partial x_k} & \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 f(x)}{\partial x_k \partial x_1} & \frac{\partial^2 f(x)}{\partial x_k \partial x_2} & \cdots \frac{\partial^2 f(x)}{\partial x_k^2} & \end{pmatrix} \in R^{k \times k}$$

Reference: `https://ccrma.stanford.edu/~dattorro/matrixcalc.pdf`

Review: Gradient

The gradient of vector-valued function $v(x) : R \to R^N$ on real domain is a row-vector

$$\nabla v(x) = \begin{pmatrix} \frac{\partial v(x)}{\partial x_1} & \frac{\partial v(x)}{\partial x_2} & \dots & \frac{\partial v(x)}{\partial x_k} \end{pmatrix} \in R^N$$

while the second-order gradient is

$$\nabla^2 v(x) = \begin{pmatrix} \frac{\partial^2 v(x)}{\partial x_1^2} & \frac{\partial v(x)}{\partial x_2^2} & \dots & \frac{\partial v(x)}{\partial x_k^2} \end{pmatrix} \in R^N$$

Gradient of vector-valued function $h(x) : R^k \to R^N$ on vector domain is

$$\nabla h(x) = \begin{pmatrix} \frac{\partial h_1(x)}{\partial x_1} & \frac{\partial h_2(x)}{\partial x_1} & \dots & \frac{\partial h_N(x)}{\partial x_1} & \\ \frac{\partial h_1(x)}{\partial x_2} & \frac{\partial h_2(x)}{\partial x_x} & \dots & \frac{\partial h_N(x)}{\partial x_x} & \\ \vdots & \vdots & & & \vdots \\ \frac{\partial h_1(x)}{\partial x_k} & \frac{\partial h_2(x)}{\partial x_k} & \dots & \frac{\partial h_N(x)}{\partial x_k} & \end{pmatrix}_{K \times N} = \begin{bmatrix} \nabla h_1(x) & \dots & \nabla h_N(x) \end{bmatrix}$$

Reference: `https://ccrma.stanford.edu/~dattorro/matrixcalc.pdf`

Heteroscedasticity

$$Y_{n\times1} = X_{n\times p}\beta_{p\times1} + \varepsilon_{n\times1}$$

$\begin{cases} \cdot \text{ transform } X : \text{ fix nonlinearity problem} \\ \cdot \text{ transform } Y : \text{ fix non-constant variance, and linearity too.} \end{cases}$

- Non-constant variance
- When the residual variance is not constant, a transformation might stabilize it
  - This will change the form of the relationship
  - You might not want to do that
- OLS estimates are still unbiased and consistent, but no longer have minimum variance $\longrightarrow$ BLUE : $Var(\hat{\beta}) = \sigma^2 (X^TX)^{-1}$ for same σ^2
- Could we just deal with the changing variance directly, without a transformation?

$$MSE = E((\hat{\theta}-\theta)^2) = E[(\hat{\theta}-E(\hat{\theta})+E(\hat{\theta})-\theta)^2]$$

$$= \underbrace{E((\hat{\theta}-E(\hat{\theta}))^2)} + \underbrace{E((E(\hat{\theta})-\theta)^2)} + 2(E(\hat{\theta})-\theta)\underbrace{E(\hat{\theta}-E(\hat{\theta}))}_{0}$$

$$= Var(\hat{\theta}) + Bias(\hat{\theta})^2$$

Know error variance

$$Var(\varepsilon) = \begin{pmatrix} \sigma_1^2 & & & 0 \\ & \sigma_2^2 & & \\ & & \ddots & \\ 0 & & & \sigma_n^2 \end{pmatrix}$$

$$W^{\frac{1}{2}} = \begin{pmatrix} 1/\sigma_1 & & & 0 \\ & 1/\sigma_2 & & \\ & & \ddots & \\ 0 & & & 1/\sigma_n \end{pmatrix}$$

- Let $Y = X\beta + \epsilon$
  - Y is a vector of random variable
  - X is a matrix of known fixed constant
  - β is an unknown vector of fixed but parameters
  - ϵ is an unknown vector of random errors or random noises
    - $\epsilon_i \sim_\perp N(0, \sigma_i^2)$

$$\Rightarrow Var(W^{\frac{1}{2}}\varepsilon) = W^{\frac{1}{2}} var(\varepsilon)(W^{\frac{1}{2}})^T$$
$$= I$$

