Exercise 8.5: Blocking on the rat diet dataset - solution

Lieven Clement and Jeroen Gilis

statOmics, Ghent University (https://statomics.github.io)

Contents

1 Background

Researchers are studying the impact of protein sources and protein levels in the diet on the weight of rats. They feed the rats with diets of beef, cereal and pork and use a low and high protein level for each diet type. The researchers can include 60 rats in the experiment. Prior to the experiment, the rats were divided in 10 homogeneous groups of 6 rats based on characteristics such as initial weight, appetite, etc.

Within each group a rat is randomly assigned to a diet. The rats are fed during a month and the weight gain in grams is recorded for each rat.

The researchers want to assess the effect of the type of diet and the protein level on the weight of the rats.

In this exercise we will perform the data exploration using all diets, but, to keep the data analysis simple we will only assess the beef and cereal diets.

2 Experimental design

- There are three explanatory variables in the experiment: the factor diet type with two levels (beef and cereal), factor protein level with levels (low and high) and a group blocking factor with 10 levels.
- There are 6 treatments: beef-high, cereal-high, pork-high, beef-low, cereal-low, pork-low protein.
- The rats are the experimental units (the unit to which a treatment is applied): in this design, there is a randomisation restriction: Within a block, a rat is randomly assigned to a diet.
- The rats are the observational units (the unit on which the response is measured): The weight is weighted for each rat.
- The weight gain is the response variable.
- The experiment is a randomized complete block (RCB) design

Load libraries

```
library(tidyverse)
```

3 Data import

```
diet <- read.table("https://raw.githubusercontent.com/statOmics/PSLS21/data/dietRats.txt",
                   header=TRUE)
head(diet)
```

```
##   weightGain protSource protLevel block
## 1        107          b         h     1
## 2         96          c         h     1
## 3        112          p         h     1
## 4         83          b         l     1
## 5         87          c         l     1
## 6         90          p         l     1
```

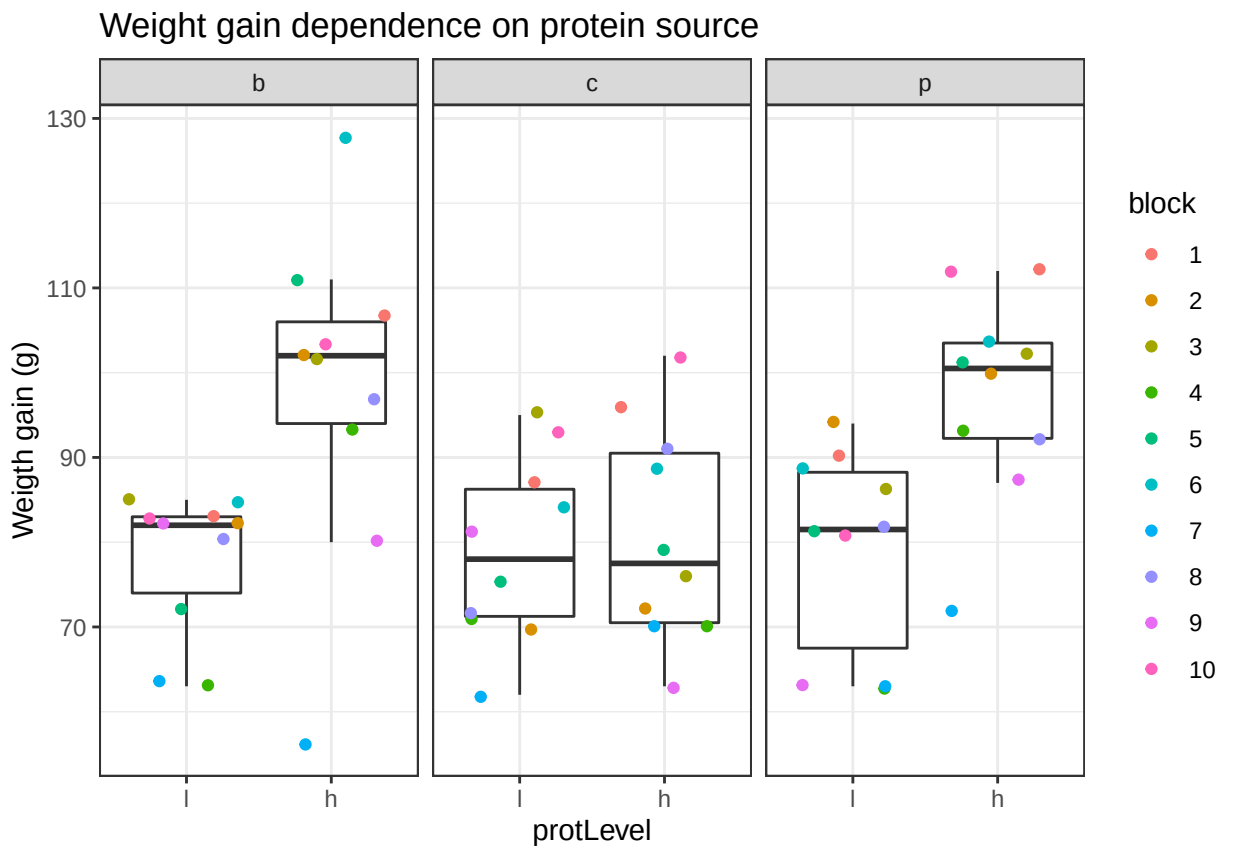
4 Tidy data

```
diet <- diet %>%
  mutate(block = as.factor(block),
         protSource = as.factor(protSource),
         protLevel = as.factor(protLevel)) %>%
  mutate(protLevel = fct_relevel(protLevel, "l"))
```

5 Data exploration

- Boxplot of the weight gain against protein source, protein level with coloring according to block

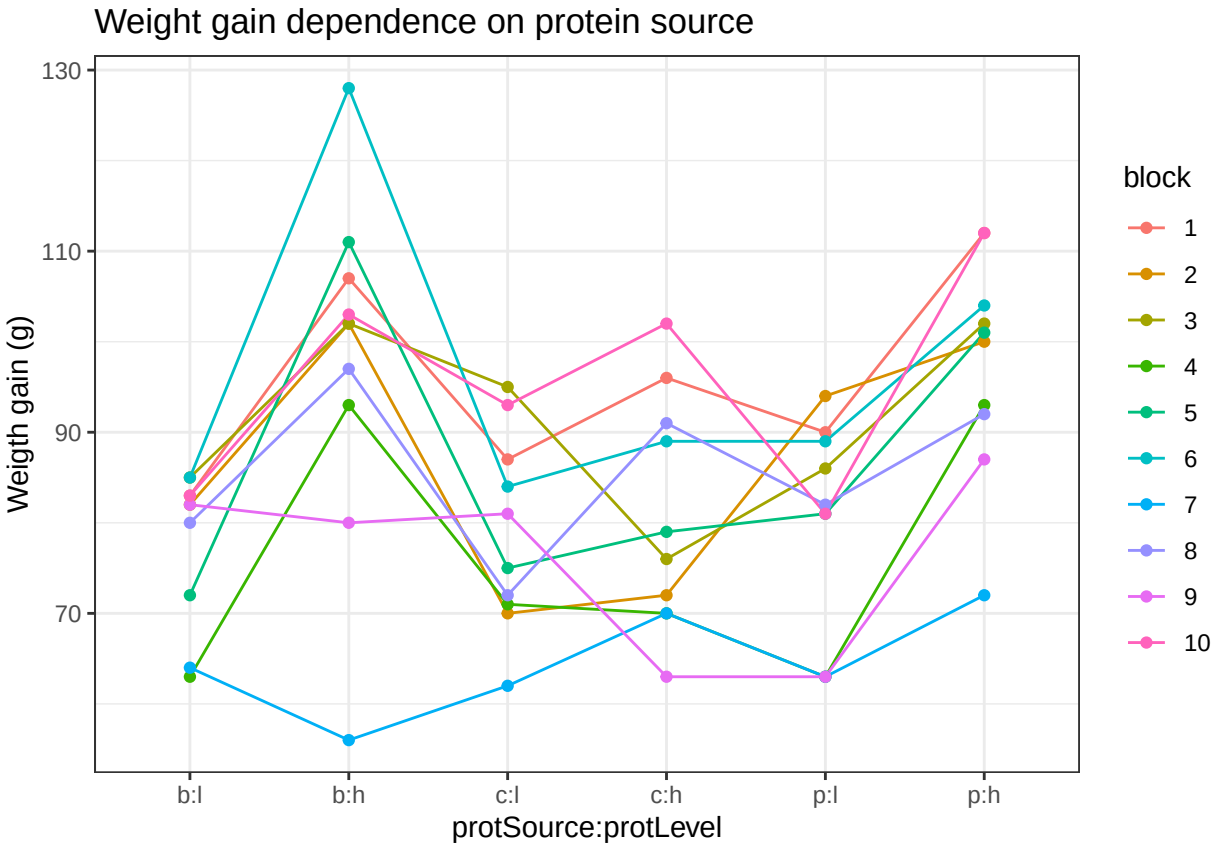
```
diet %>%
  ggplot(aes(x=protLevel, y=weightGain)) +
    scale_fill_brewer(palette="RdGy") +
    theme_bw() +
    geom_boxplot(outlier.shape=NA) +
    geom_jitter(aes(color=block)) +
    ggtitle("Weight gain dependence on protein source") +
    ylab("Weigth gain (g)") +
    #stat_summary(fun = mean, geom="point", shape=5, size=3, color="black", fill="black") +
    facet_wrap(.~protSource)
```



- Lineplot of the weight gain against protein source, protein level with coloring and grouping according to block

```
diet %>%
  ggplot(aes(x=protSource:protLevel, y=weightGain)) +
    scale_fill_brewer(palette="RdGy") +
    theme_bw() +
    geom_line(aes(group=block, color=block)) +
    geom_point(aes(color=block))+
```

```
    ggtitle("Weight gain dependence on protein source") +
    ylab("Weigth gain (g)")
```




```
    #stat_summary(fun = mean, geom="point", shape=5, size=3, color="black", fill="black")
```

- An increase in the weight of the rats seems to depend on the protein source received in the diet.
- The increase in the weight of the rats seems to depend on the level of protein received in the diet
- There also seems to be an interaction effect between the protein level and the protein source on the gain in weight of the rats. For the beef and the pork diets the effect of high protein levels in the data seems to be much stronger than in the cereal diet.
- There is also a strong effect of the block. Blocking implies a randomisation restriction, hence, we will have to include the block effect anyway.

