R403: Probabilistic and Statistical Computations with R

Lecture 16: Analysis of Covariance (ANCOVA)

Kaloyan Ganev

2022/2023

Lecture Contents

Introduction

Introduction

- Invented also by R. Fisher
- Similar to ANOVA: also used to test equality of population means
- Yet different: integrates ANOVA and regression analysis
- Applicability: both to observational and to experimental studies
- Involves at least three variables: one dependent variable, one independent variable, and one covariate
- The independent variable is a categorical one (also called a *treatment*, as in ANOVA)
- The covariate[1] is a variable which is likely to be correlated with the dependent variable
- In brief: ANCOVA is a special type of regression analysis including both quantitative and categorical regressors

[1] Also called *concomitant variable*, *nuisance variable*, etc. There may be more that one covariate.

Introduction (2)

- Covariates are used to control for effects that are not of primary interest
- This simultaneously leads to the reduction of the variance of the error term in ANOVA
- Also, they help obtain conditionally unbiased effects of treatments
- In experimental studies, their inclusion helps reduce the bias that may result from random differences between groups *before* the application of treatments

Single-Factor ANCOVA

Single-Factor ANCOVA

- Let there be $i = 1, \ldots, a$ samples (factor levels, groups)
- Let each sample contain n_i elements (cases)
- The total number of cases is $n = \sum_{i=1}^{a} n_i$
- Let Y be the response (dependent) variable; Y_{ij} then denotes its jth case in the ith group
- We will consider the version with only one covariate, call it X (X_{ij} is the value of the covariate that corresponds to the jth case in the ith group)

Single-Factor ANCOVA (2)

- Start with the single-factor fixed-effects ANOVA model:

$$Y_{ij} = \mu_. + \tau_i + \varepsilon_{ij}, \quad i = 1, 2, \ldots, a, \; j = 1, 2, \ldots, n_i$$

  where $\varepsilon_{ij} \sim N(0, \sigma^2)$
- $\mu_.$ is the overall (grand) mean
- τ_i are the fixed treatment effects, and $\sum \tau_i = 0$

Single-Factor ANCOVA (3)

- To this we add a covariate that has a relationship with the dependent variable:

$$Y_{ij} = \mu_{\cdot} + \tau_i + \gamma X_{ij} + \varepsilon_{ij}$$

where γ is a regression coefficient, and X_{ij} are pre-determined (fixed, non-stochastic)

Single-Factor ANCOVA (4)

Warnings on the Choice of Covariates

- Should be made very carefully so as to 'add value' to the model
- Covariates which bear no relation to the dependent variable will bring no additional insight or quality of estimates; in such cases it is better to stick to the simpler ANOVA
- Covariates should either be observed before the application of the treatment(s) or there should be guarantees that they won't be affected by the study
- If covariates are affected by the treatment(s), then ANCOVA might produce results that do not adequately reflect the presence/absence of effects

Single-Factor ANCOVA (5)

- After including X_{ij} in the relationship, $\mu_{\cdot}$ is no longer the overall mean of Y_{ij}
- If, however, we center X_{ij} around their grand mean, then $\mu_{\cdot}$ again becomes Y's grand mean:

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

Single-Factor ANCOVA (6)

- Taking the mathematical expectation of the latter yields:

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

- The variance is respectively:

$$\mathrm{Var}(Y_{ij}) = \sigma^2$$

- Overall, since $\varepsilon_{ij} \sim n.i.d.(0, \sigma^2)$, it follows that:

