ST 518 Homework 8

Eric Warren

November 4, 2023

Contents

1 Problem 1 **1**

1.1 Part A . . . 2
1.2 Part B . . . 2
1.3 Part C . . . 3
1.4 Part D . . . 3
1.5 Part E . . . 3
1.6 Part F . . . 3
1.7 Part G . . . 4
1.8 Part H . . . 5
1.9 Part I (Not Graded) . . . 5

2 Problem 2 **7**

2.1 Part A . . . 7
2.2 Part B . . . 8

1 Problem 1

An experiment randomizes seeds from `t = 10` plants from an F2 generation of soybeans to `N = 30` homogeneous plots. The percentage protein content is measured in the seds from the plants produced in each plot with results below and in the file **"protein-content.dat"**.

First we are going to read in the data.

```
library(tidyverse)
(protein <- read_table("protein-content.dat"))
```

```
## # A tibble: 10 x 4
##    Plant Plot1 Plot2 Plot3
##    <dbl> <dbl> <dbl> <dbl>
##  1     1  42.4  41    39.6
##  2     2  28.6  36.3  42.2
```

```
##  3     3  43.2  42.1  40.2
##  4     4  40.8  41    38.9
##  5     5  41    38.3  41.1
##  6     6  39.4  39.5  37.2
##  7     7  39.6  40.4  38.9
##  8     8  38.1  38.3  37.9
##  9     9  35.9  36.1  35.6
## 10    10  39.6  39.9  39.7
```

Consider a random effects model for these protein contents: $Y_{ij} = \mu + T_i + E_{ij}$.

1.1 Part A

Give all distributional assumptions and limits on indices/subscripts i and j.

- $i = 1, 2, ..., t$ or in this case $i = 1, 2, ..., 10$
- $j = 1, 2, ..., n$ or in this case $j = 1, 2, 3$
- $T_1, T_2, ..., T_t \sim^{iid} N(0, \sigma_T^2)$ or in this case $T_1, T_2, ..., T_10 \sim^{iid} N(0, \sigma_T^2)$
- $E_{11}, ..., E_{tn} \sim^{iid} N(0, \sigma^2)$ or in this case $E_{11}, ..., T_{10,3} \sim^{iid} N(0, \sigma^2)$
- $T_1, T_2, ..., T_t$ are independent of $E_{11}, ..., E_{tn}$ or in this case $T_1, T_2, ..., T_10$ are independent of $E_{11}, ..., E_{10,3}$

1.2 Part B

Test for a genetic component to protein content. That is, test $H_0 : Var(T_i) = 0$.

We are going to test $H_0 : Var(T_i) = 0$ with $H_A : Var(T_i) \neq 0$. We know that should do a F-test where $F = \frac{MS(T)}{MS(E)}$. We are going to get the anova table of these values below but first we have to transform the data to do this.

```
# Transform data
protein_transform <- protein %>%
  pivot_longer(Plot1:Plot3, names_to = "plot", values_to = "content")

# Make the model
model1 <- lm(content ~ Plant, protein_transform)

# Get anova table
(anova_model1 <- anova(model1))
```

```
## Analysis of Variance Table
##
## Response: content
##           Df  Sum Sq Mean Sq F value Pr(>F)
## Plant      1   9.387  9.3868  1.2138   0.28
## Residuals 28 216.532  7.7333
```

As we can see our F-value $F = \frac{MS(T)}{MS(E)} = 1.2138$ with 2, 27 degrees of freedom. We can see our p-value is 0.2799572 which is higher than most reasonable significance levels we can decide. Thus, we fail to reject our H_0 and can say that it is plausible that $Var(T_i) = 0$.

1.3 Part C

We need to find A, A^{-1}, and $\hat{\theta}$.

Now we know that $A\theta = MS$. Therefore to find just A, we can say that $A = MS\,\theta'$. We know that $MS = \begin{pmatrix} MS(T) \\ MS(E) \end{pmatrix}$ which from **Part C** directions we can plug in $MS(T) = 3\sigma_T^2 + \sigma^2$ and $MS(E) = \sigma^2$. Thus, $MS = \begin{pmatrix} MS(T) \\ MS(E) \end{pmatrix} = \begin{pmatrix} 3\sigma_T^2 + \sigma^2 \\ \sigma^2 \end{pmatrix}$. Now we know that $\theta = \begin{pmatrix} \sigma_T^2 & \sigma^2 \end{pmatrix}$.

