STAT 3119

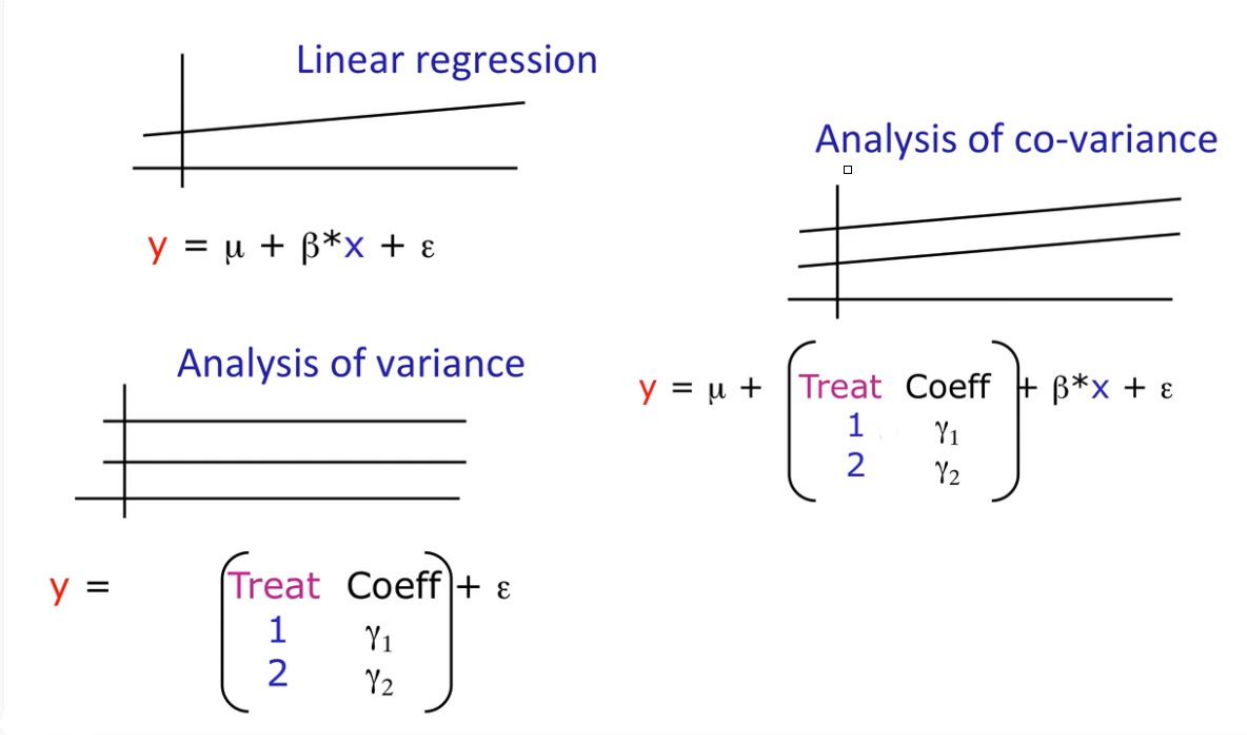
Week 10: 10/29/2019 @GWU

Outline

- Analysis of Covariance (ANCOVA) (Chapter 22)
- Single factor ANCOVA model
- Estimation and inference
- Data example with ANCOVA analysis and model diagnostics
- Extension of single-factor ANCOVA

Introduction

Analysis of covariance (ANCOVA) is a technique that combines features of analysis of variance and regression. The idea is to augment the analysis of variance model containing the factor effects (categorical variable) with **one or more additional quantitative (continuous) variables** that are related to the response variable.



The objectives of ANCOVA

- To reduce the variance of the error terms in the model, which will improve the efficiency of the analysis and make the estimate of the treatment effects more precise.

- By controlling the known related covariates , we can gain greater insight into the effects of the treatment factor on the response, e.g. a clinical trial to study drug to treat hypertension, with baseline BP taken into account.

Choices of covariates (concomitant variables) in ANCOVA

- We consider **continuous covariates related to the response**. If other categorical variables are considered, we will use the regular two-factor or multi-factor ANOVA model. Covariates frequently considered:
  - **with human subjects**: pre-study or baseline information, such as baseline age, lab values and socioeconomic status before the clinical treatments.
  - **with retail stores/sites as study units**: last period's sales or number of employees.
- **Covariates Unaffected by Treatments**: For a clear interpretation of the results, a covariate should be observed before the study; or if observed during the study, it should NOT be influenced by the treatments.
  - Although the interactions of treatments and covariates can be adjusted in the statistical model, it will make the interpretation of treatment effects (on the response) more difficult.

Example of Single-factor ANCOVA (Ch 22.3)

Example(page 926): A company studied the effects of **three different types of promotions** on *sales of its crackers*:

- Treatment 1: Sampling of product by customers in store and regular shelf space
- Treatment 2: Additional shelf space in regular location
- Treatment 3: Special display shelves at ends of aisle in addition to regular shelf space
- **Fifteen stores** were selected for the study, and a **completely randomized design (CRD)** was utilized. Each store was randomly assigned one of the promotion types, with five stores assigned to each type of promotion.
- Other relevant conditions under the control of the company, such as price and advertising, were kept the same for all stores in the study.
- Data show on the **number of cases of the product sold during the promotional period (Y)**, and the **sales of the product in the preceding period (X- covariate of interest)**.

TABLE 22.1 Data—Cracker Promotion Example (number of cases sold).