$$Y_{ij} \sim n.i.d.(\mu_{ij}, \sigma^2)$$

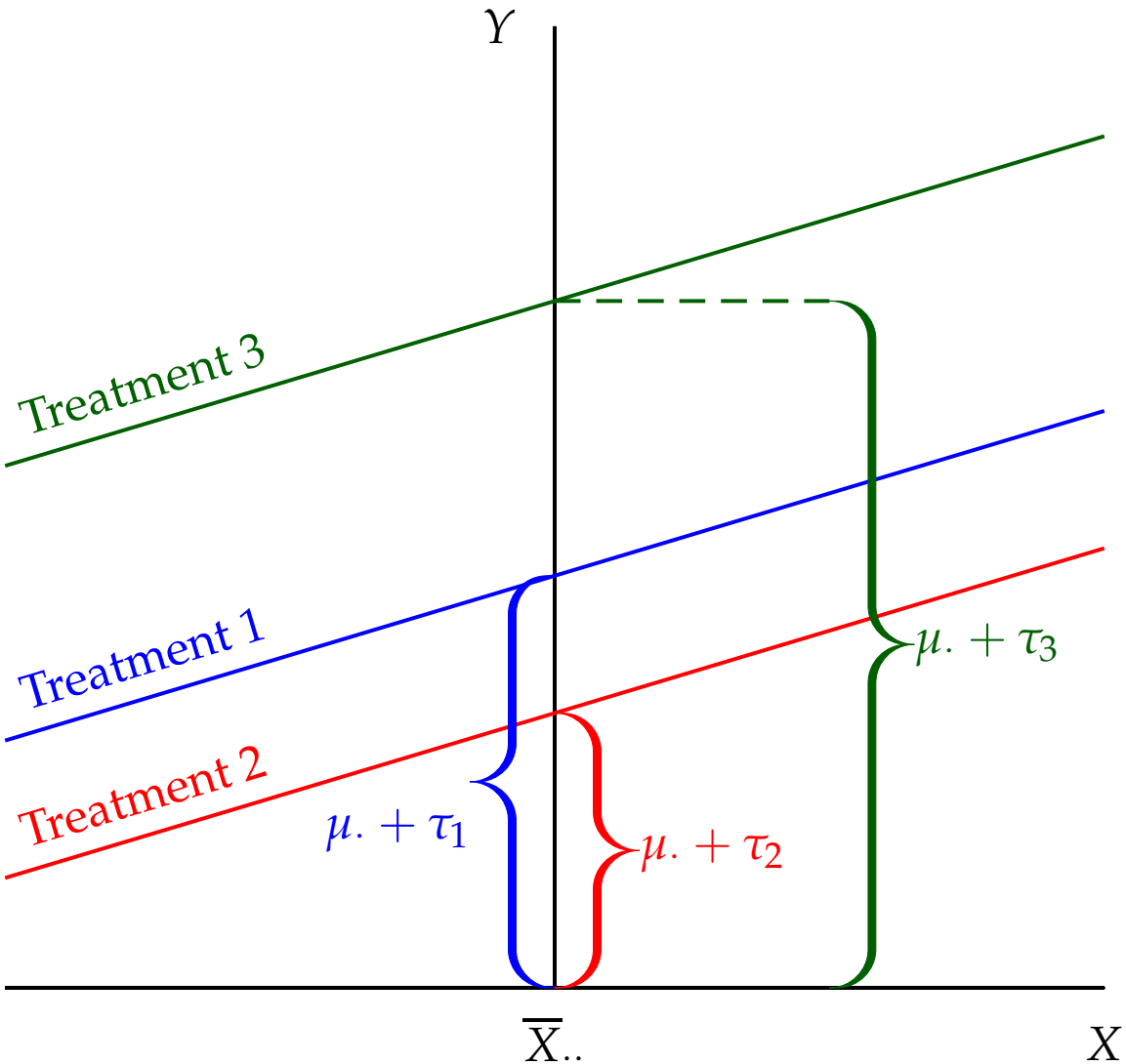
where $\mu_{ij} = \mu_{\cdot} + \tau_i + \gamma(X_{ij} - \overline{X}_{\cdot\cdot})$.

Single-Factor ANCOVA (7)

- Note that this result is different from the one in the the ANOVA model (recall it was $\mathrm{E}(Y_{ij}) = \mu_i$)
- The major difference stems from the dependence on the values of the covariate (in addition to the dependence on the treatment)
- The dependent variable in the ANCOVA model therefore is the expected response to the ith treatment
- As τ_i have (potentially) different values, that responses are given by regression lines (one line per treatment):

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

Single-Factor ANCOVA (8)



Single-Factor ANCOVA (9)

- Note that the 'origin' is $X - \overline{X}_{..} = 0$, i.e. $X = \overline{X}_{..}$
- Also, pay attention that all three regression lines are parallel to each other (slope equals γ)
- From the latter follows that there is no more a fixed mean response to treatment i
- The effects are measured by the vertical distances between two regression lines
- For example, the comparison of effects of Treatment 1 and Treatment 2 is estimated as:

$$\mu_{.} + \tau_1 - (\mu_{.} + \tau_2) = \tau_1 - \tau_2$$

- If all treatments have the same mean response for any X, then all such effects equal zero

Single-Factor ANCOVA: Possible Extensions

- **Stochastic covariates:** the ANCOVA still remains valid, if for any possible values of X the model can be interpreted as a conditional one
- **Non-linearity of the relationship between Y and X:** linearity is not essential so for example a cubic relationship can be used:

$$Y_{ij} = \mu_{\cdot} + \tau_i + \gamma_1(X_{ij} - \overline{X}_{\cdot\cdot}) + \gamma_2(X_{ij} - \overline{X}_{\cdot\cdot})^2 + \gamma_3(X_{ij} - \overline{X}_{\cdot\cdot})^3 + \varepsilon_{ij}$$

  What is essential is the constancy of the γ's, i.e. the response functions should be parallel to each other[2]
- **Multiple covariates:** inclusion is straightforward, e.g.:

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

[2] This also means there is no interaction among treatments.