Now we can let us plug in and then solve for A. Our base equation is $A\theta = MS$. Plugging everything in we get $\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} \sigma_T^2 & \sigma^2 \end{pmatrix} = \begin{pmatrix} 3\sigma_T^2 + \sigma^2 \\ \sigma^2 \end{pmatrix}$. Using what we know about matrix multiplication we can get that this turns into equations $a_{11}\sigma_T^2 + a_{12}\sigma^2 = 3\sigma_T^2 + \sigma^2$ and $a_{21}\sigma_T^2 + a_{22}\sigma^2 = \sigma^2$. We can clearly solve for the following to be true which is $a_{11} = 3, a_{12} = 1, a_{21} = 0, a_{22} = 1$. Thus, $A = \begin{pmatrix} 3 & 1 \\ 0 & 1 \end{pmatrix}$

Now we can find A' as the inverse of A. Thus, $A^{-1} = \begin{pmatrix} 3 & 1 \\ 0 & 1 \end{pmatrix}^{-1} = \frac{1}{det\begin{pmatrix} 3 & 1 \\ 0 & 1 \end{pmatrix}} \begin{pmatrix} 1 & -1 \\ 0 & 3 \end{pmatrix} = \begin{pmatrix} \frac{1}{3} & \frac{-1}{3} \\ 0 & 1 \end{pmatrix}$.

Now we can find $\hat{\theta} = A^{-1}MS = \begin{pmatrix} \frac{1}{3} & \frac{-1}{3} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 3\hat{\sigma_T^2} + \hat{\sigma^2} \\ \hat{\sigma^2} \end{pmatrix} = \begin{pmatrix} \hat{\sigma_T^2} \\ \hat{\sigma^2} \end{pmatrix}$ which makes sense since we expect $\hat{\theta} = \begin{pmatrix} \hat{\sigma_T^2} \\ \hat{\sigma^2} \end{pmatrix}$ because $\theta = \begin{pmatrix} \sigma_T^2 \\ \sigma^2 \end{pmatrix}$. In this case $\hat{\sigma^2} = MS(E) = 7.7333$ and $\hat{\sigma_T^2} = \frac{MS(T) - MS(E)}{n} = \frac{9.3868 - 7.7333}{3} = 0.5511667$ so $\hat{\theta} = \begin{pmatrix} \hat{\sigma_T^2} \\ \hat{\sigma^2} \end{pmatrix} = \begin{pmatrix} 0.5511667 \\ 7.7333 \end{pmatrix}$

1.4 Part D

Estimate the coefficient of variation among protein contents.

We need to know $\hat{\sigma_T^2}$ which we can find is $\hat{\sigma_T^2} = \frac{MS(T) - MS(E)}{3} = \frac{9.3868 - 7.7333}{3} = 0.5511667$. We also need to know $\hat{\mu} = E(\hat{Y_{ij}}) =$ `mean(protein_transform$content)` $= 39.0933333$. The coefficient of variation among protein contents is $CV(Y_{ij}) = \frac{\sqrt{\sigma_T^2 + \sigma^2}}{|\hat{\mu}|} = \frac{\sqrt{\sigma_T^2 + \sigma^2}}{|\bar{y_{..}}|} = \frac{\sqrt{0.5511667 + 7.7333}}{|39.09333|} = 0.07362573$. Thus, the coefficient of variation among protein contents is 0.07362573 or 7.362573%.

1.5 Part E

Estimate $Corr(Y_{21}, Y_{22})$, the so-called *heritability* of protein content.

Note from **Part D** we found that $\hat{\sigma_T^2} = 0.5511667$.

We know we can find this as $Corr(Y_{21}, Y_{22}) = \frac{Cov(Y_{21}, Y_{22})}{\sqrt{Var(Y_{21})Var(Y_{22})}} = \frac{\hat{\sigma_T^2}}{\hat{\sigma_T^2} + \hat{\sigma^2}} = \frac{0.5511667}{0.5511667 + 7.7333} = 0.06653014$. Therefore, the so-called *heritability* of protein content is 0.06653014.

1.6 Part F

Report a 95% confidence interval for the mean protein content μ of seeds grown from a randomly sampled Plant.