- If we define the reciprocal of each variance, σ_i^2, as the weight, $w_i = 1/\sigma_i^2$, and let W be a diagonal matrix containing these weights
  - Instead of minimizing the Sum of squared residuals, $Q = \sum_i e_i^2 = e'e$
  - We minimize the **weighted (SS) of residual** instead, $Q_w = \sum_i w_i e_i^2 = e'We$
  - Resulting parameter
    - $b_w = (X'WX)^{-1}X'WY$
    - $Var(b_w) = (X'WX)^{-1}$

$$W^{\frac{1}{2}}Y = W^{\frac{1}{2}}X\beta + W^{\frac{1}{2}}\varepsilon$$
$$Y^* = X^*\beta + \varepsilon^*$$
$$\hat{\beta} = (X^{*\prime}X^*)^{-1}X^{*\prime}Y^*$$

$$b_w = \hat{\beta}_w = (X'WX)^{-1}X'WY \leftarrow = (X'W^{\frac{1}{2}}W^{\frac{1}{2}}X)^{-1}XW^{\frac{1}{2}}W^{\frac{1}{2}}Y$$

Unknow error variance

However, $\sigma_1^2, \ldots, \sigma_n^2$ are almost always unknown. Way to deal with it

- If you have replicates at various levels of X, you can just estimate the error variances by the sample variance at each level.
- If you don't have replicates, you can try to estimate how the variance (or SD) is changing as a function of X.

1. Regress Y against preditor variable(s) as usual (OLS) & obtain residual e_i and $\hat{Y}_i$ $\quad e_i^2 \sim \sigma_i^2$, $|e_i| \sim \sigma_i$, $|e_i|$ works better in presence of outliers.
2. Regress $|e_i|$ against predictors $x_1, \ldots, x_k$ or fitted value $\hat{Y}_i$
3. Let $\hat{s}_i$ be the fitted value for the regression in 2
4. Define $w_i = 1/\hat{s}_i^2$ for i=1,...,n
5. Use $b_w = (X'WX)^{-1}X'WY$ as estimated coefficient.

$$W = \begin{pmatrix} w_1 & & 0 \\ & \ddots & \\ 0 & & w_n \end{pmatrix}$$

Weighted Least Square

$W = var(e)^{-1} = \Sigma^{-1} = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{pmatrix}^{-1}$

Show $Q_w = \sum_i w_i e_i^2 = e' W e, w_i = 1/\sigma_i^2$

Proof:

$$Q_W = \sum_{i=1}^n (\sqrt{w_i}\, e_i)^2$$

$$= (\sqrt{w_1} e_1 \quad \sqrt{w_2} e_2 \quad \cdots \quad \sqrt{w_n} e_n) \begin{pmatrix} \sqrt{w_1} e_1 \\ \sqrt{w_2} e_2 \\ \vdots \\ \sqrt{w_n} e_n \end{pmatrix}$$

$$= \left[\begin{pmatrix} \sqrt{w_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{w_n} \end{pmatrix}_{n \times n} \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}_{n \times 1} \right]^T \left[\underbrace{\begin{pmatrix} \sqrt{w_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{w_n} \end{pmatrix}}_{W^{\frac{1}{2}}} \underbrace{\begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}}_{e} \right]_{1 \times n}$$

$$= e' W^{\frac{1}{2}} W^{\frac{1}{2}} e$$

$$= e' W e \qquad \text{Q.E.D.}$$

Weighted Least Square in general

Symmetric, since it is var-cov matrix.

In general, we assume that $Var(\epsilon) = \sigma^2\Sigma$, w.l.o.g, $\sigma^2 = 1$

Show $b_w = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y$

Proof:

We want to minimize $Q = e'\Sigma^{-1}e$

$$= (Y - X\beta)^T \Sigma^{-1} (Y - X\beta)$$

$$(X^T\Sigma^{-1}X)^T = X^T\Sigma^{-1}X$$

$$= (Y^T - \beta^T X^T)\Sigma^{-1}(Y - X\beta)$$

$$= Y^T\Sigma^{-1}Y - Y^T\Sigma^{-1}X\beta - \beta^T X^T\Sigma^{-1}Y + \beta^T \underline{X^T\Sigma^{-1}X}\beta$$

$$\frac{\partial Q}{\partial \beta} = -Y^T\Sigma^{-1}\overset{MD02}{X - (X^T\Sigma^{-1}Y)'} + 2\underline{\beta^T X^T\Sigma^{-1}X} \quad MD04$$

$$= -2Y^T\Sigma^{-1}X + 2\beta^T(X^T\Sigma^{-1}X)$$

$$\Rightarrow \quad X^T\Sigma^{-1}Y = (X^T\Sigma^{-1}X)\beta$$

$$\Rightarrow \quad \hat{\beta}_w = b_w = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y$$