6 Filter the data to only use the beef and cereal diet

```
diet.bc <- diet %>% filter(protSource != "p")
```

7 Multivariate linear regression analysis

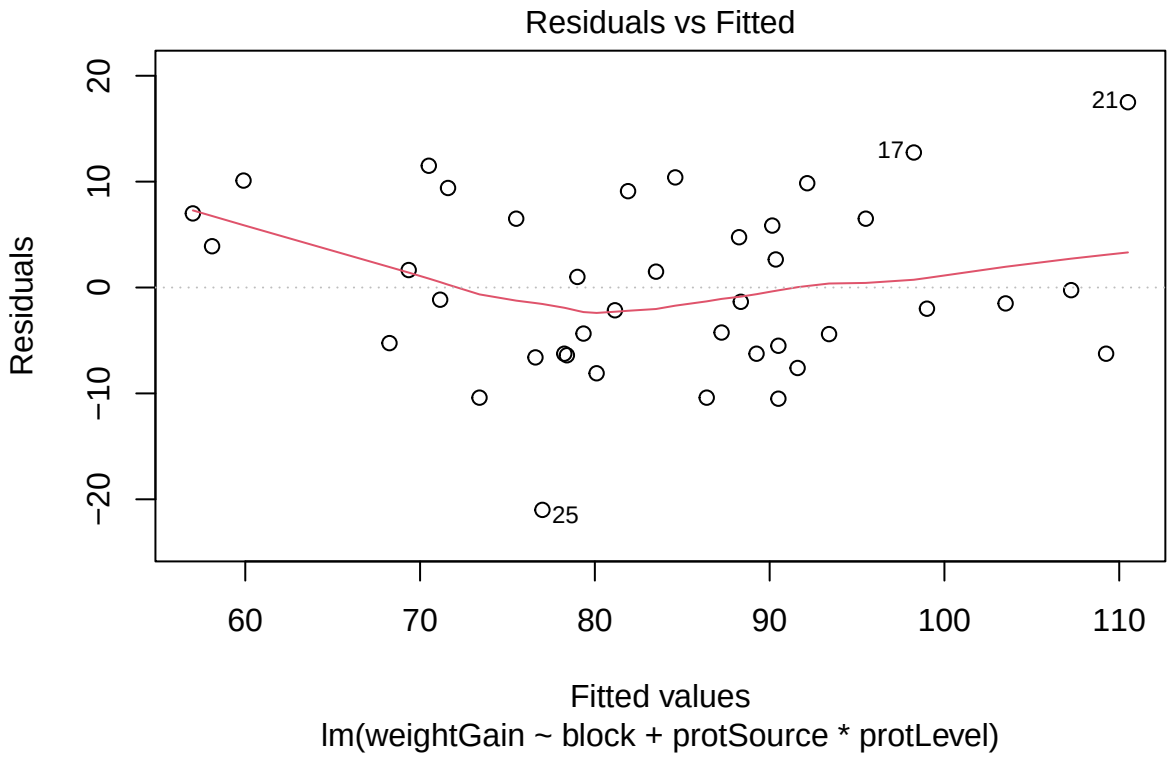
7.1 Assumptions

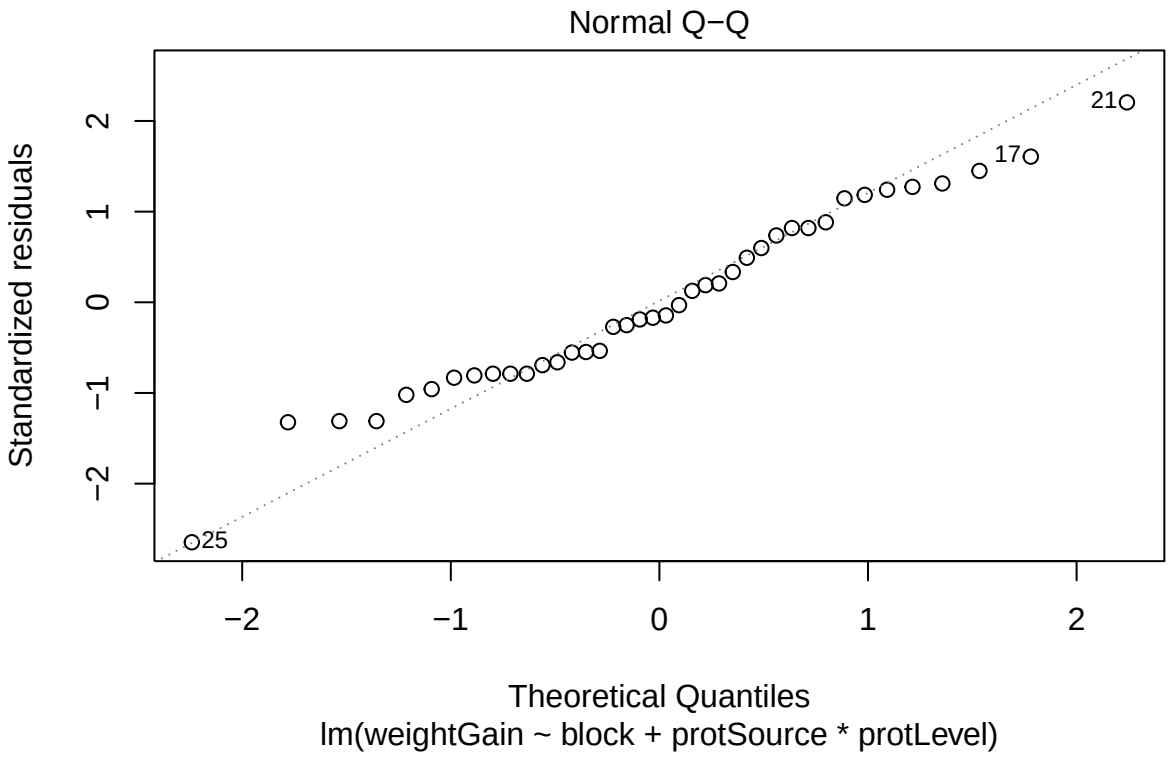
List assumptions:

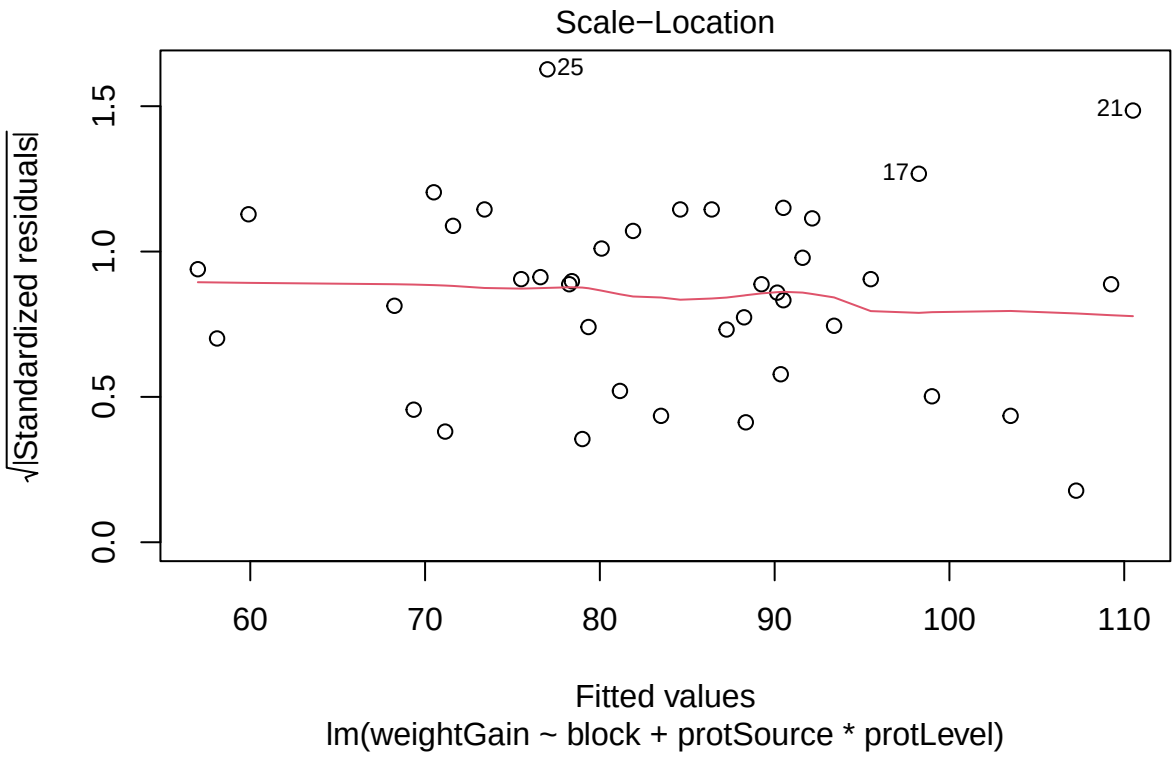
1. The observations are independent
2. Linearity between the response and predictor variable
3. The residuals of the model must be normally distributed
4. Homoscedasticity of the data