Treatment	Store (j)									
	1		2		3		4		5	
i	Y_{i1}	X_{i1}	Y_{i2}	X_{i2}	Y_{i3}	X_{i3}	Y_{i4}	X_{i4}	Y_{i5}	X_{i5}
1	38	21	39	26	36	22	45	28	33	19
2	43	34	38	26	38	29	27	18	34	25
3	24	23	32	29	31	30	21	16	28	29

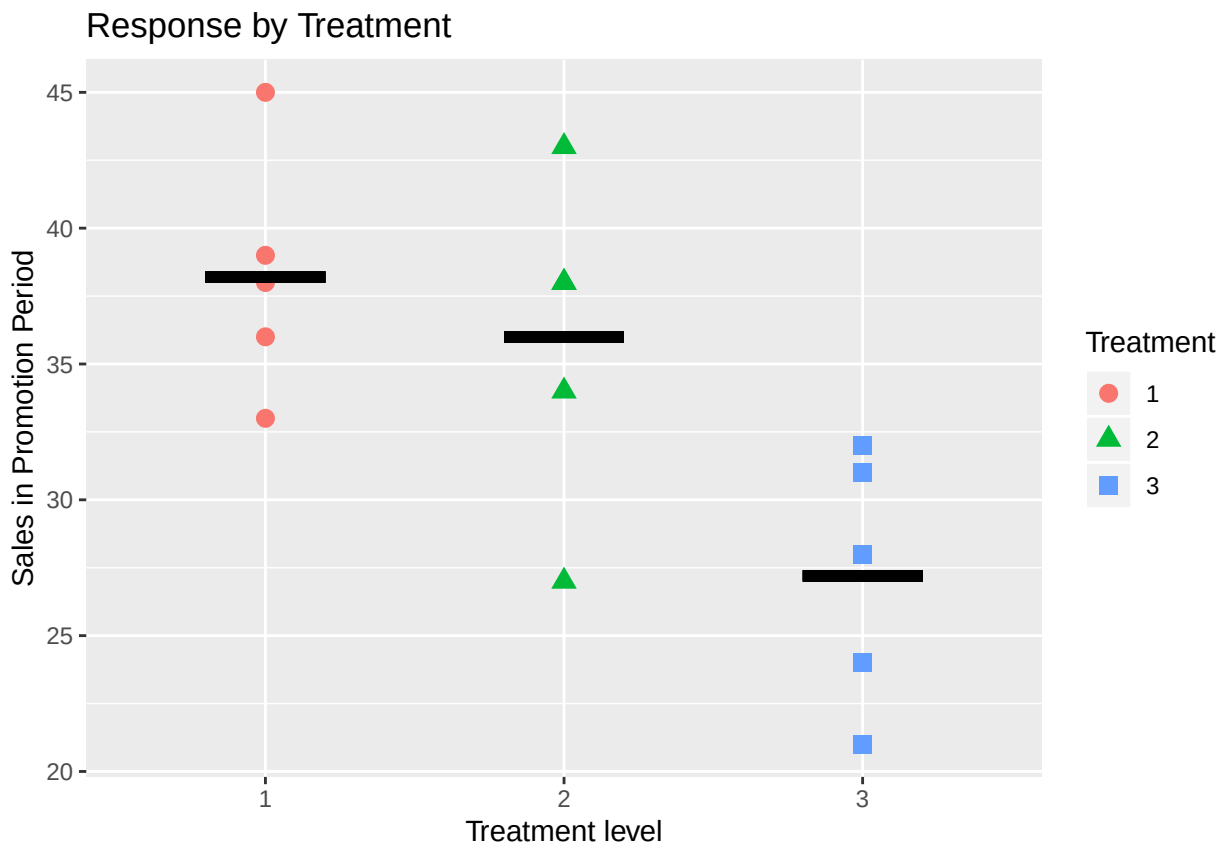
R analysis: Data checking

1. Read the data

```
# read data from week5 folder online
Ex22  =read.table(
        url("https://raw.githubusercontent.com/npmldabook/Stat3119/master/Week8/CH22TA01.txt"))

names(Ex22) =  c("New_Sales", "Old_Sales", "Treatment","Units")

# make the treatment as a factor variable
Ex22$Treatment =  as.factor(Ex22$Treatment)
```

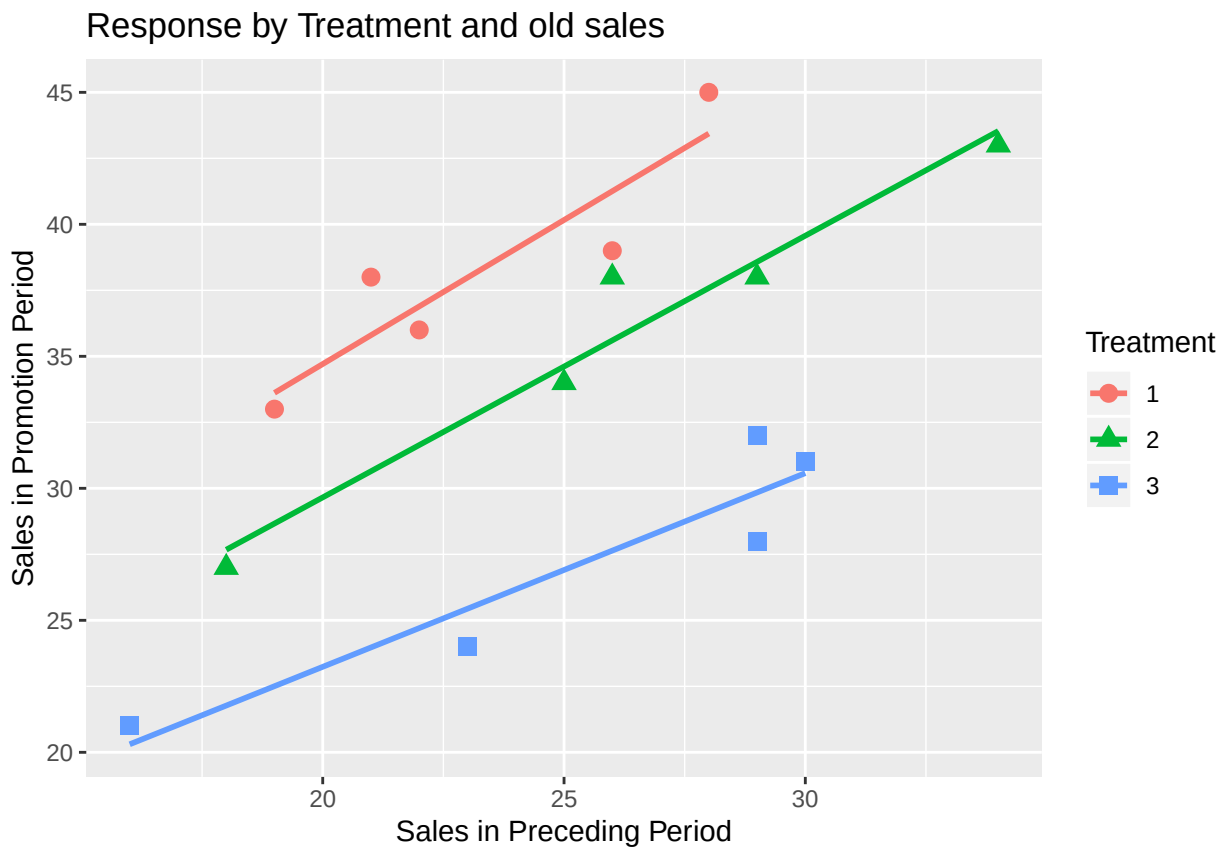
2A. Plot the response by treatment/factor levels

```
library(ggplot2)

Trt.means= as.numeric(with(Ex22, by(New_Sales, Treatment, mean) ))

ggplot(data=Ex22, aes(x= Treatment, y= New_Sales,color= Treatment, shape=  Treatment))+
  geom_point(size=3) +
  geom_segment(aes(x = 0.8, y = Trt.means[1], xend = 1.2, yend = Trt.means[1]),size=2, col=1)+
  geom_segment(aes(x = 1.8, y = Trt.means[2], xend = 2.2, yend = Trt.means[2]),size=2, col=1)+
  geom_segment(aes(x = 2.8, y = Trt.means[3], xend = 3.2, yend = Trt.means[3]),size=2, col=1)+
  labs(title="Response by Treatment",
       x="Treatment level", y="Sales in Promotion Period")
```