ANCOVA: Regression Formulation

- A convenient way to estimate the model parameters
- Specifically, this convenience is implemented in statistical packages, including R
- Take the single-factor, single-covariate ANCOVA model
- Let the number of treatments (level) be r
- Define $r-1$ indicator variables in the following way:

$$I_k = \begin{cases} 1, \text{ in case of treatment } k \\ -1, \text{ in case of treatment } r \\ 0, \text{ otherwise} \end{cases},$$

where $k = 1, \ldots, r-1$

ANCOVA: Regression Formulation (2)

- Denote:
$$x_{ij} = X_{ij} - \overline{X}_{..}$$
- Then the ANCOVA model can be written as:
$$Y_{ij} = \mu_{.} + \tau_1 I_{ij,1} + \ldots + \tau_{r-1} I_{ij,r-1} + \gamma x_{ij} + \varepsilon_{ij}$$
- The τ's turn out to be the regression coefficients in this formulation
- Note: In R we will use the programmed routines to estimate the coefficients. They are somewhat different and slightly roundabout. This also means that they do not follow the way of solving the problem found in the book

ANCOVA: Regression Formulation (3)

The following assumptions for appropriateness should hold before running the regression:

1. $\varepsilon_{ij} \sim n.i.d.(0, \sigma^2)$
2. Linearity of regression with respect to covariate
3. Constancy of γ (i.e. slopes of regression lines across treatments should be equal)

ANCOVA: Regression Formulation (4)

- The main hypothesis to be tested coincides with that of ANOVA, i.e.:

$$H_0 : \tau_1 = \tau_2 = \ldots = \tau_r = 0$$

  against:

$$H_1 : \text{at least one } \tau \neq 0$$

- In the regression setting, the test boils down to testing whether a group of regression coefficients are simultaneously equal to zero
- This is done by means of a partial F-test
- If the result is statistically significant, the cause should be investigated
- For example, pairwise comparisons of the τ's should be made

Single-factor ANCOVA: An Example in R

An Example in R

- Adapted to R from Kutner et al. (2004), p. 926
- A company sells crackers, and wants to know the effects of three different types of promotions:
  1. Treatment 1: Sampling by customers in store
  2. Treatment 2: Additional shelf space
  3. Treatment 3: Special display shelves in addition to regular shelf space
- Fifteen stores' counts of sales[3] were studied; five stores were assigned to each treatment
- The following table contains data on the number of sales Y during the promotional period, and on the number of sales X in the preceding period (note: X is the covariate!)

[3] i.e. number of sales

Single-factor ANCOVA: An Example in R (2)

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

- The data are contained also in `stores.csv`
- Load this file in R via:

```
stores <- read.csv2(stores.csv)
```

Single-factor ANCOVA: An Example in R (3)

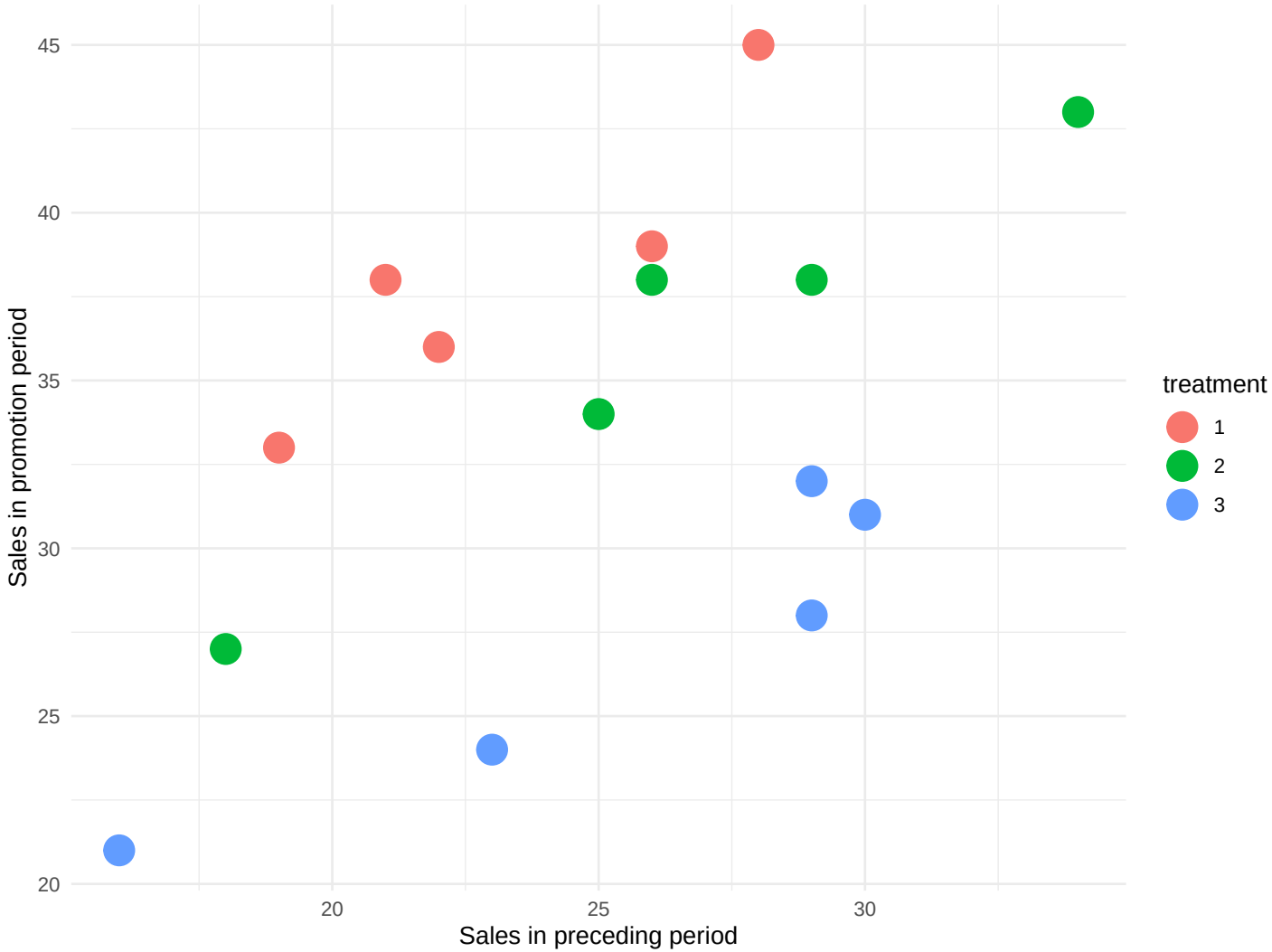
- ...and set factors:

```
stores$treatment <- as.factor(stores$treatment)
stores$store <- as.factor(stores$store)
attach(stores)
```