We know this confidence interval formula is $\bar{y..} \pm t(0.025, t-1)\sqrt{\frac{MS(T)}{nt}}$. We know that $t = 10, n = 3, \bar{y..} = 39.09333, MS(T) = 9.3868$ so our confidence interval is $39.09333 \pm t(0.025, 10-1)\sqrt{\frac{9.3868}{3*10}}$ ===> $39.09333 \pm t(0.025, 9)\sqrt{\frac{9.3868}{30}}$ ===> $39.09333 \pm 2.262157 * 0.5593687$ ===> 39.09333 ± 1.26538 which gives us a lower interval of 37.82795 and upper confidence interval of 40.35871. Thus our 95% confidence interval for the mean protein content μ of seeds grown from a randomly sampled Plant is 37.82795 to 40.35871.

1.7 Part G

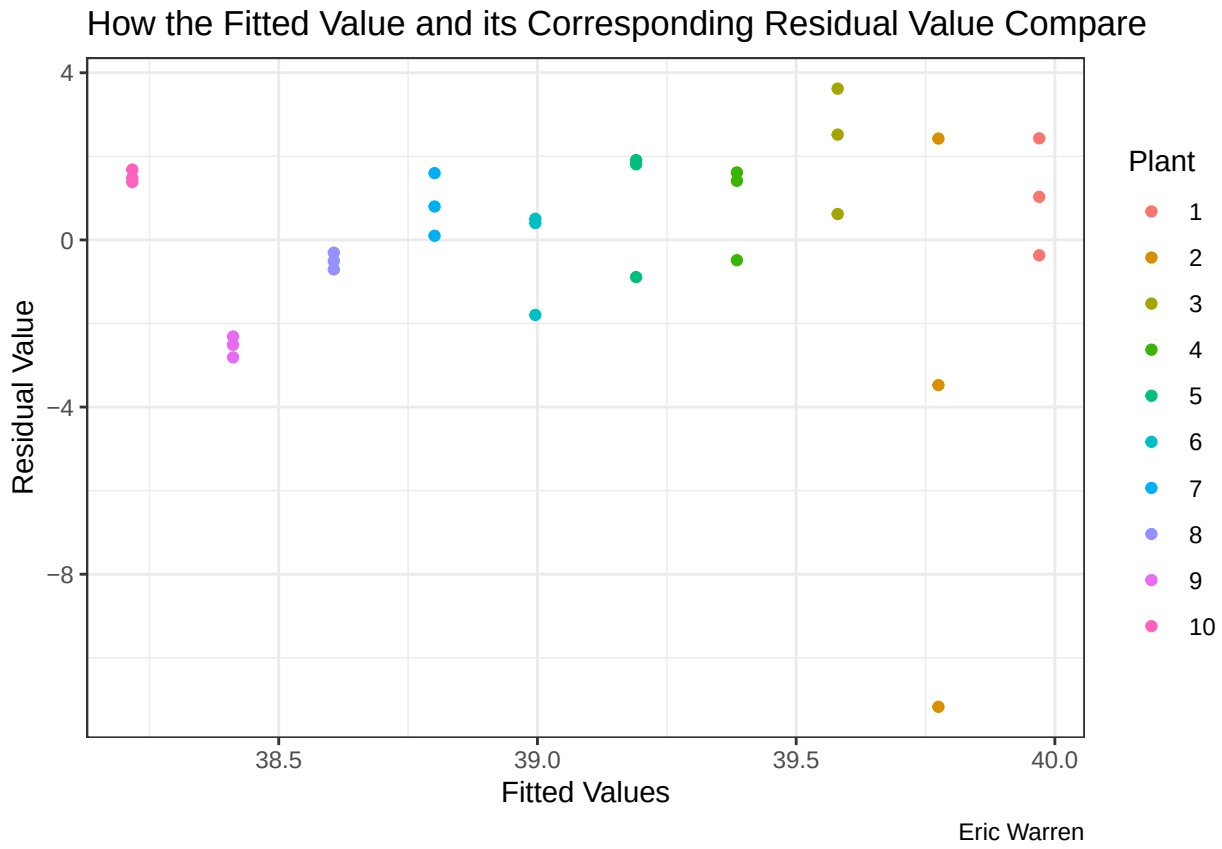
Plot the residuals from the fitted model, $y_{ij} - \bar{yi}.$ against the fitted values, , $\hat{y_{ij}} = \bar{yi}..$ Use a different plotting symbol or color for each of the 10 plants.