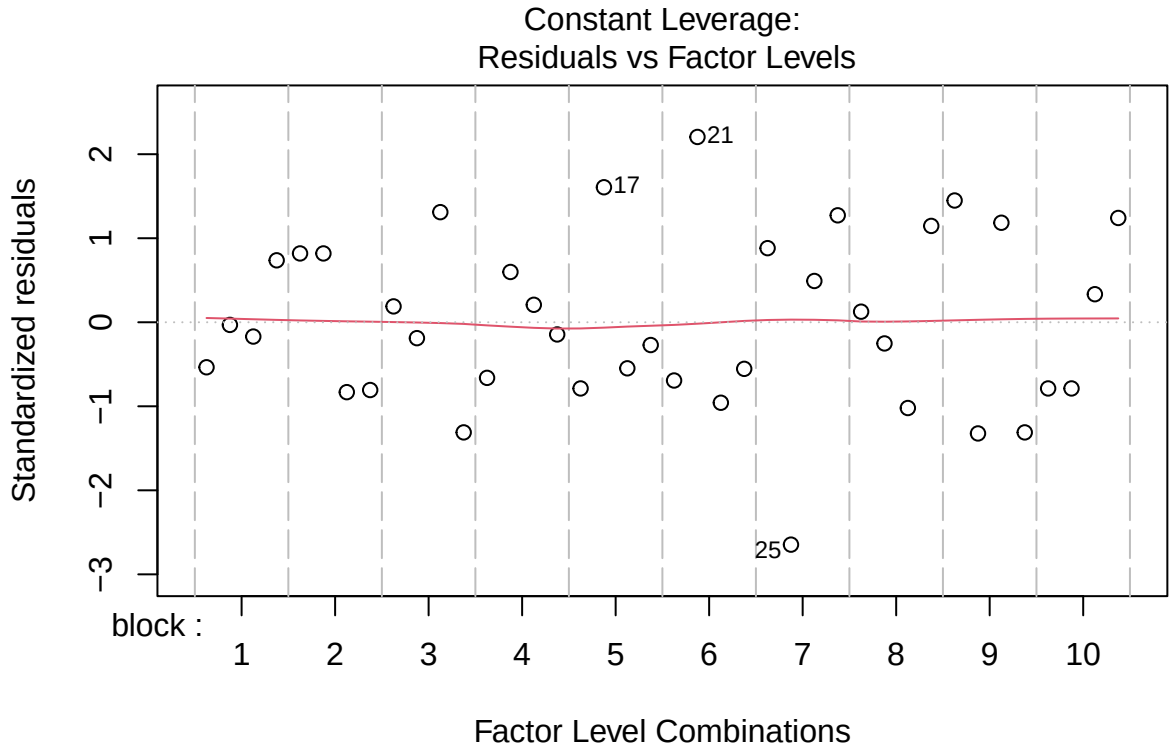
The first assumption is met if we correct for block in the model because the rats were randomized to the treatment within block. The other three assumptions can be assessed by fitting the linear model and calling the `plot()` function as follows.

```
lm1 <- lm(weightGain ~ block + protSource*protLevel, data=diet.bc)
plot(lm1)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(weightGain ~ block + protSource * protLevel)
25
17
21
−2
−1
0
1
2


Scale−Location
25
21
17
√|Standardized residuals|
0.0
0.5
1.0
1.5
60
70
80
90
100
110
Fitted values
lm(weightGain ~ block + protSource * protLevel)



All assumptions are met for this dataset.

7.2 Hypothesis testing

We here fit a linear model with a blocking factor for `block` and main and interaction effects for protein source and protein level.

```
summary(lm1)
```

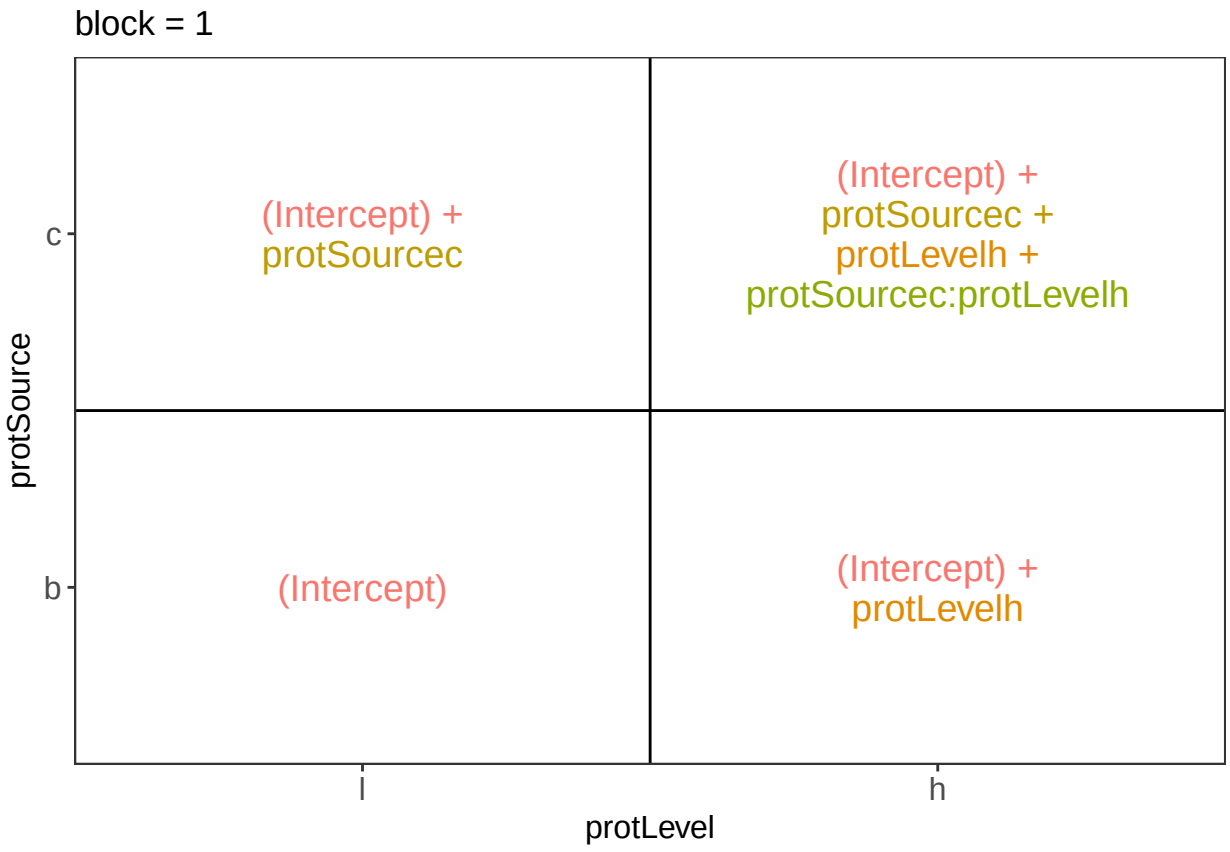
```
##
## Call:
## lm(formula = weightGain ~ block + protSource * protLevel, data = diet.bc)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21.00  -6.25  -1.25   6.50  17.50
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             87.250      5.506  15.846 3.39e-15 ***
## block2                 -11.750      6.830  -1.720 0.096797 .
## block3                  -3.750      6.830  -0.549 0.587467
## block4                 -19.000      6.830  -2.782 0.009735 **
## block5                  -9.000      6.830  -1.318 0.198650
## block6                   3.250      6.830   0.476 0.637999
```