Weighted Least Square in general

In general, we assume that $Var(\epsilon) = \sigma^2\Sigma$

Show $E(b_w) = \beta$, $Var(b_w) = \sigma^2(X'\Sigma^{-1}X)^{-1}$

Proof:

$$b_w = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y, \quad Y = X\beta + \varepsilon, \quad \underline{E(\varepsilon) = 0, \; Var(\varepsilon) = \sigma^2\Sigma}$$

- $$E(b_w) = E[(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y)$$
$$= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}E(Y)$$
$$= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}X\beta$$
$$= \beta$$

$$A' = \Sigma^{-1}X(X'\Sigma^{-1}X)^{-1}$$

- $$Var(b_w) = Var(AY) = A\,Var(Y)A^T, \quad A = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}$$
$$= A\sigma^2\Sigma A^T = \sigma^2 \underbrace{(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}}_{A}\Sigma\underbrace{\Sigma^{-1}X(X'\Sigma^{-1}X)^{-1}}_{A^T}$$
$$= \sigma^2(X'\Sigma^{-1}X)^{-1}$$

WLS data example: Computer-Assisted Learning Dataset

Data:

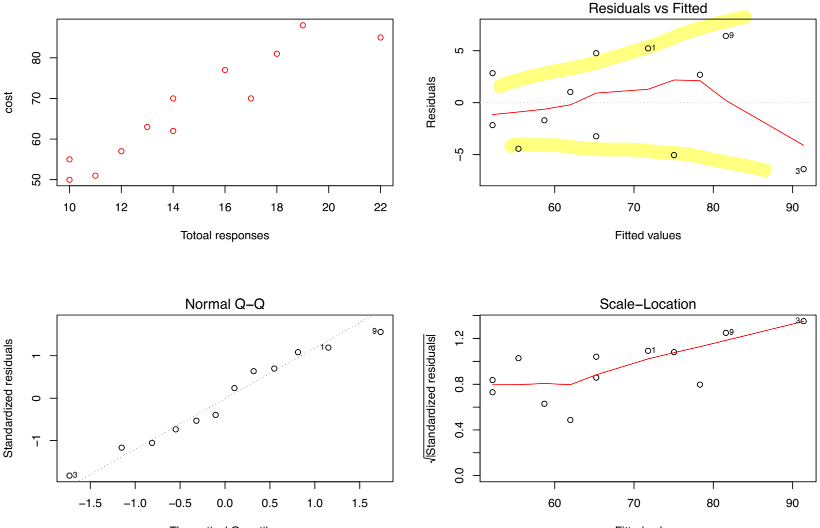
- cost (Y): the cost of the computer time (Y)
- totresp (X): the predictor is the total number of responses in completing a lesson (X).

```
cost = c(77, 70, 85, 50, 62, 70, 55, 63, 88, 57, 81, 51)
totresp =c(16, 14, 22, 10, 14, 17, 10, 13, 19, 12, 18, 11)
ols = lm(cost~totresp)
summary(ols)
```

```
Call:
lm(formula = cost ~ totresp)

Residuals:
   Min     1Q Median     3Q    Max
-6.389 -3.536 -0.334  3.319  6.418

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.4727     5.5162   3.530  0.00545 **
totresp       3.2689     0.3651   8.955 4.33e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

WLS data example: Computer-Assisted Learning Dataset

```
par(mfrow=c(2,2))
plot(totresp,cost,type="p",col="red",xlab="Totoal responses")
plot(ols,which=c(2,1,3))
```



WLS data example: Computer-Assisted Learning Dataset

The weights we will use will be based on regressing the absolute residuals from OLS versus the predictor.