- Then plot the result:

```
library(ggplot2)
ggplot(data=stores,aes(x=X,y=Y,col=treatment)) +
  geom_point(size=I(6)) +
  xlab("Sales in preceding period") +
  ylab("Sales in promotion period") +
  theme_minimal()
```

Single-factor ANCOVA: An Example in R (4)



Single-factor ANCOVA: An Example in R (5)

- The grand mean equals 25:

```
mean(X)
```

- Generate the centred covariate:

```
xcov <- X - mean(X)
```

- Run the ANCOVA regression:

```
ancova1 <- lm(Y ~ treatment + xcov)
```

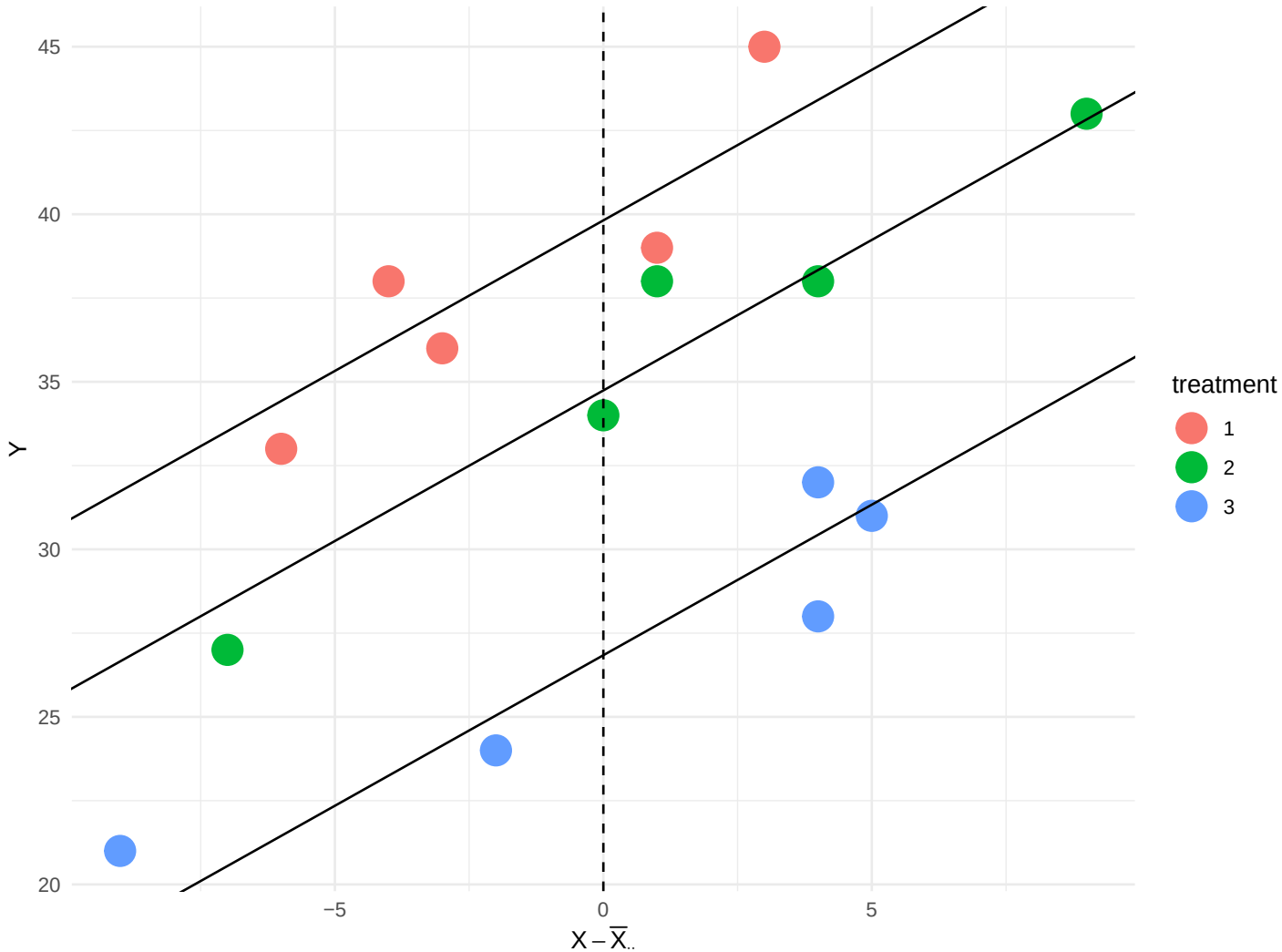
Single-factor ANCOVA: An Example in R (6)