```
## block7                   -30.250      6.830  -4.429 0.000141 ***
## block8                    -8.250      6.830  -1.208 0.237536
## block9                   -16.750      6.830  -2.453 0.020933 *
## block10                    2.000      6.830   0.293 0.771883
## protSourcec                1.100      4.319   0.255 0.800914
## protLevelh                20.000      4.319   4.630 8.23e-05 ***
## protSourcec:protLevelh   -18.200      6.109  -2.979 0.006043 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.659 on 27 degrees of freedom
## Multiple R-squared:  0.7252, Adjusted R-squared:  0.6031
## F-statistic: 5.939 on 12 and 27 DF,  p-value: 6.122e-05
```

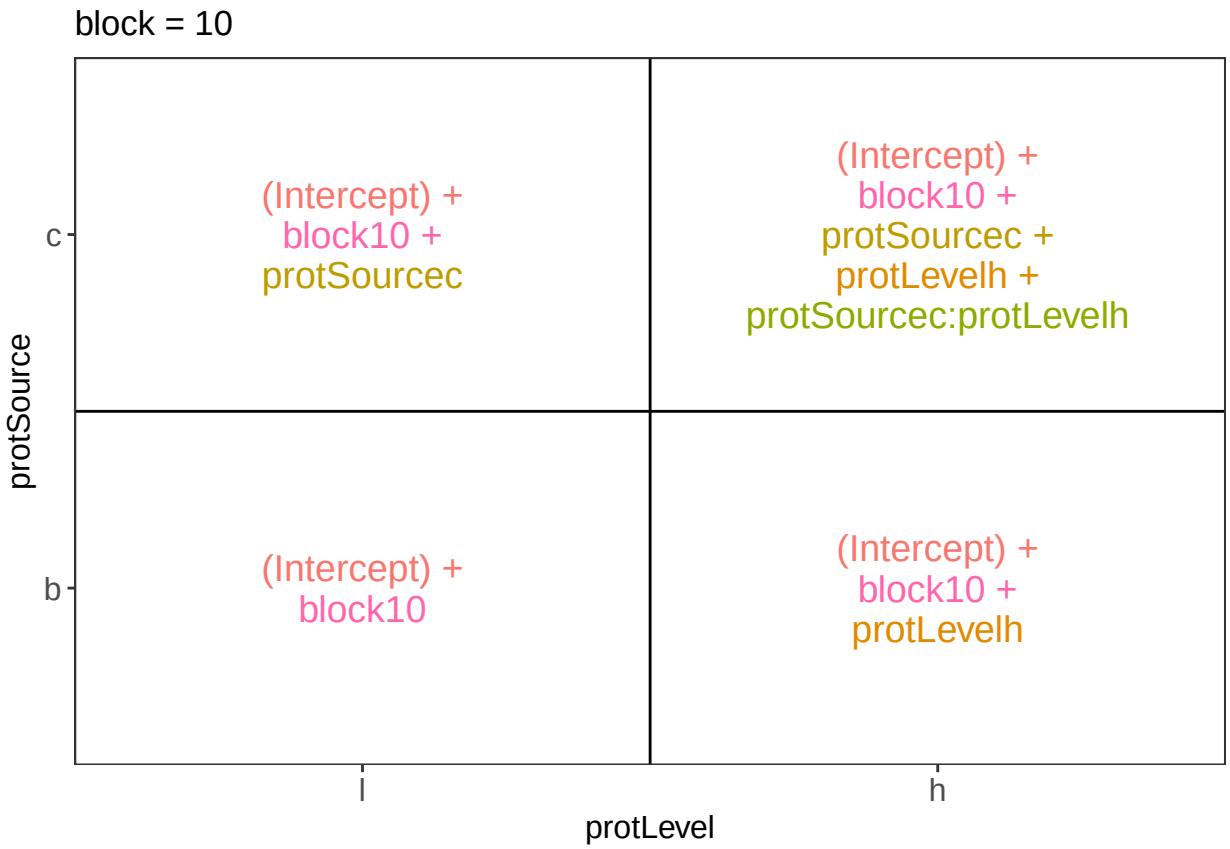
7.3 Interpretation of the regression parameters

```
library(ExploreModelMatrix)
ExploreModelMatrix::VisualizeDesign(diet.bc, ~ block + protSource * protLevel)$plotlist
```

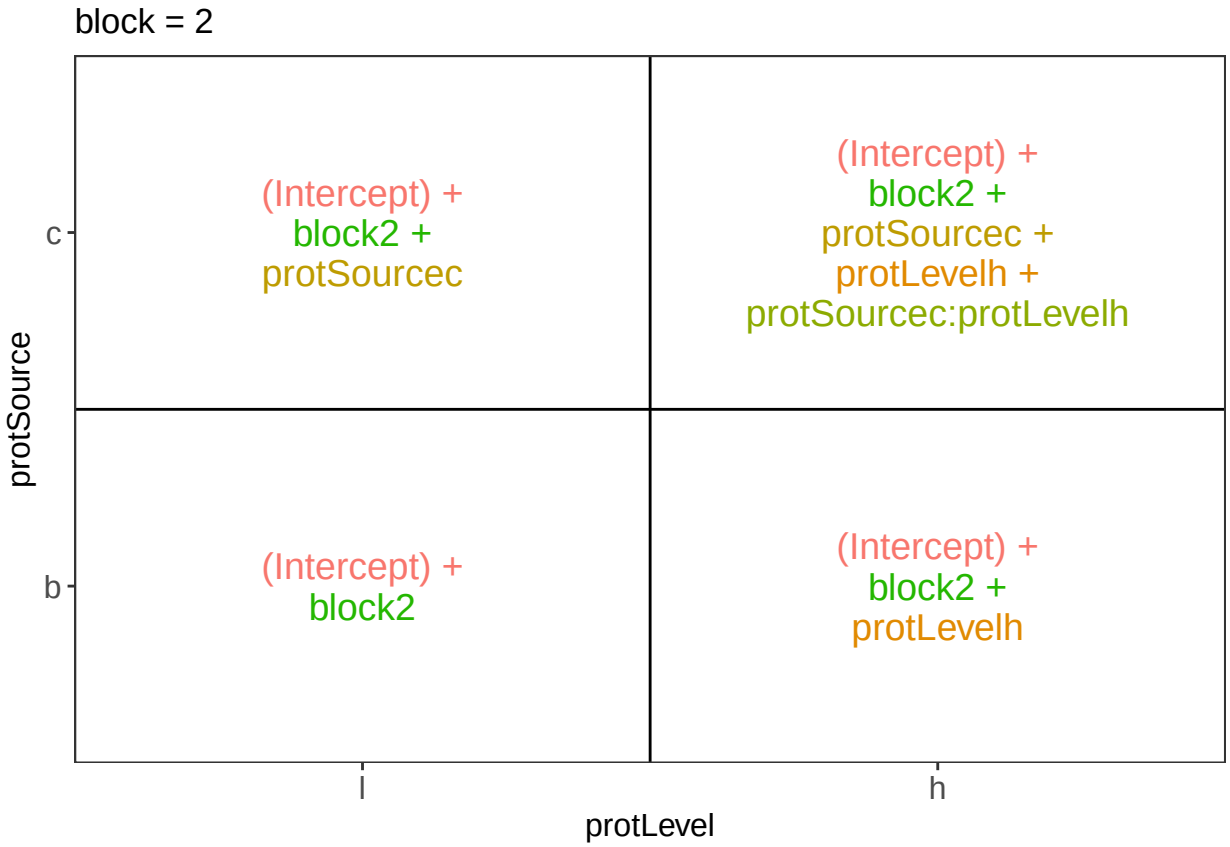
```
## $`block = 1`
```



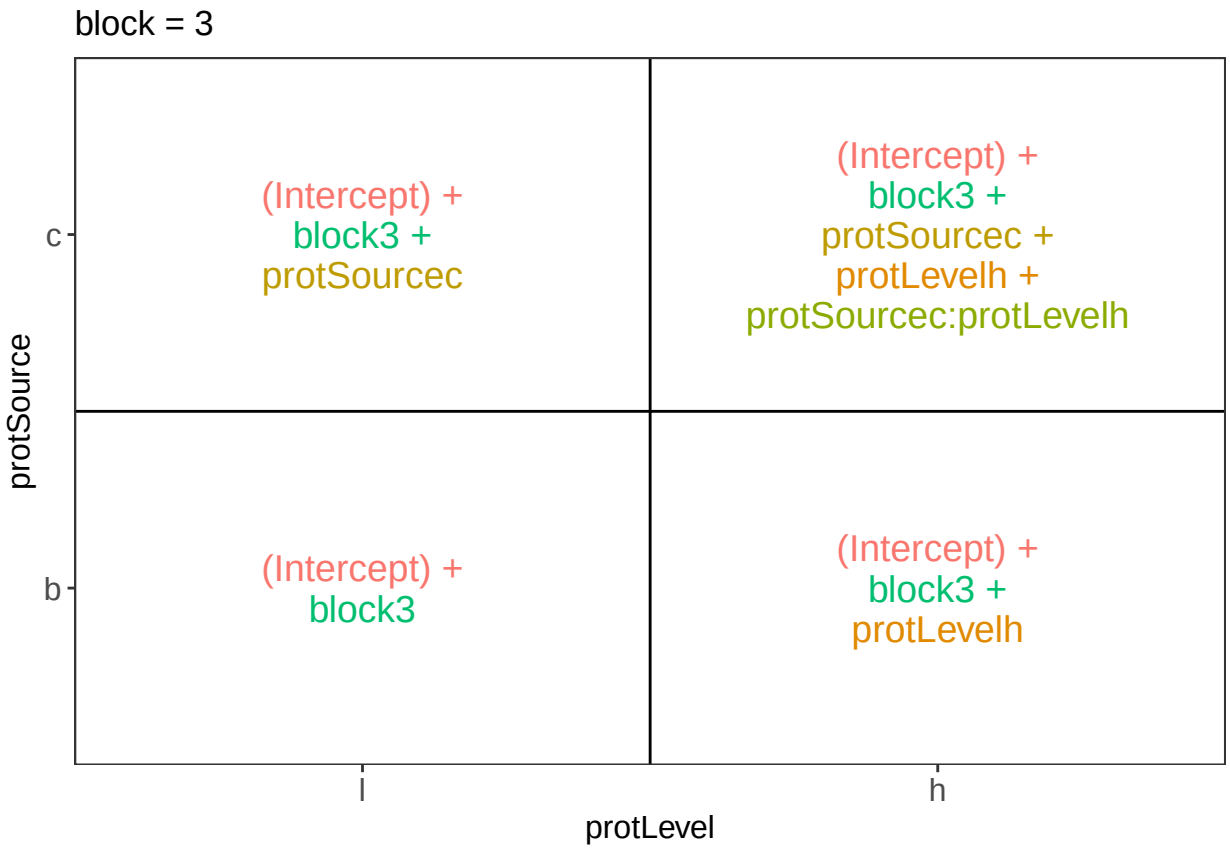