2B. Plot the response by previous sale X and group by treatment, with an added LS line per group.

```
ggplot(Ex22, aes(x= Old_Sales, y= New_Sales, color= Treatment, shape=  Treatment ))+
  geom_point(size=3) +
  geom_smooth(method = lm, se= F)+
  labs(title="Response by Treatment and old sales",
       x="Sales in Preceding Period", y="Sales in Promotion Period")
```



Findings:

1) It seems the previous sale X is linearly related to the response for each treatment group, with similar slopes among different treatment levels (i.e., no evidence of interaction between previous sale and treatments).

2) The points around the within-treatment regression lines are less than the scatter around the treatment means (the first figure). This shows that considering this covariate (previous sale X) in a model will reduce the residual error variability and make the analysis more efficient to estimate treatment difference

Single-Factor Covariance Model (Ch 22.2)

Notation:

- n_i= the number of cases for the ith factor level ($i = 1, ..., r$).
- $n_T = \sum n_i$ = the total number of cases (total sample size).
- Y_{ij} = the jth observation on the response variable for the ith factor level.
- X_{ij} = the covariate value associated with the jth case for the ith factor level.

The usual **ANCOVA model for a single-factor study** with r fixed levels (22.3) is

$$Y_{ij} = \mu_. + \tau_i + \gamma(X_{ij} - \overline{X}_{..}) + \epsilon_{ij} \qquad (22.3)$$

where

- $\mu_.$ = an overall mean
- τ_i = the fixed treatment effects subject to the restriction $\sum \tau_i = 0$.
- γ = the regression coefficient for the relation between Y and X.
- Error term ϵ_{ij} = independent $N(0, \sigma^2)$, $i = 1, ..., r; j = 1, ..., n_i$.

Notes:

1. If we just plug in the observation X_{ij} without centering around its overall mean $\overline{X}_{..}$, then regression coefficient γ is still the same to reflect the linear relationship of Y and X, but $\mu_.$ is no longer the overall mean and it might not be interpretable when $X_{ij} = 0$ (Some lab values such as BP are always positive).
2. After applying the centering to X_{ij} by subtracting its overall mean $\overline{X}_{..}$, it then follows

$$E(Y_{ij}) = \mu_{ij} = \mu_. + \tau_i + \gamma(X_{ij} - \overline{X}_{..})$$

$$\text{var}(Y_{ij}) = \sigma^2$$

Then, $\mu_. + \tau_i$ represents the mean of response Y for an "average" value of the covariate given the factor (treatment) level i.

3. We consider here the covariates X_{ij} are given observations. Because of i.i.d error term ϵ_{ij}, we can also state the ANCOVA model as the response variable

$$Y_{ij} \text{ are independent } \sim N(\mu_{ij}, \sigma^2)$$

where $\mu_{ij} = \mu_. + \tau_i + \gamma(X_{ij} - \overline{X}_{..})$ with $\sum \tau_i = 0$.

Assumptions for ANCOVA Model

I. Regular ANOVA assumptions

- Independent, normally distributed errors
- Homogeneity of variances among treatment levels

II. Additional assumptions for ANCOVA

- **Linearity**: For each treatment i, the expected response, μ_{ij}, is given by a regression line.

$$\mu_{ij} = \mu_. + \tau_i + \gamma(X_{ij} - \overline{X}_{..})$$

- **Constancy (Homogeneity) of the slope**: all treatment regression lines have the same slope.

Thus, τ_i is no longer the main effect for treatment i since the mean response varies wtih X, but the comparison between any two treatment levels i and i' for the same covariate value X is given by $D = \tau_i - \tau_i'$, which is also the vertical distance between the two corresponding parallel regression lines.