- Graph data and fitted lines:

```
library(latex2exp)

ggplot(data=stores,aes(x=xcov,y=Y,col=treatment)) +
  geom_point(size=6) + geom_abline(aes(slope = ancova1$
      coefficients[4],intercept = ancova1$coefficients[1])) +
  geom_abline(aes(slope = ancova1$coefficients[4],intercept =
      ancova1$coefficients[1]+ancova1$coefficients[2])) +
  geom_abline(aes(slope = ancova1$coefficients[4],intercept =
      ancova1$coefficients[1]+ancova1$coefficients[3])) +
  xlab(TeX('$X - \\bar{X}_{\\cdot\\cdot}$')) + ylab(TeX('Y')) +
  geom_vline(xintercept = 0, lty = 2) +
  theme_minimal()
```

Single-factor ANCOVA: An Example in R (7)



Single-factor ANCOVA: An Example in R (8)

- How to interpret those results? First, look again at the estimation output
- Note that $\gamma = 0.8986$ (this is the slope)
- We want to find $\mu.$ and all τ's
- (Intercept) ($= 39.8174$) in the regression output gives the intersection point of the regression line and the vertical line at $X - \overline{X}..$; it also equals $\mu. + \tau_1$
- The next two coefficients provide the respective differences of intercepts of treatment 2 and treatment 3 from the intercept of treatment 1
- Therefore, $\mu. + \tau_2 = 39.8174 - 5.0754 = 34.742$, and $\mu. + \tau_3 = 39.8174 - 12.9768 = 26.8406$

Single-factor ANCOVA: An Example in R (9)

- Summing the three intercepts yields:

$$\mu. + \tau_1 + \mu. + \tau_2 + \mu. + \tau_3 = 101.4$$

- But we also know that $\sum_i \tau_i = 0$, so:

$$\mu. + \tau_1 + \mu. + \tau_2 + \mu. + \tau_3 = 3\mu.$$

- Combining the two results, we find:

$$3\mu. = 101.4 \Rightarrow \mu. = 33.8$$

- Finally, $\tau_1 = 6.0174$, $\tau_2 = 0.942$, and $\tau_3 = -6.9594$

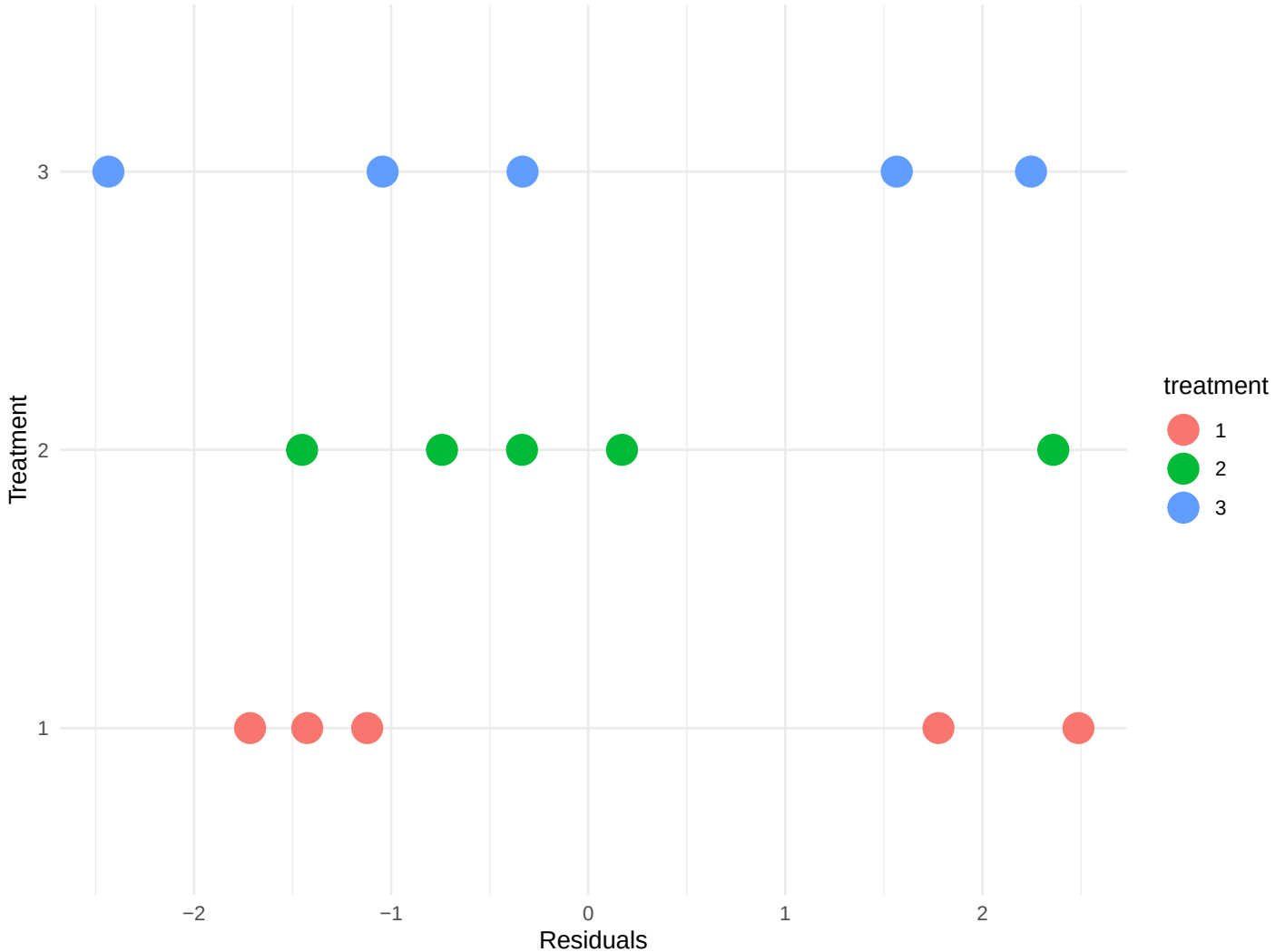
Single-factor ANCOVA: An Example in R (10)

- The adequacy of the fitted model can be assessed first visually using some residual plots
- The first plot contains the values of residuals aligned to each corresponding treatment
- Code to reproduce the plot:

```
ancova1.df <- data.frame(Fitted = fitted(ancova1), Residuals =
    resid(ancova1), Treatment = treatment)
ggplot(ancova1.df, aes(x = Residuals, y = Treatment,color=
    treatment)) +
  geom_point(size = 6) +
  theme_minimal()
```

- It can be easily seen that the ranges of the three groups of residuals are similar which does not suggest unequal variances

Single-factor ANCOVA: An Example in R (11)