```
##
## $`block = 10`
```


block = 10
(Intercept) +
block10 +
protSourcec
(Intercept) +
block10 +
protSourcec +
protLevelh +
protSourcec:protLevelh
(Intercept) +
block10
(Intercept) +
block10 +
protLevelh
protSource
c
b
l
h
protLevel

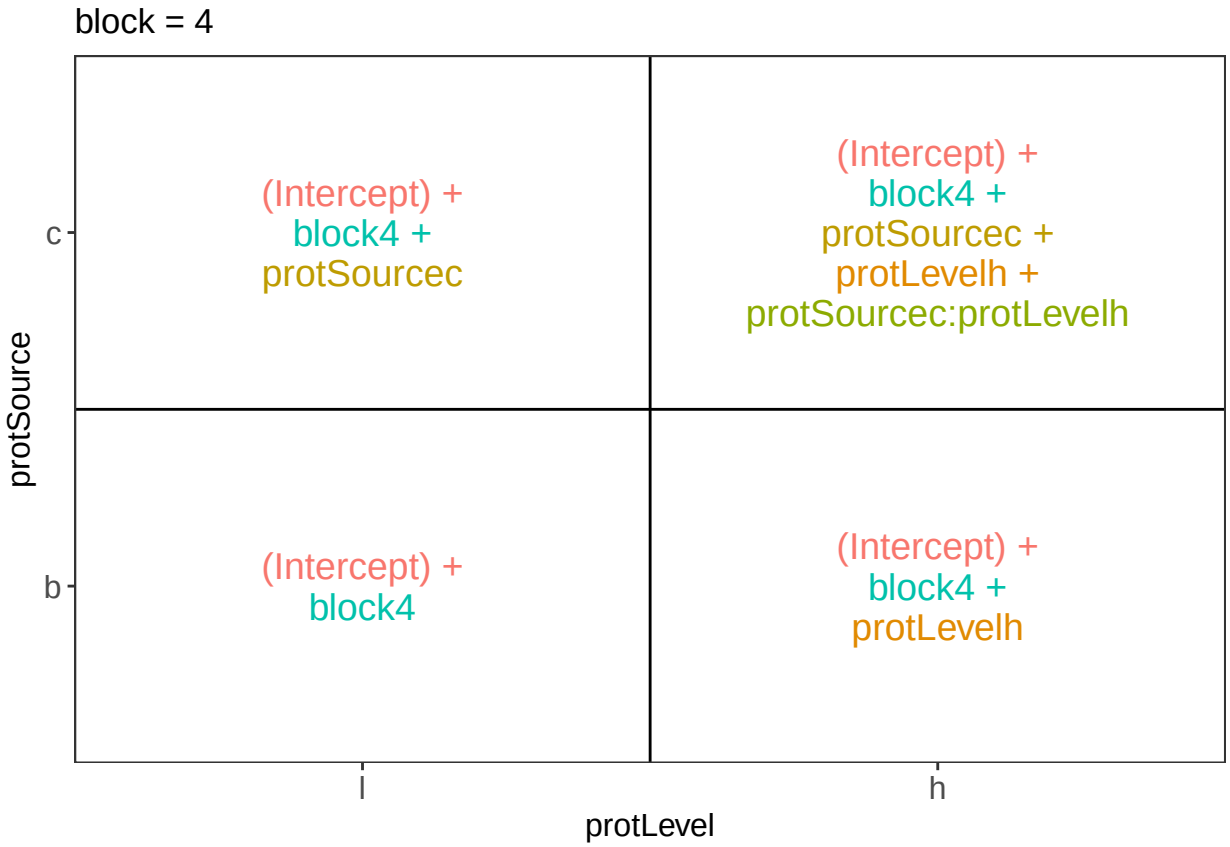
```
##
## $`block = 2`
```


block = 2
(Intercept) +
block2 +
protSourcec
(Intercept) +
block2 +
protSourcec +
protLevelh +
protSourcec:protLevelh
(Intercept) +
block2
(Intercept) +
block2 +
protLevelh
protSource
c
b
l
h
protLevel

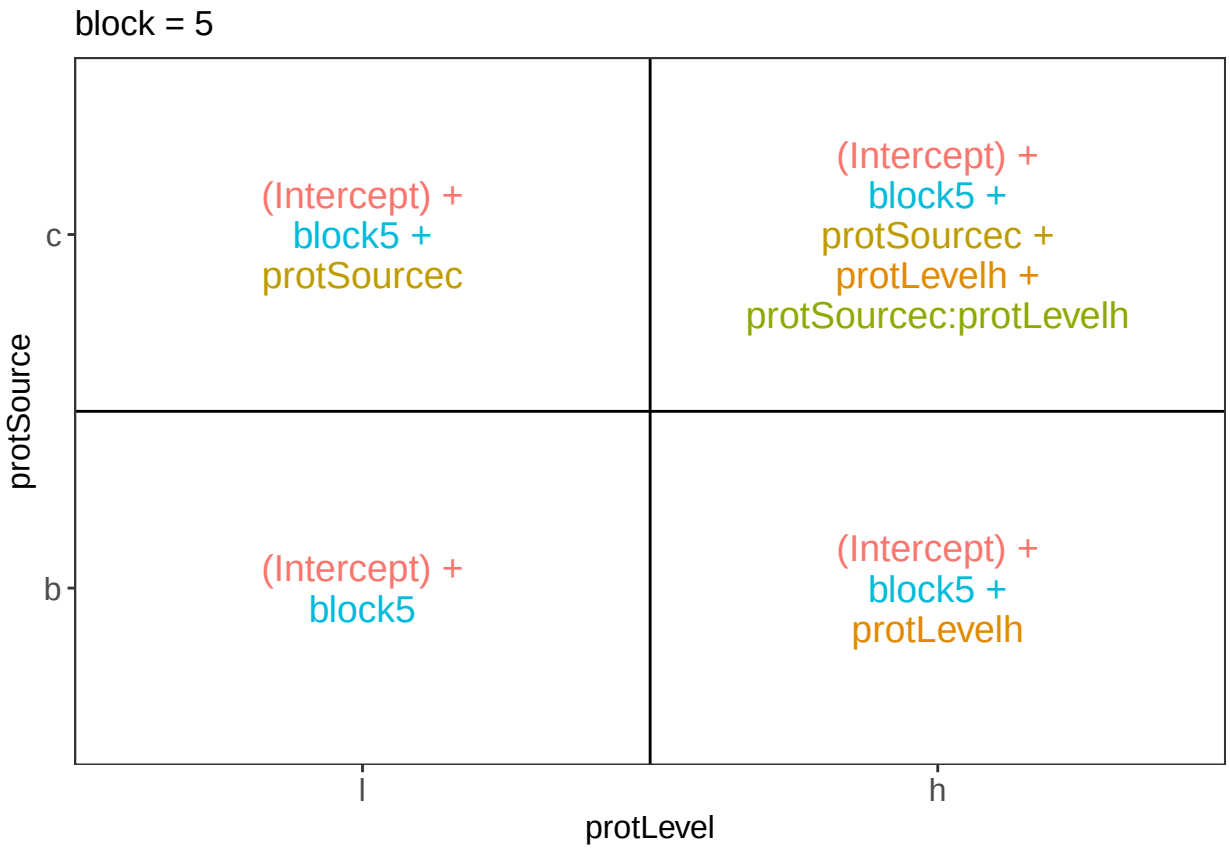
```
##
## $`block = 3`
```


block = 3
(Intercept) +
block3 +
protSourcec
(Intercept) +
block3 +
protSourcec +
protLevelh +
protSourcec:protLevelh
(Intercept) +
block3
(Intercept) +
block3 +
protLevelh
protSource
c
b
l
h
protLevel

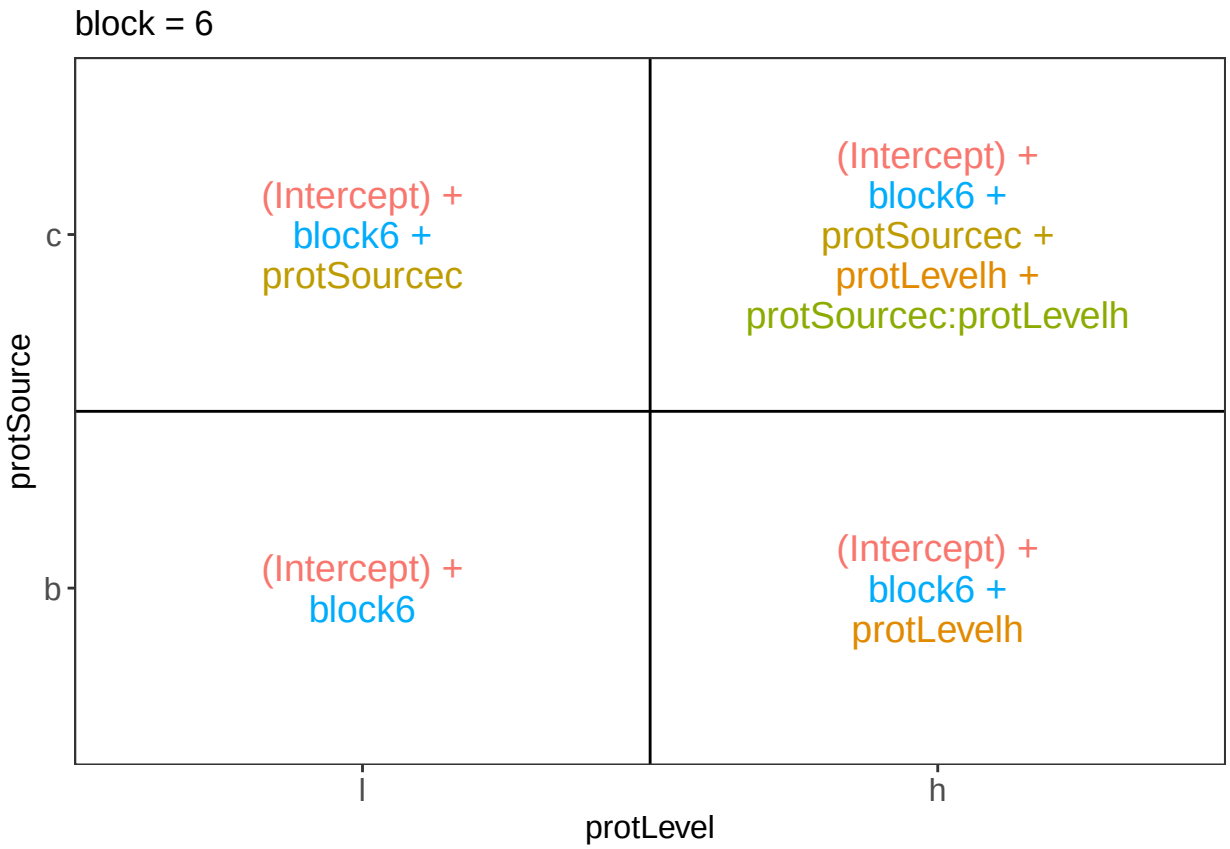
```
##
## $`block = 4`
```


block = 4
(Intercept) +
block4 +
protSourcec
(Intercept) +
block4 +
protSourcec +