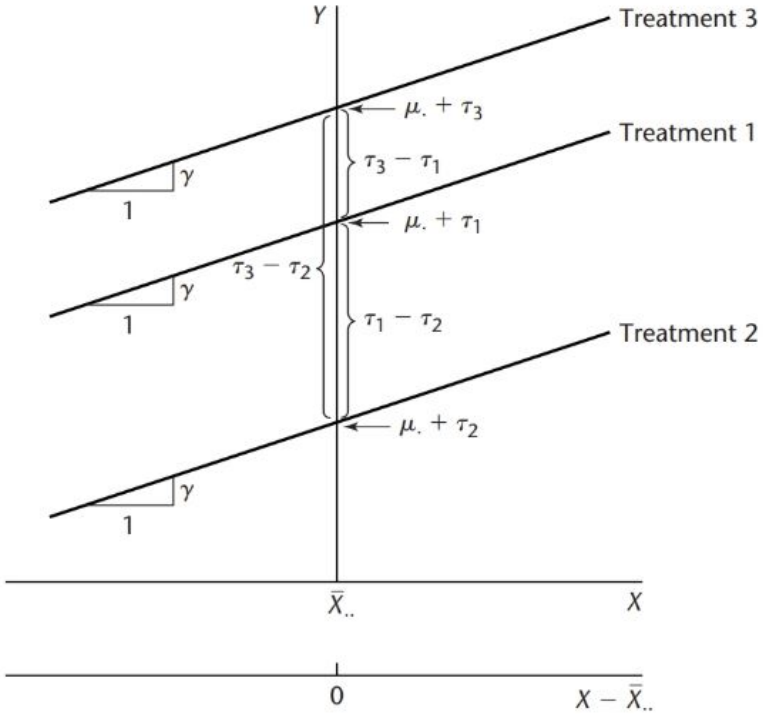


FIGURE 22.3
Example of Treatment Regression Lines with Covariance Model (22.3).

Estimation of ANCOVA model parameters

- An easy way to estimate the model parameters and make inferences is through the **regression** approach.
- The same regression technique based on matrix calculation can be applied here given the specific design matrix for the ANCOVA model.
- Most statistical software such as R and SAS can be used for ANCOVA and regression analysis.

Regression Formulation: We can express the ANCOVA model (22.3) as follows

$$Y_{ij} = \mu_. + \tau_1 I_{ij1} + ... + \tau_{r-1} I_{ij,r-1} + \gamma x_{ij} + \epsilon_{ij}$$

where

- $x_{ij} = X_{ij} \overline{X}_{..}$, centered observations
- we use $r-1$ indicator functions to represent the r treatment levels, then the treatment effects $\tau_1, ..., \tau_{r-1}$ are exactly the regression coefficients for the corresponding indicator variables. We don't need the indicator for the rth level because $\tau_r = -(\tau_1 + ..\tau_{r-1})$, due to the zero-sum constraint.

$$I_1 = \begin{cases} 1 & \text{if case from treatment 1} \\ -1 & \text{if case from treatment } r \\ 0 & \text{otherwise} \end{cases}$$

$$\vdots$$

$$I_{r-1} = \begin{cases} 1 & \text{if case from treatment } r-1 \\ -1 & \text{if case from treatment } r \\ 0 & \text{otherwise} \end{cases}$$

- Consequently, we can fit this regression model to get the coefficent estimates (i.e., the estimates of the ANCOVA model parameters).

Inferences of ANCOVA analysis

- The **key** statistical inferences of interest in ANCOVA are the same as with ANOVA models, i.e., whether the treatments have any effects on the response, and if so what these effects are.
- **Testing**: For fixed treatment effects, it is the same as the hypothesis testing in ANOVA:

$$H_0 : \tau_1 = \tau_2 = ... = \tau_r = 0 \quad vs.$$

$$H_a : \text{not all} \quad \tau_i \quad \text{equal zero}$$

  We can test these hypotheses as testing whether several regression coefficients equal zero in its equivalent regression model.

- **Further estimation**: If we reject the above null and find the treatment effects differ, the next step is to investigate the nature of these effects, such as, pairwise comparisons , some contrasts, or linear combinations of of treatment effects. Similar multiple comparison procedure may apply, as appropriate.
- Occasionally, the nature of the relationship between Y and X is of interest, for which we can make inference on the regrression coefficient γ.

Crakers Example: ANCOVA analysis

- 3 treatment levels, 15 stores
- Y= number of cases of the product sold during the promotional period
- X= the sales of the product in the preceding period

TABLE 22.1 Data—Cracker Promotion Example (number of cases sold).

Treatment	Store (j)									
	1		2		3		4		5	
i	Y_{i1}	X_{i1}	Y_{i2}	X_{i2}	Y_{i3}	X_{i3}	Y_{i4}	X_{i4}	Y_{i5}	X_{i5}
1	38	21	39	26	36	22	45	28	33	19
2	43	34	38	26	38	29	27	18	34	25
3	24	23	32	29	31	30	21	16	28	29

We set the regression model below:

$$Y_{ij} = \mu_. + \tau_1 I_{ij1} + \tau_2 I_{ij2} + \gamma x_{ij} + \varepsilon_{ij} \qquad \text{Full model} \qquad \textbf{(22.13)}$$

where:

$$I_1 = \begin{cases} 1 & \text{if store received treatment 1} \\ -1 & \text{if store received treatment 3} \\ 0 & \text{otherwise} \end{cases}$$

$$I_2 = \begin{cases} 1 & \text{if store received treatment 2} \\ -1 & \text{if store received treatment 3} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{ij} = X_{ij} - \bar{X}_{..}$$

R implementation

1. generate the I_j and centered X and fit the regression model

```
Indicator1 = (Ex22$Treatment=="1")*1 + (Ex22$Treatment=="3")*(-1)
Indicator2 = (Ex22$Treatment=="2")*1 + (Ex22$Treatment=="3")*(-1)

# center the observation
(meanX= mean( Ex22$Old_Sales))
```

```
## [1] 25
```

```
X.centered = Ex22$Old_Sales - meanX

LM.full  =  lm( New_Sales~ Indicator1 + Indicator2 + X.centered, data=Ex22 )
```

2. generate model parameters

```
# regression overall summary
summary(LM.full)
```

```
##
## Call:
## lm(formula = New_Sales ~ Indicator1 + Indicator2 + X.centered,
##     data = Ex22)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4348 -1.2739 -0.3362  1.6710  2.4869
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.8000     0.4835  69.908 6.37e-16 ***
## Indicator1    6.0174     0.7083   8.496 3.67e-06 ***
## Indicator2    0.9420     0.6987   1.348    0.205
## X.centered    0.8986     0.1026   8.759 2.73e-06 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.873 on 11 degrees of freedom
## Multiple R-squared:  0.9403, Adjusted R-squared:  0.9241
## F-statistic: 57.78 on 3 and 11 DF,  p-value: 5.082e-07
```

```
# only the point estimates, mu, tau_i and gamma
(Coef=coef(LM.full))
```

```
## (Intercept)  Indicator1  Indicator2  X.centered
##  33.8000000   6.0174070   0.9420168   0.8985594
```

```
# Get the 95% confidence interval for each model parameter
confint(LM.full)
```

```
##                  2.5 %    97.5 %
## (Intercept) 32.7358390 34.864161
## Indicator1   4.4585443  7.576270
## Indicator2  -0.5957732  2.479807
## X.centered   0.6727716  1.124347
```

Example: To test any treatment effect, $\tau_1 = \tau_2 = ... = \tau_r = 0$.