First we need to get the fitted values for our data. Then we will graph the residuals.

```
# Get the fitted value
protein_transform$predictions_model1 <- predict(model1, newdata = protein_transform)

# Get the residuals
protein_transform$residuals_model1 <- protein_transform$content - protein_transform$predictions_model1

# Graph the residuals
protein_transform %>%
  ggplot(aes(x = predictions_model1, y = residuals_model1)) +
  geom_point(aes(color = as.factor(Plant))) +
  labs(x = "Fitted Values",
       y = "Residual Value",
       color = "Plant",
       title = "How the Fitted Value and its Corresponding Residual Value Compare",
       caption = "Eric Warren") +
  theme_bw()
```



1.8 Part H

One of the observations is an outlier that occurred because of a transposition error. Do your best to identify this observation. The tens digit is wrong. Without changing the value of this observation, the estimated intraplant correlation is $\hat{\rho} = 0.2823$. If the tens digit is changed to correct the transposition error, the plot above looks much better and $\hat{\rho} = 0.5199$. How do you suggest that the outlier be modified, keeping in mind that a data analyst should never to this without consulting with the experimentalist.

This question is basically answered in the prompt. We know that a data analyst should never to modify an experiment without consulting with the experimentalist. So that is what we should do first. We should identify the outlier which seems to be the result of *Plant 2 Plot 1* as something in the 20s seems very low. Now before making this change, we need to talk with the experimentalist and tell them that we found something that we believe to be an error. After discussing with them, we should ask them what steps we should do next (i.e. if we should change the result to the correct value) before actually doing it. Then once we get their approval we should change our result and do our analysis again. If they say do not change it, then do not do so but maybe make a note of the outlier and the suspicion that an error could have occurred but not confirmed (if the experimentalist agrees this is the best way to proceed). Again we need to remember we are working with/for the experimentalist so we need to proceed in the way they want, so we should always double check and ask for approval before jumping into changes.

1.9 Part I (Not Graded)

Repeat problems (b), (c), (d) and (f) after correcting the transposition error mentioned in (h).