protLevelh +
protSourcec:protLevelh
(Intercept) +
block4
(Intercept) +
block4 +
protLevelh
protSource
c
b
l
h
protLevel

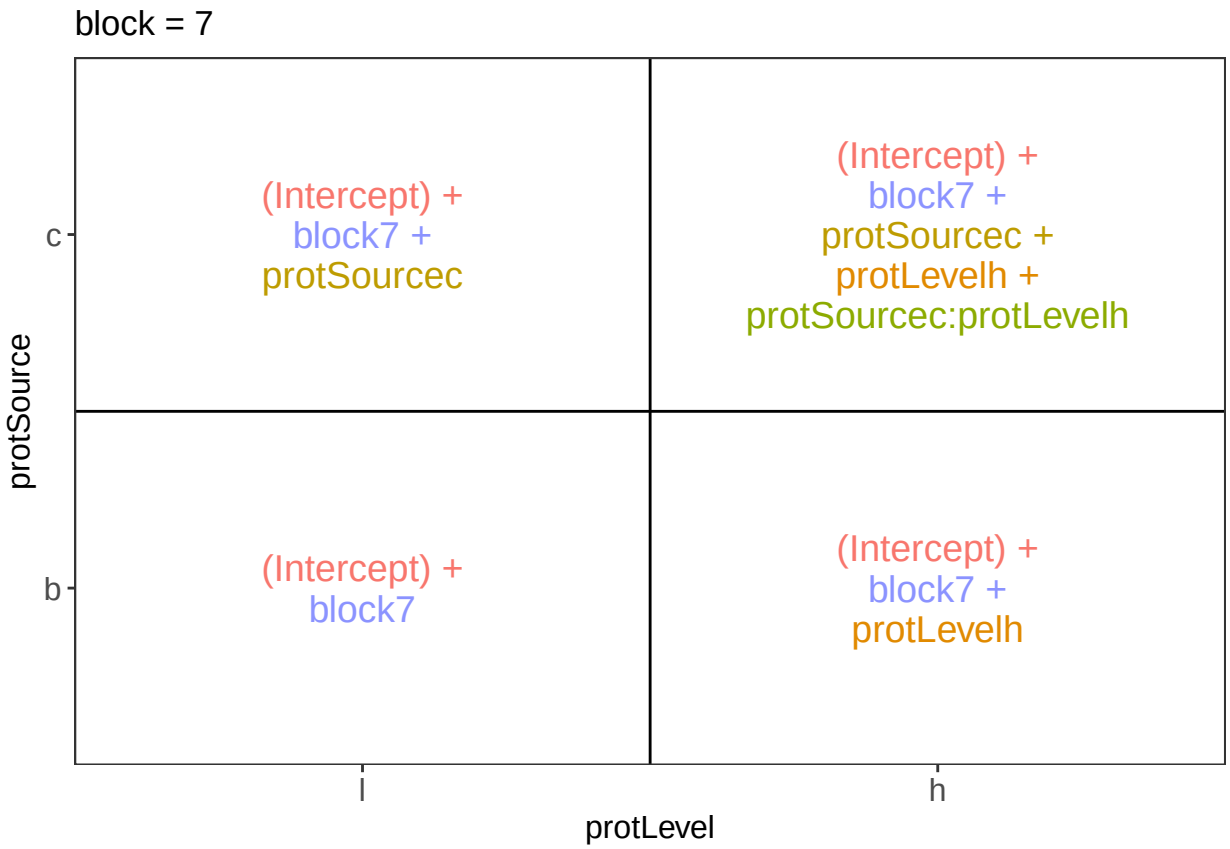
```
##
## $`block = 5`
```


block = 5
(Intercept) +
block5 +
protSourcec
(Intercept) +
block5 +
protSourcec +
protLevelh +
protSourcec:protLevelh
(Intercept) +
block5
(Intercept) +
block5 +
protLevelh
protSource
c
b
l
h
protLevel

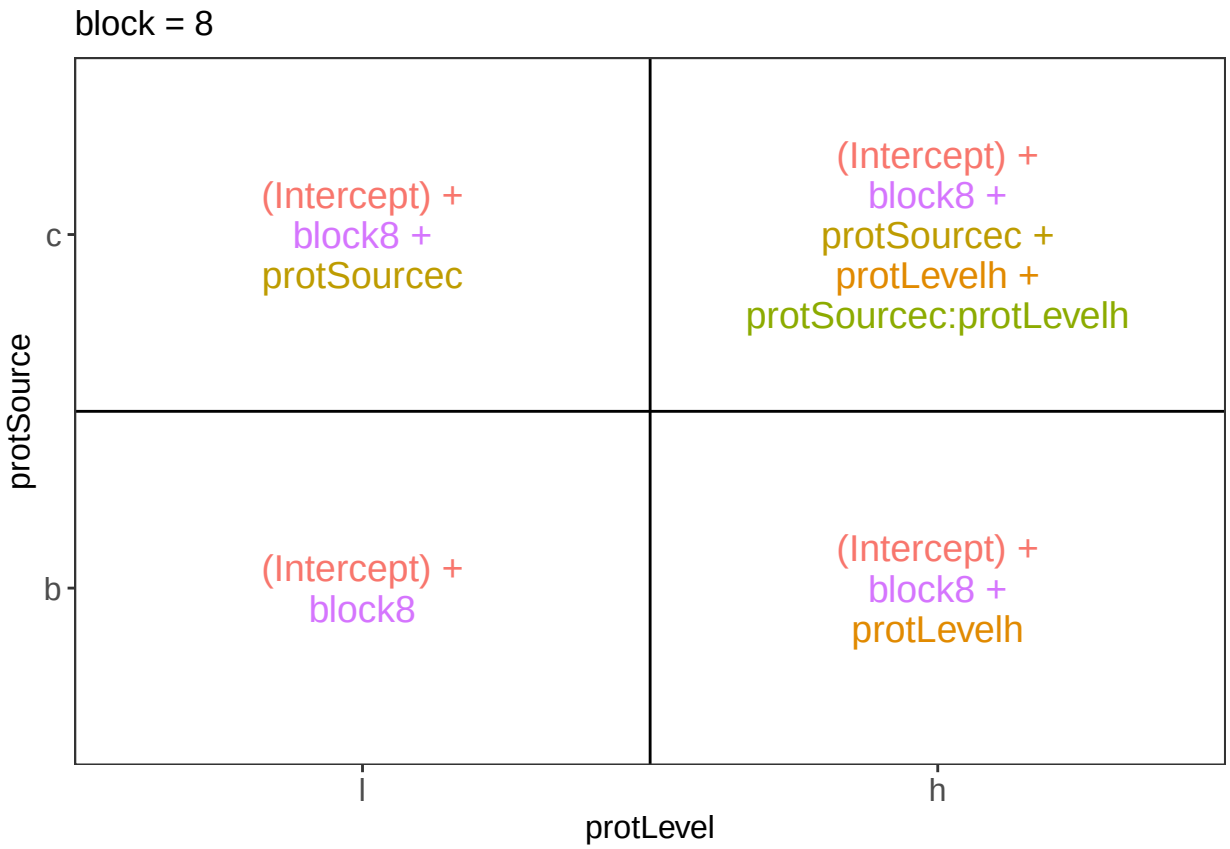
```
##
## $`block = 6`
```


block = 6
(Intercept) +
block6 +
protSourcec
(Intercept) +
block6 +
protSourcec +
protLevelh +
protSourcec:protLevelh
(Intercept) +
block6
(Intercept) +
block6 +
protLevelh
protSource
c
b
l
h
protLevel

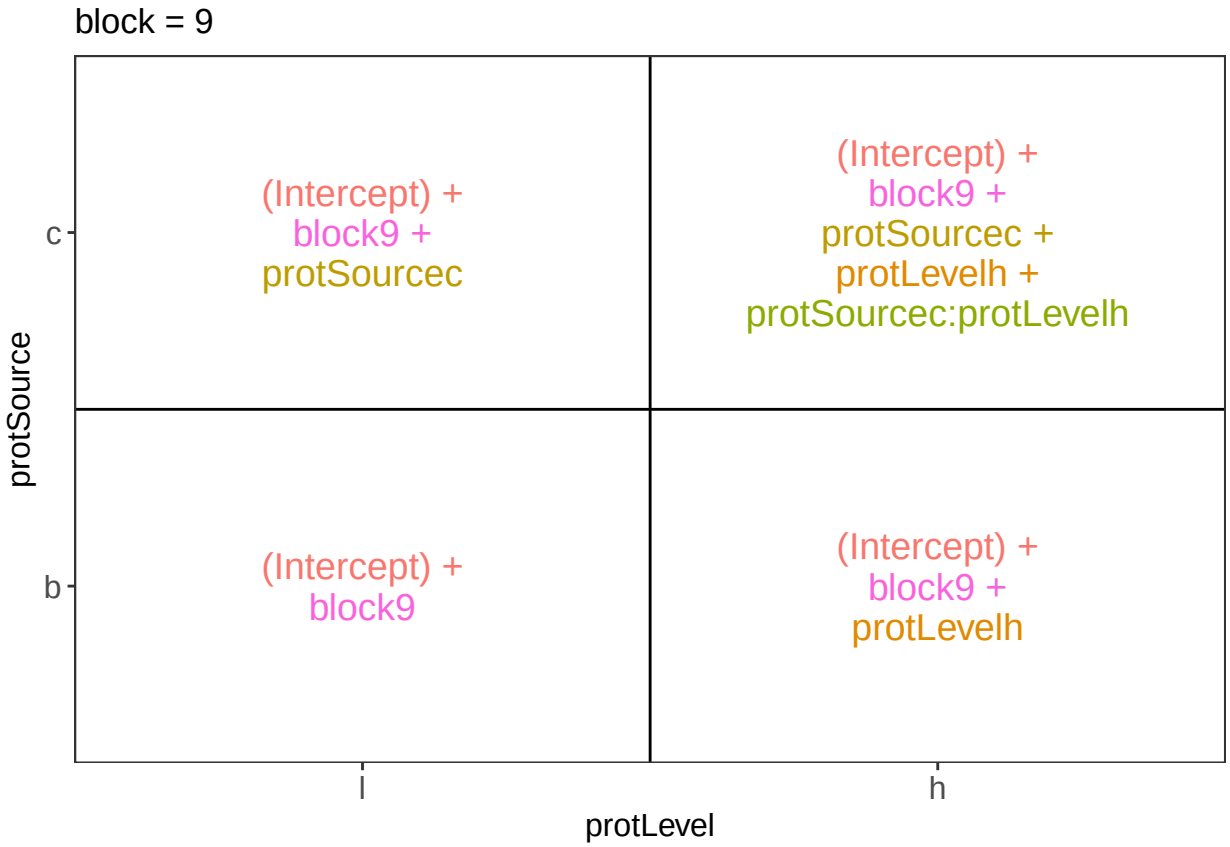
```
##
## $`block = 7`
```


block = 7
(Intercept) +
block7 +
protSourcec
(Intercept) +
block7 +
protSourcec +
protLevelh +
protSourcec:protLevelh
(Intercept) +
block7
(Intercept) +
block7 +
protLevelh
protSource
c
b
l
h
protLevel