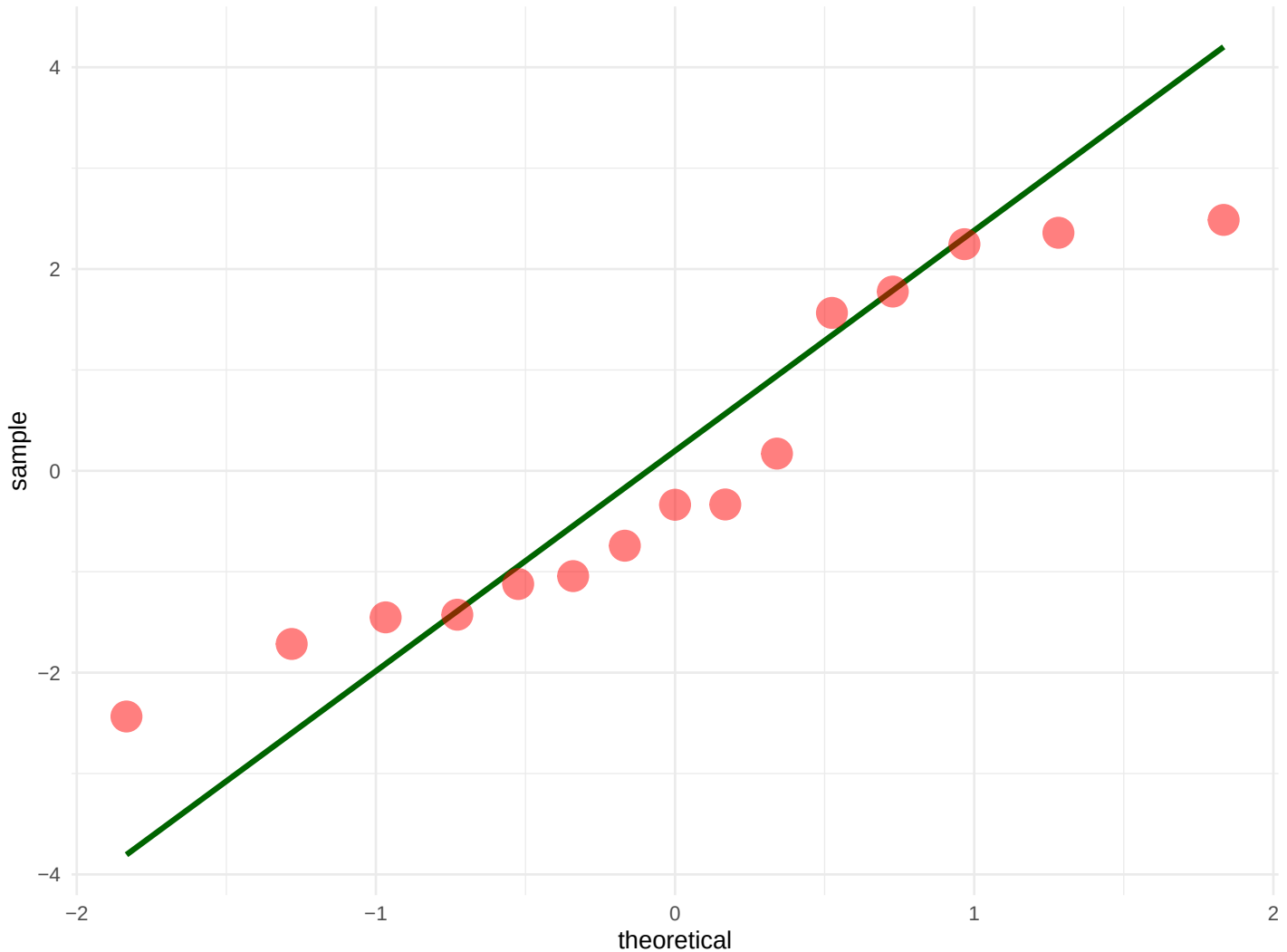
Single-factor ANCOVA: An Example in R (12)

- The second plot is a normal q-q plot

```
ggplot(data = ancova1.df, aes(sample = Residuals)) +
  stat_qq_line(size = 1.2, color = "darkgreen") +
  stat_qq(size=6,color="red", alpha = 0.5) +
  theme_minimal()
```

- It compares the actual values with the theoretical values of the normal distribution
- If the actual values seem to lie on the 45°-line, then the empirical distribution is approximately normal
- In the current case there is some visual evidence on non-normality

Single-factor ANCOVA: An Example in R (13)



Single-factor ANCOVA: An Example in R (14)

- However, applying a formal test such as the Jarque-Bera one, does not lead to the rejection of the normality hypothesis

```
library(moments)
jarque.test(ancova1.df$Residuals)
```

- Then, overall, we can say that the model is appropriate
- Now, we proceed to testing the significance of treatment effects
- The null hypothesis is:

$$H_0 : \tau_1 = \tau_2 = 0$$

- This automatically implies that $\tau_3 = 0$ (Why?)
- The alternative is that at least one of the τ's is not equal to zero

Single-factor ANCOVA: An Example in R (15)

- To make the inference, first run the restricted model:

$$Y_{ij} = \mu_{\cdot} + \gamma x_{ij} + \varepsilon_{ij}$$

- Run the restricted model in R, then perform ANOVA on it:

```
ancova1r <- lm(Y ~ xcov)
anova(ancova1r)
```

- The sum of squares of the residuals is 455.72
- Performing ANOVA on the unrestricted model:

```
anova(ancova1)
```

yields a value for the sum of squares of the residuals equal to 38.57

Single-factor ANCOVA: An Example in R (16)