We will compare the previous full model with the reduced model under H_0:

$$Y_{ij} = \mu_{.} + \gamma x_{ij} + \epsilon_{ij}$$

```
#Reduced model
LM.reduced =  lm( New_Sales~   X.centered, data=Ex22 )

#compare the two models
anova(LM.reduced, LM.full)
```

```
## Analysis of Variance Table
##
## Model 1: New_Sales ~ X.centered
## Model 2: New_Sales ~ Indicator1 + Indicator2 + X.centered
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     13 455.72
## 2     11  38.57  2    417.15 59.483 1.264e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Results: By comparing the two model RSS, we obtain the F-test statistic= 49.48 , following a F-distribution $F(df_1 = 2, df_2 = 11)$. The p-value <0.0001, therefore we reject the null and conclude that the three cracker promotions (treatments) differ in sales effectiveness.

Futher inference on treatment effects (p. 930-930)

In general, using regression technique

1. **Point Estimate**: From the full regression model, we can obtain the parameter estimate for the overall mean, $\hat{\mu}_.$ and the two treatment effects $\hat{\tau}_1$ and $\hat{\tau}_2$. Then we can also get the estimate for any linear combination of $L = c_0\mu_. + c_1\tau_1 + c_2\tau_2 + c_3\tau_3$ (including the pairwise difference or any contrasts) as

$$\hat{L} = c_0\hat{\mu}_. + c_1\hat{\tau}_1 + c_2\hat{\tau}_2 + c_3\hat{\tau}_3 = c_0\hat{\mu}_. + (c_1 - c_3)\hat{\tau}_1 + (c_2 - c_3)\hat{\tau}_2$$

since $\hat{\tau}_3 = -(\hat{\tau}_1 + \hat{\tau}_2)$ from the zero-sum constraint.

2. **Variance of $\hat{L}$**: we write $\hat{L} = \mathbf{c^T}\hat{\theta}$, where $c = (c_0, c_1 - c_3, c_2 - c_3)^T$ is the constant coefficient vector and $\hat{\theta} = (\hat{\mu}_., \hat{\tau}_1, \hat{\tau}_2)^T$, then the variance of $\hat{L}$ is

$$\sigma^2(\hat{L}) = var(\hat{L}) = \mathbf{c^T}\Sigma\mathbf{c}$$

where Σ is the 3 by 3 variance-covariance matrix of $\hat{\theta}$, which can be estimated and obtained from the regression output.

R analysis

```
# The parameter estimate for mu, tau1 and tau2
(Coef=coef(LM.full))
```

```
## (Intercept)  Indicator1  Indicator2  X.centered
##  33.8000000   6.0174070   0.9420168   0.8985594
```

```
data.frame(mu=Coef[1],tau1= Coef[2], tau2= Coef[3],
           tau3=-Coef[2]-Coef[3])
```

```
##                mu     tau1      tau2      tau3
## (Intercept) 33.8 6.017407 0.9420168 -6.959424
```

```
# First we get the variance-covariance matrix of coefficient estimate 4 by 4
 vcov(LM.full)
```

```
##             (Intercept)  Indicator1  Indicator2  X.centered
## (Intercept)   0.2337655  0.00000000  0.00000000  0.00000000
## Indicator1    0.0000000  0.50162766 -0.26028512  0.01894258
## Indicator2    0.0000000 -0.26028512  0.48815738 -0.01473312
## X.centered    0.0000000  0.01894258 -0.01473312  0.01052366
```

```
# Since we only need the first 3*3 submatrix for mu, tau1, tau2
 (Sigma= vcov(LM.full)[1:3, 1:3])
```

```
##             (Intercept) Indicator1 Indicator2
## (Intercept)   0.2337655  0.0000000  0.0000000
## Indicator1    0.0000000  0.5016277 -0.2602851
## Indicator2    0.0000000 -0.2602851  0.4881574
```

For example, as special cases:

- For pairwise differences: as in (22.16a), when $L = \tau_1 - \tau_2$, its estimate is $\hat{L} = \hat{\tau}_1 - \hat{\tau}_2 = 6.017 - 0.9420 = 5.075$, its variance estimate is $\widehat{var}(\hat{L}) = \widehat{var}(\hat{\tau}_1) + \widehat{var}(\hat{\tau}_2) - 2\widehat{cov}(\hat{\tau}_1, \hat{\tau}_2) = .5016 + .48822(.2603) = 1.5104$.
- For adjusted treatment means (least squares means): as in (22.22), when $L = \mu_. + \tau_1$, its estimate $\hat{L} = \hat{\mu}_. + \hat{\tau}_1 = 33.8 + 6.017 = 39.817$, and its variance estimate is $\widehat{var}(\hat{L}) = \widehat{var}(\hat{\mu}_.) + \widehat{var}(\hat{\tau}_1) + 2\widehat{cov}(\hat{\mu}_., \hat{\tau}_1) = .2338 + .5016 + 2(0) = .7354$.

1. Therefore for any linear combination of treatment effects L, with the point estimate $\hat{L}$ and the variance estimate $s^2(\hat{L})$, we can get the $(1-\alpha)$ confidence interval as $\hat{L} \pm t(1-\alpha/2, n_T - r - 1)s(\hat{L})$, where the corresponding t-distribution has $df = n_T - r - 1$ (1 df is used to estimate the covariate coefficient compared to the standard ANOVA model).
2. When we make inferences on multiple parameters or several linear combinations, $L_1, L_2, ...$, we need to adjust for mutiple comparisons using those procedure that we learned previous. For examples,

- Bonferroni method, with its multiple to replace $t(1-\alpha/2, n_T - r - 1)$ in the CI formula,

$$B = t(1 - \alpha/(2g); n_T - r - 1)$$

- Scheffe procedure (for a number of contrasts), with its multiple

$$S = \sqrt{(r-1)F(1-\alpha; r-1, n_T - r - 1)}$$

Example: Estimate of treatment means using R

We have just discussed how to use the regression and matrix operation to calculate it point estimates and confidence interval with adjustment for mutiple comparison, if needed. The computations above can be performed much more conveniently using the R package **emmeans**.

1. Estimate the adjusted treatment mean (or "Least squares (LS) means")

The mean response at $X = \overline{X}_{..}$ for ith treatment is $\mu_. + \tau_i$. We can get the mean response and the estimated CI as follows:

```
library(emmeans)
fit<- lm(New_Sales~ Treatment + X.centered, data=Ex22 )
fit.emm <- emmeans( fit, ~ Treatment)
# CI without adjustment for MCP
fit.emm
```

```
##  Treatment emmean    SE df lower.CL upper.CL
##  1          39.8 0.858 11     37.9     41.7
##  2          34.7 0.850 11     32.9     36.6
##  3          26.8 0.838 11     25.0     28.7
##
## Confidence level used: 0.95
```

```
# CI with adjustment for MCP
confint(fit.emm, adjust = "Bonferroni")
```

```
##  Treatment emmean    SE df lower.CL upper.CL
##  1           39.8 0.858 11     37.4     42.2
##  2           34.7 0.850 11     32.3     37.1
##  3           26.8 0.838 11     24.5     29.2
##
## Confidence level used: 0.95
## Conf-level adjustment: bonferroni method for 3 estimates
```

2. Pairwise comparison of treatment effects $\tau_i - \tau_{i'}$.

```
confint(pairs(fit.emm), adjust = "scheffe")
```

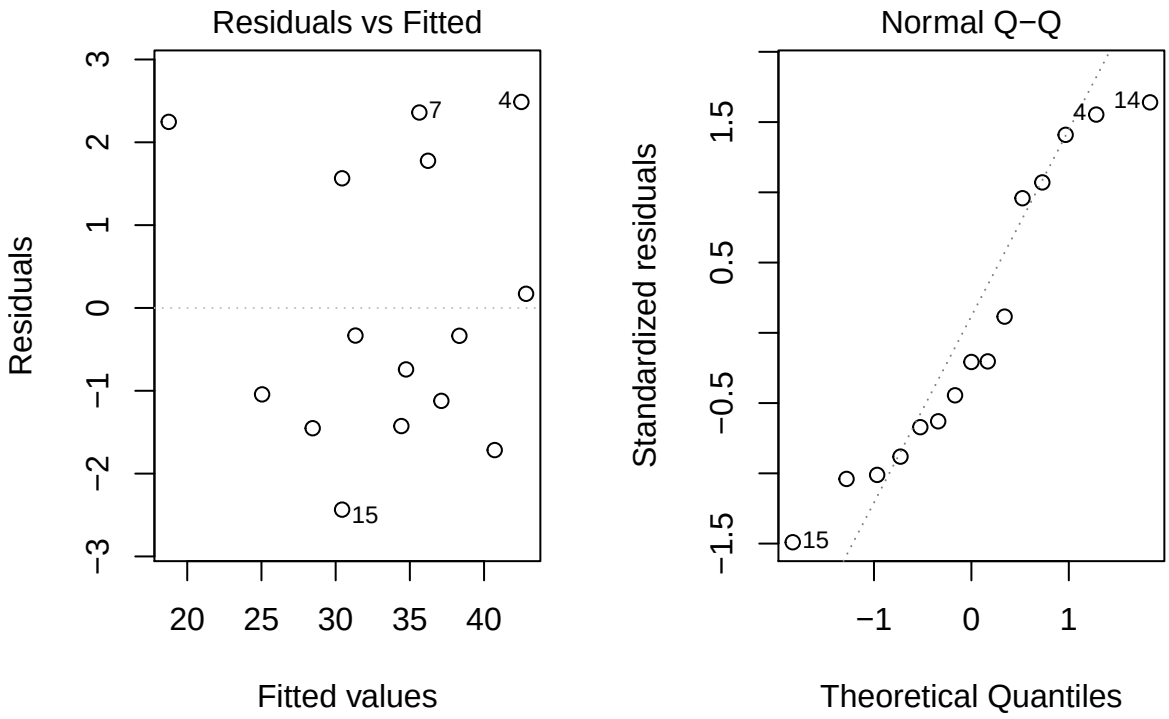
```
##  contrast estimate   SE df lower.CL upper.CL
##  1 - 2        5.08 1.23 11     1.61     8.54
##  1 - 3       12.98 1.21 11     9.57    16.38
##  2 - 3        7.90 1.19 11     4.55    11.26
##
## Confidence level used: 0.95
## Conf-level adjustment: scheffe method with dimensionality 2
```

Note: The point estimate, SE and 95% CI with Scheffe adjustment are exactly same the results in the textbook, page 931.
Since none of the confidence intervals cover zero. We can also concluded $\tau_1 > \tau_2 > \tau_3$ with 95% confidence. That is, treatment 1 is significantly better than the other two treatments, and treatment 2 is superior to treatment 3.

Example: Model diagnostic check

1A. checking standard ANOVA assumptions from residuals

```
par(mfrow=c(1,2))
plot(LM.full, 1, add.smooth = F)
plot(LM.full, 2)
```



Results: Residual plots do not suggest any major differences in the variances of the error terms or any significant departure from normal errors.

1B. check the assumptions using statistical tests

1. We have learned to use the Brown-Forsythe test the for homogeneity of variances. The R package "**car**" has a function **leveneTest** can be used to directly apply this test.