```
##
## $`block = 8`
```


block = 8
(Intercept) +
block8 +
protSourcec
(Intercept) +
block8 +
protSourcec +
protLevelh +
protSourcec:protLevelh
(Intercept) +
block8
(Intercept) +
block8 +
protLevelh
protSource
c
b
l
h
protLevel

```
##
## $`block = 9`
```



There are 3 levels for protein source, 2 levels for protein level and 10 levels for the blocking variable. We will have one reference level for each respective variable: beef, low, block 1. So we need 2, 1 and 9 dummy variables to introduce the factors protein source, protein level and block in the linear model, respectively.

Hence, we can write down the linear model as follows:

$y_i = \beta_0 + \beta_c x_{i,c} + \beta_h x_{i,h} + \beta_{ch} x_{i,c} x_{i,h} + \beta_{b2} x_{i,b2} + ... + \beta_{b10} x_{i,b10} + \epsilon_i$

with: y_i the observed weight gain for rat i, $x_{i,h}$ a dummy variable which is 1 if rat i receives a high protein diet and is 0 otherwise,
$x_{i,c}$ a dummy variable which is 1 if rat i receives a cereal diet and is 0 otherwise,
$x_{i,bk}$ is a dummy variable which is 1 if rat i belongs to block bk and is 0 otherwise, with $k \in 2, ..., 10$, and ϵ_i an error term which is normally distributed with mean 0 and variance σ^2, i.e. $\epsilon_i \sim N(0, \sigma^2)$.

- Rats that are assigned to block k and receive a beef based low protein diet have a covariate pattern $x_{i,h} = 0$, $x_{i,c} = 0$, $x_{i,bm} = 0$ with $m \neq k$ and $x_{i,bk} = 1$. Their mean weight gain is thus equal to $\mu_{l,b,bk} = \beta_0 + \beta_{bk}$
- Rats that are assigned to block k and receive a beef based high protein diet have a covariate pattern $x_{i,h} = 1$, $x_{i,c} = 0$, $x_{i,bm} = 0$ with $m \neq k$ and $x_{i,bk} = 1$. Their mean weight gain is thus equal to $\mu_{h,b,bk} = \beta_0 + \beta_h + \beta_{bk}$
- Rats that are assigned to block k and receive a cereal based low protein diet have a covariate pattern $x_{i,h} = 0$, $x_{i,c} = 1$, $x_{i,bm} = 0$ with $m \neq k$ and $x_{i,bk} = 1$. Their mean weight gain is thus equal to $\mu_{h,c,bk} = \beta_0 + \beta_c + \beta_{bk}$
- Rats that are assigned to block k and receive a cereal based heigh protein diet have a covariate pattern $x_{i,h} = 1$, $x_{i,c} = 1$ $x_{i,bm} = 0$ with $m \neq k$ and $x_{i,bk} = 1$. Their mean weight gain is thus equal to $\mu_{h,c,bk} = \beta_0 + \beta_h + \beta_c + \beta_{ch} + \beta_{bk}$

We can now relate this to the output of the `lm` function:

- The intercept β_0 is thus the average weight gain in the low beef diet for rats in block 1.
- The parameter β_c: the average weight gain difference between cereal-low and beef-low diet is 1.1g.
- The parameter β_h: the average weight gain difference between beef-high and beef-Low diet is 20g.
- The parameter β_{ch} is the difference in the average weight gain difference due to the high protein level as compared to the low protein level for cereal diets as compared to the weight gain difference that occurs due to the protein level in the reference class (here beef diet). Here this is negative, i.e. -18.2g, thus the weight gain for the cereal protein source increases on average less between high and low protein diets than in beef based diets.

7.4 Testing the overall (combined) effect of diet

Because there are multiple factors with different levels in the model, we can first assess the effect of the diet (protein Level, protein source and the interaction) by using anova. With this test we will assess the null hypothesis that the average weight gain in each treatment is equal: i.e. $H_0 : \mu_{b,l} = \mu_{b,h} = \mu_{c,h} = \mu_{c,l}$ versus the alternative hyptohesis H_1 : that at least two treatment means are different.

```
lm0 <- lm(weightGain ~ block, data=diet.bc)
anova(lm0, lm1)
```

```
## Analysis of Variance Table
##
## Model 1: weightGain ~ block
## Model 2: weightGain ~ block + protSource * protLevel
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     30 5175.0
## 2     27 2518.8  3    2656.2 9.4909 0.000189 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can conclude that there is an very significant effect of the diet type (protein source and/or protein level and/or protein source-protein level interaction) on the weight gain of rats (p = 2e-04).