- We will use the following statistic to test significance:

$$F = \frac{SSE(R) - SSE(U)}{SSE(U)} \Big/ \frac{df_R - df_U}{df_U}$$

  where SSE denotes sum of squares of errors, df denotes degrees of freedom, and R and U stand respectively for 'restricted' and 'unrestricted'
- This statistic follows an F distribution when H_0 holds
- In the current example:

$$F = \frac{455.75 - 38.57}{38.57} \Big/ \frac{13 - 11}{11} = 59.5$$

- At the 5% significance level this exceeds the critical value, so we have a significant result
- This leads to the rejection of the null

Single-factor ANCOVA: An Example in R (17)

- The **mean treatment effects** are easy to estimate knowing the values of the τ's; they are as follows:

$$\begin{aligned} \tau_1 - \tau_2 &= 6.0174 - 0.9420 &&= 5.0754 \\ \tau_1 - \tau_3 &= 6.0174 + 6.9594 &&= 12.9768 \\ \tau_2 - \tau_3 &= 0.9420 + 6.9594 &&= 7.9014 \end{aligned}$$

- Note that the first two figures are just the negatives of the second and the third regression coefficients
- Print the variance-covariance matrix of the ANCOVA regression:

```
vcov(ancova1)
```

Single-factor ANCOVA: An Example in R (18)

- From it (and using the regression output) we can directly see that:

$$\begin{aligned} \text{Var}(\tau_1 - \tau_2) &= 1.5104 \\ \text{Var}(\tau_1 - \tau_3) &= 1.4535 \end{aligned}$$

- We have to only find $\text{Var}(\tau_2 - \tau_3)$
- Noting that:

$$(\tau_1 - \tau_3) - (\tau_1 - \tau_2) = \tau_2 - \tau_3,$$

  we easily see that:

$$\begin{aligned} \text{Var}(\tau_2 - \tau_3) &= \text{Var}[(\tau_1 - \tau_3) - (\tau_1 - \tau_2)] = \\ &= \text{Var}(\tau_1 - \tau_3) + \text{Var}(\tau_1 - \tau_2) - 2\text{cov}(\tau_1 - \tau_3, \tau_1 - \tau_2) \end{aligned}$$

- The covariance is also available from R, so:

$$\text{Var}(\tau_2 - \tau_3) = 1.4131$$

Single-factor ANCOVA: An Example in R (19)

- Using means and variances, we can construct a family of confidence intervals to make a multiple comparison of effects
- This is done using Scheffé's approach:

$$\text{E(effect)} - S\sqrt{\text{Var(effect)}} \leq \text{E(effect)} \leq \text{E(effect)} + S\sqrt{\text{Var(effect)}},$$

where $S^2 = (r-1)F(1-\alpha; r-1, n-r-1)$.

Single-factor ANCOVA: An Example in R (20)

- As a hint, note that in the current case, with $\alpha = 0.05$:

$$S^2 = (3-1)F(0.95; 3-1, 15-3-1) = 2F(0.95; 2, 11)$$

- You can use R again to find the F-value:

```
qf(0.95,2,11)
```

- The remaining calculations are left as an exercise (straightforward)
- What the results show is that Treatment 1 (tasting) is more successful that the other two

Single-factor ANCOVA: An Example in R (21)

- So far we simply assumed that slopes are parallel
- What if we have to assert it?
- We introduce interaction terms in the model to allow for different slopes per treatment:

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

- This can be done in R as follows:

```
ancova2 <- lm(Y ~ treatment*xcov)
```

or, alternatively:

```
ancova2 <- lm(Y ~ treatment + xcov + treatment:xcov)
```

- See estimation output

Single-factor ANCOVA: An Example in R (22)

- You can easily see from the output that the interaction terms are not statistically significant
- As an exercise, plot the data and add regression lines assuming that the interaction terms are statistically significant

Two-Factor ANCOVA

Two-Factor ANCOVA

- We will just briefly sketch it
- Start from the fixed-effects ANOVA model with two factors and balanced data:

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk},\ i = 1, \ldots, a,\ j = 1, \ldots, b,\ k = 1, \ldots, n$$

  where α_i is the main effect of factor A at level i, β_j is the main effect of factor B at level j, and $\alpha\beta$ is an interaction effect
- With the same notation, the two-factor ANCOVA model is:

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma(X_{ijk} - \overline{X}_{...}) + \varepsilon_{ijk}$$

Two-Factor ANCOVA (2)

- From this point onwards, analogical considerations are applied
- In R, adding a factor variable in a regression model is straightforward
- Just do not forget the interaction term!
- Matters get a little bit more complicated in unbalanced datasets but we leave that and other complications to your curiosity

References

References

- Kutner, Nachtsheim, and Neter (2005): *Applied Linear Statistical Models*, McGraw-Hill, 5th edn.
- Cochran (1957): Analysis of Covariance: Its Nature and Uses, *Biometrics*, Vol. 13, No. 3, Special Issue on the Analysis of Covariance (Sep., 1957), pp. 261-281
- Huitema (2011): *The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies*, Wiley, 2nd edn.