```
# estimate weight
absresid = abs(ols$residuals)
# |e|~ x
reg =lm(absresid~totresp)
wt=1/(reg$fitted.values^2)
wls = lm(cost~totresp,weights=wt)
summary(wls)
```

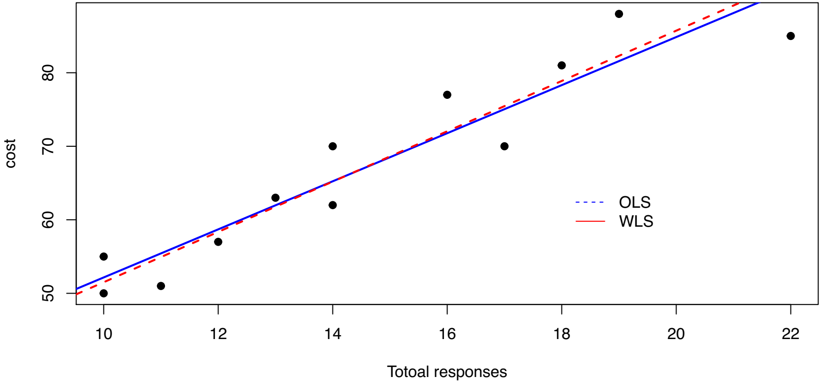
	slope	
OLS	3.2689	close
WLS	3.4211	

```
Call:
lm(formula = cost ~ totresp, weights = wt)

Weighted Residuals:
     Min       1Q   Median       3Q      Max
-1.48741 -0.96167 -0.04198  1.10930  1.50265

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.3006     4.8277   3.584  0.00498 **
totresp       3.4211     0.3703   9.238 3.27e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

WLS data example: Computer-Assisted Learning Dataset

The weights we will use will be based on regressing the absolute residuals from OLS versus the predictor.

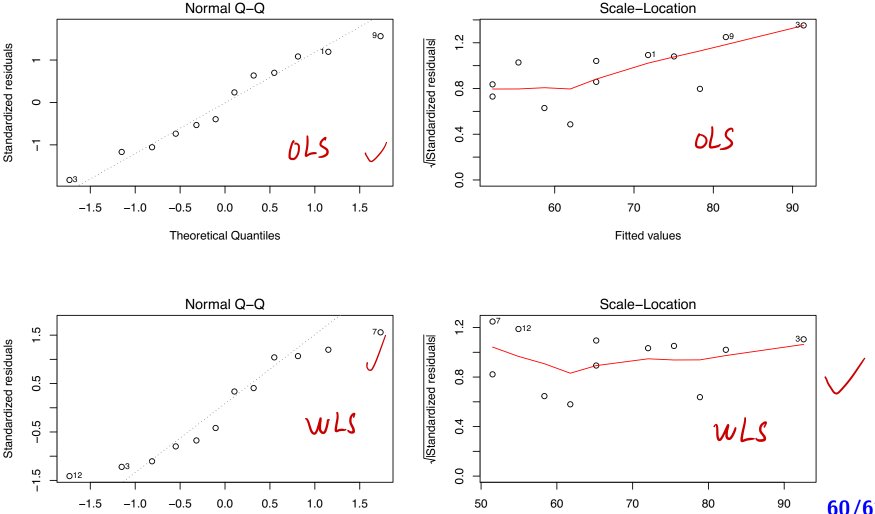
```r
plot(totresp,cost,type="p",col="black",xlab="Totoal responses", pch=19)
abline(ols,col="blue",lty=1,lwd=2)
abline(wls,col="red",lty=2,lwd=2)
legend(18,65,c("OLS","WLS"),col=c("blue","red"),lty=c(2,1),bty="n")
```



WLS data example: Computer-Assisted Learning Dataset

Compare the residual plots

```r
par(mfrow=c(2,2))
plot(ols,which=c(2,3))
plot(wls,which=c(2,3))
```



Some key points regarding weighted least squares are:

How to obtain W? Key to WLS

- The difficulty, in practice, is determining estimates of the error variances (or standard deviations).
- Weighted least squares estimates of the coefficients will usually be nearly the same as the "ordinary" unweighted estimates. In cases where they differ substantially, the procedure can be iterated until estimated coefficients stabilize (often in no more than one or two iterations); this is called **iteratively reweighted least squares**.
- In some cases, the values of the weights may be based on theory or prior research.
- In designed experiments with large numbers of replicates, weights can be estimated directly from sample variances of the response variable at each combination of predictor variables.
- Use of weights will (legitimately) impact the widths of statistical intervals.

After Lecture This Week

Practice problems

- ANCOVA: CH22 in KNNL book, CH4 in Appl. Stat By Fan.
- Weighted Least Square: CH 18.4 in KNNL book. Section 1.4 in Appl. Stat By Fan
- Try all the R example in slides.
- Review all types of ANOVA and ANCOVA analysis.

Topics for next week:

- Ridge regression
- Logistic regression model