First we need to modify the data. Change the result to what it should be and remodel.

```r
# Correct data
protein_transform2 <- protein_transform
protein_transform2[4, 3] <- 38.6

# Make a new model
# Make the model
model2 <- lm(content ~ Plant, protein_transform2)

# Get anova table
(anova_model2 <- anova(model2))
```

```
## Analysis of Variance Table
##
## Response: content
##           Df Sum Sq Mean Sq F value   Pr(>F)
## Plant      1 27.969 27.9686  9.2404 0.005089 **
## Residuals 28 84.750  3.0268
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Test for a genetic component to protein content. That is, test $H_0 : Var(T_i) = 0$.
  - As we can see our F-value $F = \frac{MS(T)}{MS(E)} = 9.2404$ with 2, 27 degrees of freedom. We can see our p-value is 0.0050894 which is lower than most reasonable significance levels we can decide. Thus, we reject our H_0 and can say that with statistically significant evidence we can say that $Var(T_i) \neq 0$.
- We need to find A, A^{-1}, and $\hat{\theta}$.
  - Now we know that $A\theta = MS$. Therefore to find just A, we can say that $A = MS\,\theta'$. We know that $MS = \begin{pmatrix} MS(T) \\ MS(E) \end{pmatrix}$ which from **Part C** directions we can plug in $MS(T) = 3\sigma_T^2 + \sigma^2$ and $MS(E) = \sigma^2$. Thus, $MS = \begin{pmatrix} MS(T) \\ MS(E) \end{pmatrix} = \begin{pmatrix} 3\sigma_T^2 + \sigma^2 \\ \sigma^2 \end{pmatrix}$. Now we know that $\theta = \begin{pmatrix} \sigma_T^2 & \sigma^2 \end{pmatrix}$. Now we can let us plug in and then solve for A. Our base equation is $A\theta = MS$. Plugging everything in we get $\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} \sigma_T^2 & \sigma^2 \end{pmatrix} = \begin{pmatrix} 3\sigma_T^2 + \sigma^2 \\ \sigma^2 \end{pmatrix}$. Using what we know about matrix multiplication we can get that this turns into equations $a_{11}\sigma_T^2 + a_{12}\sigma^2 = 3\sigma_T^2 + \sigma^2$ and $a_{21}\sigma_T^2 + a_{22}\sigma^2 = \sigma^2$. We can clearly solve for the following to be true which is $a_{11} = 3, a_{12} = 1, a_{21} = 0, a_{22} = 1$. Thus, $A = \begin{pmatrix} 3 & 1 \\ 0 & 1 \end{pmatrix}$ Now we can find A' as the inverse of A. Thus, $A^{-1} = \begin{pmatrix} 3 & 1 \\ 0 & 1 \end{pmatrix}^{-1} = \frac{1}{det\begin{pmatrix} 3 & 1 \\ 0 & 1 \end{pmatrix}} \begin{pmatrix} 1 & -1 \\ 0 & 3 \end{pmatrix} = \begin{pmatrix} \frac{1}{3} & \frac{-1}{3} \\ 0 & 1 \end{pmatrix}$. Now we can find $\hat{\theta} = A^{-1}MS = \begin{pmatrix} \frac{1}{3} & \frac{-1}{3} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 3\hat{\sigma_T^2} + \hat{\sigma^2} \\ \hat{\sigma^2} \end{pmatrix} = \begin{pmatrix} \hat{\sigma_T^2} \\ \hat{\sigma^2} \end{pmatrix}$ which makes sense since we expect $\hat{\theta} = \begin{pmatrix} \hat{\sigma_T^2} \\ \hat{\sigma^2} \end{pmatrix}$ because $\theta = \begin{pmatrix} \sigma_T^2 \\ \sigma^2 \end{pmatrix}$. In this case $\hat{\sigma^2} = MS(E) = 3.0268$ and $\hat{\sigma_T^2} = \frac{MS(T) - MS(E)}{n} = \frac{27.9686 - 3.0268}{3} = 8.313933$ so $\hat{\theta} = \begin{pmatrix} \hat{\sigma_T^2} \\ \hat{\sigma^2} \end{pmatrix} = \begin{pmatrix} 8.313933 \\ 3.0268 \end{pmatrix}$.
- Estimate the coefficient of variation among protein contents.

- Note we just found that $\hat{\sigma_T^2} = 8.313933$. We also need to know $\hat{\mu} = E(\hat{Y_{ij}}) =$ `mean(protein_transform2$content)` $= 39.4266667$. The coefficient of variation among protein contents is $CV(Y_{ij}) = \frac{\sqrt{\sigma_T^2+\sigma^2}}{|\hat{\mu}|} = \frac{\sqrt{\sigma_T^2+\sigma^2}}{|\bar{y_{\cdot\cdot}}|} = \frac{\sqrt{8.313933+3.0268}}{|39.42667|} = 0.08541428$. Thus, the coefficient of variation among protein contents is 0.08541428 or 8.541428%.

- Report a 95% confidence interval for the mean protein content μ of seeds grown from a randomly sampled Plant.

  - We know this confidence interval formula is $\bar{y_{\cdot\cdot}} \pm t(0.025, t-1)\sqrt{\frac{MS(T)}{nt}}$. We know that $t = 10, n = 3, \bar{y_{\cdot\cdot}} = 39.42667, MS(T) = 27.9686$ so our confidence interval is $39.42667 \pm t(0.025, 10-1)\sqrt{\frac{27.9686}{3*10}} ==> 39.42667 \pm t(0.025, 9)\sqrt{\frac{27.9686}{30}} ===> 39.42667 \pm 2.262157 * 0.9655499 ===> 39.42667 \pm 2.184225$ which gives us a lower interval of 37.242445 and upper confidence interval of 41.610895. Thus our 95% confidence interval for the mean protein content μ of seeds grown from a randomly sampled Plant is 37.242445 to 41.610895.

2 Problem 2

This problem expands on our work with sample size and power calculations and makes explicit use of `PROC GLMPOWER` in **SAS**. The non-central F-distributions of F ratios for factorial effects can be quantified in a manner similar to what we did with single factor experiments. Consider designing a 2 ×4 factorial experiment with the following means:

Factor A	Factor B-1	Factor B-2	Factor B-3	Factor B-4	Marginal Means
1	100	108	116	116	110
2	86	94	102	110	98
Marginal Mean	93	101	109	113	104

2.1 Part A

Use `PROC GLMPOWER` code like that given above to obtain the power to detect

- The A × B interaction effects
- The main effects of A

when $r = 3, 4, 5$ and $\sigma = 6$ or $\sigma = 10$. Enter the powers in tables as below

Power to detect A and B interaction

σ	r = 3	r = 4	r = 5
6	.159		
10			.120

Power to detect main effects of A

(a)