```
Rpackage= "car"
if (! Rpackage %in% installed.packages()) install.packages(Rpackage)
library(car)
```

```
## Loading required package: carData
```

```
leveneTest(New_Sales~ Treatment, center='median', data=Ex22)
```

```
## Levene's Test for Homogeneity of Variance (center = "median")
##       Df F value Pr(>F)
## group  2  0.1076 0.8988
##       12
```

2. We use shapiro-Wilk test to test normality of residuals

```r
shapiro.test(LM.full$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  LM.full$residuals
## W = 0.90084, p-value = 0.09803
```

Therefore, the results from the statistical tests suggested that data meet assumptions necessary for ANOVA.

Checking linearity and the constant slope (page 932)

Visual inspection of the scatter plots in our previous data checking step.

2A. Using model to check the linearity of X and Y in the entire data and by treatment.

```r
anova(lm(New_Sales~ Old_Sales, data= Ex22))
```

```
## Analysis of Variance Table
##
## Response: New_Sales
##           Df Sum Sq Mean Sq F value  Pr(>F)
## Old_Sales  1 190.68 190.678  5.4393 0.03641 *
## Residuals 13 455.72  35.056
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(lm(New_Sales~ Old_Sales, data= subset(Ex22, Treatment=='1')))
```

```
## Analysis of Variance Table
##
## Response: New_Sales
##           Df Sum Sq Mean Sq F value  Pr(>F)
## Old_Sales  1 65.256  65.256  14.454 0.03196 *
## Residuals  3 13.544   4.515
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(lm(New_Sales~ Old_Sales, data= subset(Ex22, Treatment=='2')))
```

```
## Analysis of Variance Table
##
## Response: New_Sales
##           Df Sum Sq Mean Sq F value   Pr(>F)
## Old_Sales  1 134.81 134.810  56.253 0.004911 **
## Residuals  3   7.19   2.397
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(lm(New_Sales~ Old_Sales, data= subset(Ex22, Treatment=='3')))
```

```
## Analysis of Variance Table
##
## Response: New_Sales
##           Df Sum Sq Mean Sq F value  Pr(>F)
## Old_Sales  1 76.012  76.012  21.139 0.01934 *
## Residuals  3 10.788   3.596
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2B. To perform a statistical test for the constancy of slope

We can compare the ANCOVA model with a more complex model that allows for different slopes for the treatments by introducing cross-product interaction terms:

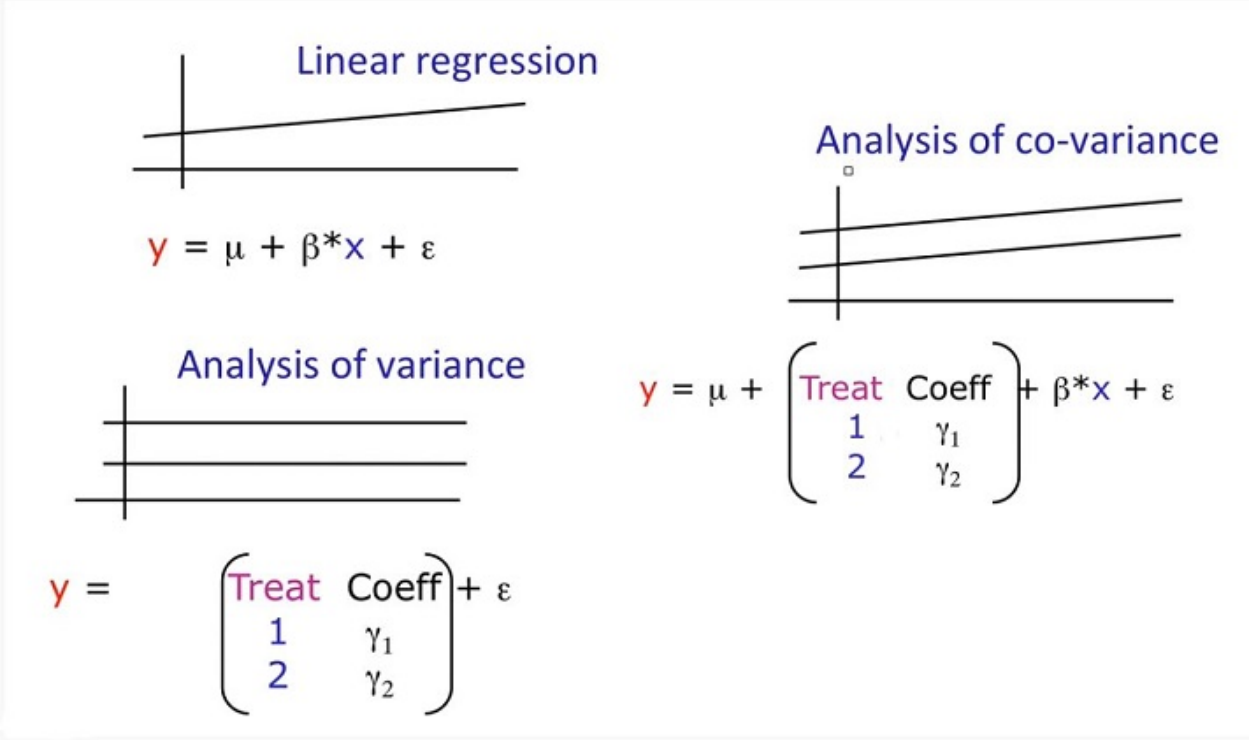
$$Y_{ij} = \mu_{.} + \tau_1 I_{ij1} + \tau_2 I_{ij2} + \gamma x_{ij} + \beta_1 I_{ij1} x_{ij} + \beta_2 I_{ij2} x_{ij} + \epsilon_{ij}$$

```r
# Adding interaction term
LM.Interaction  =  lm( New_Sales~ Indicator1 + Indicator2 + X.centered+
                         Indicator1:X.centered + Indicator2:X.centered,
                         data=Ex22 )

# compare two models
anova(LM.full, LM.Interaction)
```

```
## Analysis of Variance Table
##
## Model 1: New_Sales ~ Indicator1 + Indicator2 + X.centered
## Model 2: New_Sales ~ Indicator1 + Indicator2 + X.centered + Indicator1:X.centered +
##     Indicator2:X.centered
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     11 38.571
## 2      9 31.521  2    7.0505 1.0065 0.4032
```

Results: The F-test statistic= 1.0065 , following a F-distribution $F(df_1 = 2, df_2 = 9)$. The p-value = 0.4032, therefore we don't have any evidence to reject the null that the three treatment regression lines have the same slope.

Check back on Data fitting



The ANCOVA model seems appropriate here. We can add the fitted regression lines to the data to show the final model fit.

```
# model parameter
Coef
```

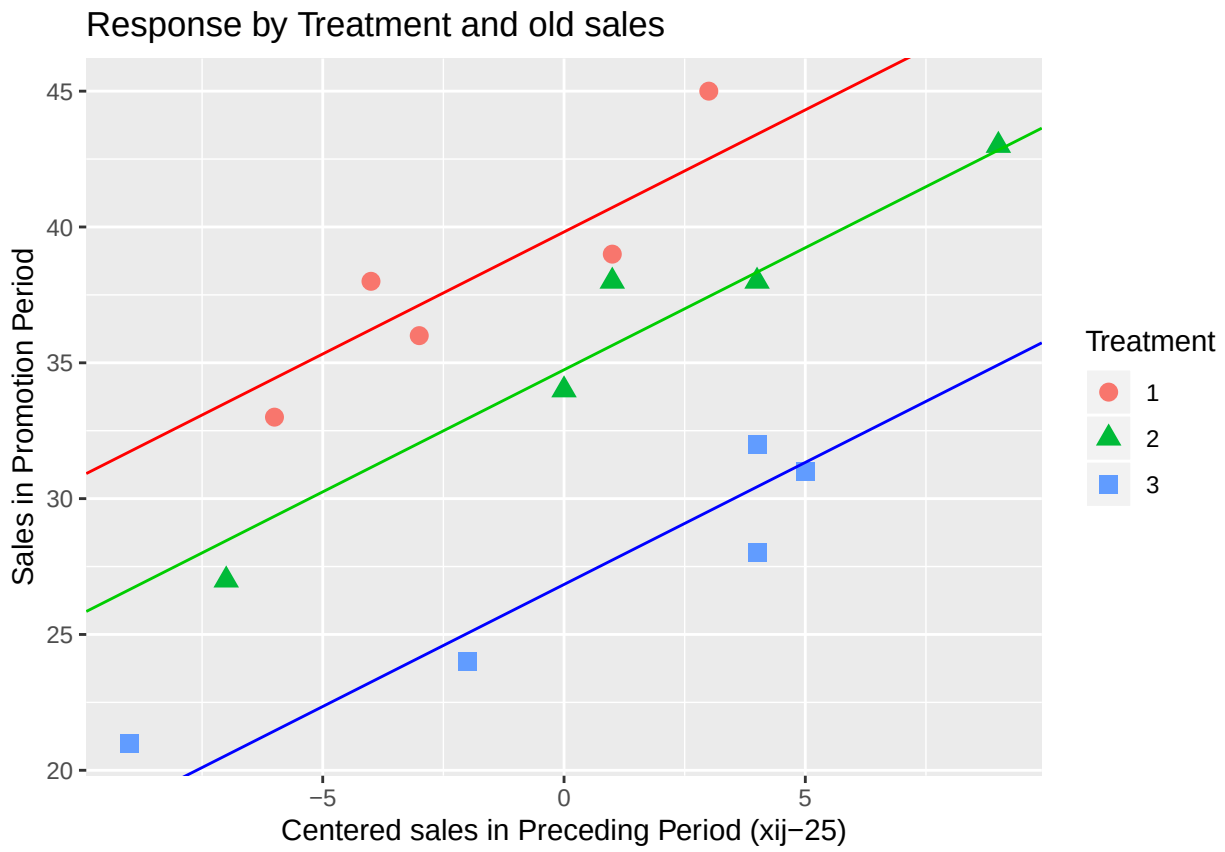
```
## (Intercept)  Indicator1  Indicator2  X.centered
##  33.8000000   6.0174070   0.9420168   0.8985594
```

```
Coef<- as.numeric(Coef)

# get slope= gamma, intercept= mu.+tau_i
Fitted= data.frame(Slope=Coef[4],
                   Intercept1=Coef[1]+Coef[2],
                   Intercept2=Coef[1]+Coef[3],
                   Intercept3=Coef[1]-Coef[2]-Coef[3])
Fitted
```

```
##       Slope Intercept1 Intercept2 Intercept3
## 1 0.8985594   39.81741   34.74202   26.84058
```

```
# Plot data and add regression lines
ggplot(data=Ex22, aes(x= Old_Sales-25, y= New_Sales, color= Treatment, shape=  Treatment ))+
  geom_point(size=3) +
  geom_abline(aes(slope=Slope , intercept=Intercept1), col=2,  data=Fitted) +
  geom_abline(aes(slope=Slope , intercept=Intercept2), col=3,  data=Fitted) +
  geom_abline(aes(slope=Slope , intercept=Intercept3), col=4,  data=Fitted) +
```

```
labs(title="Response by Treatment and old sales",
     x="Centered sales in Preceding Period (xij-25)", y="Sales in Promotion Period")
```



Generalization of ANCOVA model (page 923)

1. **Nonlinearity of Relation**. The linear relation between Y and X assumed in covariance model (22.3) is not essential. Any other relation could be used. For instance, the model for a quadratic relation is as follows:

$$Y_{ij} = \mu_. + \tau_i + \gamma_1(X_{ij} - \overline{X}_{..}) + \gamma_2(X_{ij} - \overline{X}_{..})^2 + \epsilon_{ij}$$

Linearity of the relation leads to simpler analysis and is often a sufficiently good approximation. If a linear relation is not a good approximation, however, a more adequate description of the relation should be utilized in the covariance model.

2. **Several Concomitant Variables**.

Covariance model (22.3) uses a single covariate, which is often sufficient to reduce the error variability substantially. However, the model can be extended to include ≥ 2 covariates. For two concomitant variables, X_1 and X_2, the ANCOVA model is as follows:

$$Y_{ij} = \mu_. + \tau_i + \gamma_1(X_{ij1} - \overline{X}_{..1}) + \gamma_2(X_{ij2} - \overline{X}_{..2}) + \epsilon_{ij}$$

Summary this week

- Reading: Chapter 22.1-22.3 (p.917-933) ANCOVA
- Reminder:

1) Homework (week 8 on block design) due on Thursday
2) Go over the mid-term exam on Thursday.