```
data one;
   do a=1 to 2 ; do b= 1 to 4;
      input muij @;
      output;
   end;end;
datalines;
100 108 116 116 86 94 102 110
;
run;
proc glmpower;
   class a b;
   model muij=a|b;
   power
      stddev=6 to 10 by 4
      ntotal=24 to 40 by 8
      power=.;
run;
```

Figure 1: Code to Run

σ	r = 3	r = 4	r = 5
6	.996		>.999
10			.957

When we run this code, we get the following output table:

We can use this to fill in our table. When `N Total` is 24, $r = 3$; when `N Total` is 32, $r = 4$; when `N Total` is 40, $r = 5$. When `Source = a` this is looking at the main effects of factor a; when `Source = b` this is looking at the main effects of factor b; When `Source = a*b` this is looking at $A * B$ interaction. We will now use this information and table above to fill in our table in an appropriate way.

Power to detect A and B interaction

σ	r = 3	r = 4	r = 5
6	.159	.215	.271
10	.086	.103	.120

Power to detect main effects of A

σ	r = 3	r = 4	r = 5
6	.996	>.999	>.999
10	.788	.902	.957

2.2 Part B

The mean in the i^{th} row and j^{th} column of the table above is given by $\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$. Solve for all 15 parameters under the constraints that $\sum_i \alpha_i = 0$, $\sum_j \beta_j = 0$, and interaction effects sum to zero within any row or column. Partial solution given below:

- $\mu = 104$
- $\alpha_1 = 6$
- $\beta_1 = -11$
- $\beta_2 = -3$
- $(\alpha\beta)_{11} = 1$
- $(\alpha\beta)_{12} = 1$

Computed Power						
Index	Source	Std Dev	N Total	Test DF	Error DF	Power
1	a	6	24	1	16	0.996
2	a	6	32	1	24	>.999
3	a	6	40	1	32	>.999
4	a	10	24	1	16	0.788
5	a	10	32	1	24	0.902
6	a	10	40	1	32	0.957
7	b	6	24	3	16	0.998
8	b	6	32	3	24	>.999
9	b	6	40	3	32	>.999
10	b	10	24	3	16	0.806
11	b	10	32	3	24	0.934
12	b	10	40	3	32	0.980
13	a*b	6	24	3	16	0.159
14	a*b	6	32	3	24	0.215
15	a*b	6	40	3	32	0.271
16	a*b	10	24	3	16	0.086
17	a*b	10	32	3	24	0.103
18	a*b	10	40	3	32	0.120

Figure 2: Power Calculations Output Table

column by having $(\alpha\beta)_{11} + (\alpha\beta)_{12} + (\alpha\beta)_{13} + (\alpha\beta)_{14} = 0 ===> 1 + 1 + 1 - 3 = 0 ===> 0 = 0$ is true so this term is confirmed.

We can now follow up to find $(\alpha\beta)_{23}$. We can do this because interaction effects sum to zero within any row or column. So $(\alpha\beta)_{23} = -(\alpha\beta)_{13} = -1$. We can check this by $(\alpha\beta)_{21} + (\alpha\beta)_{22} + (\alpha\beta)_{23} + (\alpha\beta)_{24} = 0 ===> -1 - 1 - 1 + 3 = 0 ===> 0 = 0$ is true. Also check $\bar{y_2}3 = \mu + \alpha_2 + \beta_3 + (\alpha\beta)_{23} ===> 102 = 104 - 6 + 5 + (\alpha\beta)_{23} ===> 102 = 103 + (\alpha\beta)_{23} ===> (\alpha\beta)_{23} = -1$. Therefore, we have confirmed that $(\alpha\beta)_{23} = -1$ is true.

Now we have found all of them such that the final values for all of them are:

- $\mu = 104$
- $\alpha_1 = 6$
- $\alpha_2 = -6$
- $\beta_1 = -11$
- $\beta_2 = -3$
- $\beta_3 = 5$
- $\beta_4 = 9$
- $(\alpha\beta)_{11} = 1$
- $(\alpha\beta)_{12} = 1$
- $(\alpha\beta)_{13} = 1$
- $(\alpha\beta)_{14} = -3$
- $(\alpha\beta)_{21} = -1$
- $(\alpha\beta)_{22} = -1$
- $(\alpha\beta)_{23} = -1$
- $(\alpha\beta)_{24} = 3$