7.5 Assessing the interaction effect between protein source and protein level

```
library(car)
summary(lm1)
```

```
##
## Call:
## lm(formula = weightGain ~ block + protSource * protLevel, data = diet.bc)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21.00  -6.25  -1.25   6.50  17.50
##
## Coefficients:
```

```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               87.250      5.506  15.846 3.39e-15 ***
## block2                   -11.750      6.830  -1.720 0.096797 .
## block3                    -3.750      6.830  -0.549 0.587467
## block4                   -19.000      6.830  -2.782 0.009735 **
## block5                    -9.000      6.830  -1.318 0.198650
## block6                     3.250      6.830   0.476 0.637999
## block7                   -30.250      6.830  -4.429 0.000141 ***
## block8                    -8.250      6.830  -1.208 0.237536
## block9                   -16.750      6.830  -2.453 0.020933 *
## block10                    2.000      6.830   0.293 0.771883
## protSourcec                1.100      4.319   0.255 0.800914
## protLevelh                20.000      4.319   4.630 8.23e-05 ***
## protSourcec:protLevelh   -18.200      6.109  -2.979 0.006043 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.659 on 27 degrees of freedom
## Multiple R-squared:  0.7252, Adjusted R-squared:  0.6031
## F-statistic: 5.939 on 12 and 27 DF,  p-value: 6.122e-05
```

```
Anova(lm1,type="III")
```

```
## Anova Table (Type III tests)
##
## Response: weightGain
##                         Sum Sq Df  F value    Pr(>F)
## (Intercept)          23423.3  1 251.0832 3.387e-15 ***
## block                 3992.6  9   4.7554 0.0007546 ***
## protSource               6.1  1   0.0649 0.8009145
## protLevel             2000.0  1  21.4388 8.228e-05 ***
## protSource:protLevel   828.1  1   8.8767 0.0060432 **
## Residuals             2518.8 27
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is a very significant interaction between the protein source and the protein level. This indicates that the average weight increase due to the protein level differs according to the protein source. Hence, we cannot assess the effect of the protein source and/or protein level independently because there effects of the protein source vary according to the protein level.

7.6 Assessing specific contrasts

Imagine that we are interested in assessing if there is an effect of

1. protein source in the low protein diets
    - $\mu_{c,l} - \mu_{b,l} = \beta_c$
2. protein source in high protein diets
    - $\mu_{c,h} - \mu_{b,h} = \beta_c + \beta_{ch}$

3. protein level for beef diets ($\mu_{b,h} - \mu_{b,l} = \beta_h$), and cereal diets ($\mu_{c,h} - \mu_{c,l} = \beta_h + \beta_{ch}$)
4. if the effect of the protein level differs between
    - beef and cereal $(\mu_{c,h} - \mu_{c,l}) - (\mu_{b,h} - \mu_{b,l}) = \beta_{ch}$

These effects of interest are so-called **contrasts, i.e. linear combinations of the parameters**.

We can define the contrasts and assess the significance of the contrasts with the code below. The contrasts are given as input in the form of symbolic descriptions to the `linfct` argument of the `glht` function.

```
library(multcomp)
set.seed(75468) # to get reproducible results (small effect if removed)
lm1MultComp <- glht(
    model = lm1,
    linfct = c("protSourcec = 0",
               "protSourcec + protSourcec:protLevelh = 0",
               "protLevelh = 0",
               "protLevelh +  protSourcec:protLevelh = 0",
               "protSourcec:protLevelh = 0")
)
```

```
summary(lm1MultComp)
```

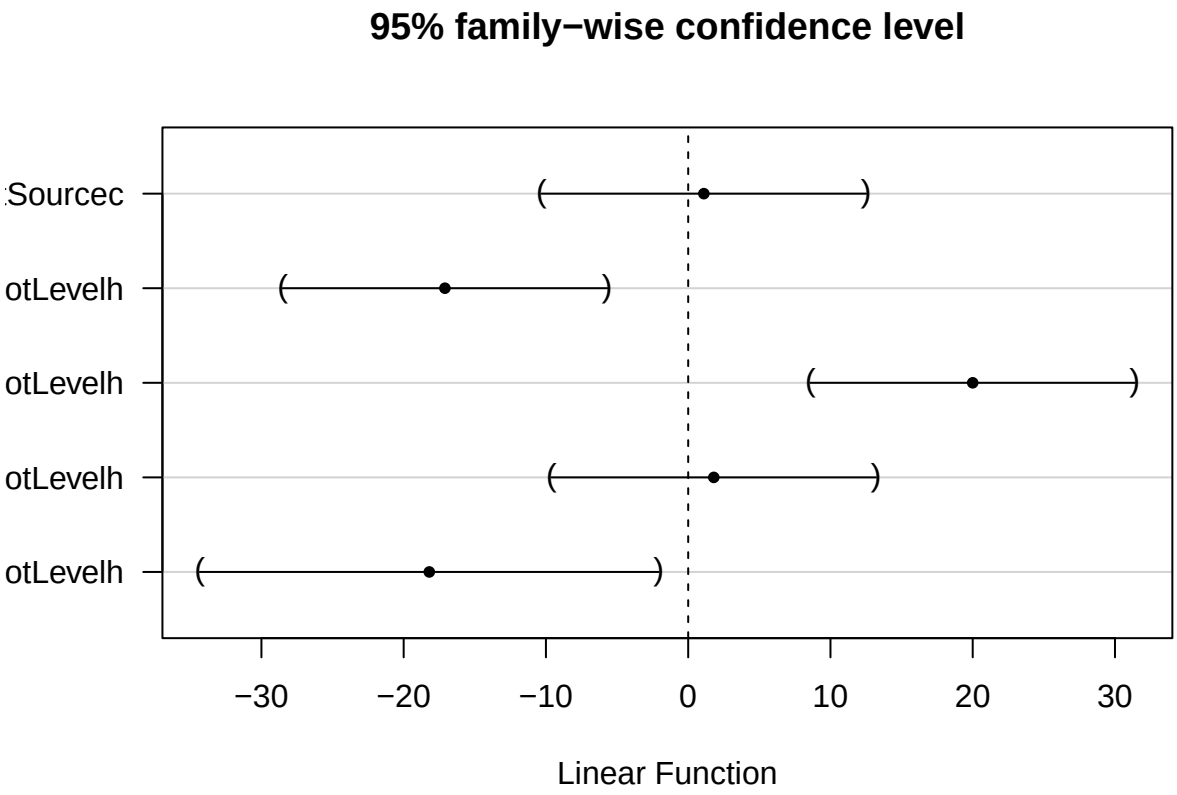
```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = weightGain ~ block + protSource * protLevel, data = diet.bc)
##
## Linear Hypotheses:
##                                          Estimate Std. Error t value Pr(>|t|)
## protSourcec == 0                            1.100      4.319   0.255  0.99235
## protSourcec + protSourcec:protLevelh == 0  -17.100      4.319  -3.959  0.00211
## protLevelh == 0                             20.000      4.319   4.630  < 0.001
## protLevelh + protSourcec:protLevelh == 0     1.800      4.319   0.417  0.96838
## protSourcec:protLevelh == 0                -18.200      6.109  -2.979  0.02351
##
## protSourcec == 0
## protSourcec + protSourcec:protLevelh == 0 **
## protLevelh == 0                           ***
## protLevelh + protSourcec:protLevelh == 0
## protSourcec:protLevelh == 0               *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

```
confint(lm1MultComp)
```

```
##
##   Simultaneous Confidence Intervals
##
## Fit: lm(formula = weightGain ~ block + protSource * protLevel, data = diet.bc)
##
```

```
## Quantile = 2.6386
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##                                            Estimate lwr      upr
## protSourcec == 0                             1.1000 -10.2975  12.4975
## protSourcec + protSourcec:protLevelh == 0 -17.1000 -28.4975  -5.7025
## protLevelh == 0                            20.0000   8.6025  31.3975
## protLevelh + protSourcec:protLevelh == 0    1.8000  -9.5975  13.1975
## protSourcec:protLevelh == 0                -18.2000 -34.3185  -2.0815
```

```
plot(lm1MultComp)
```



Note that the p-values and the confidence intervals are automatically corrected for multiple testing.

8 Conclusion

- There is an extremely significant effect of the type of protein diet on the weight gain of rats (p«1e-3).
- The average weight gain does not vary significantly according to protein source in the diets with low protein levels (p = 0.99).
- The weight gain in the cereal diet at high protein concentration is on average 17.1g lower than in the high protein beef diet (95% CI [5.7, 28.5]) and the difference is very significant (p =

format(summary(lm1MultComp)$test$pvalues[names(summary(lm1MultComp)$test$tstat)==“protSourcec + protSourcec:protLevelh”],digits=1)).

- We also discovered an extremely significant difference in weight gain according to the protein level for beef based diets (p«0.001).
  The weight gain on average increases with 20g in the high protein level as compared to the low protein beef diet (95%CI [8.6, 31.4]). The protein level effect is not significant for the cereal diet (0.97).
- Finally there is a significant interaction between protein level and protein source (p = 0.023), i.e. the increase in weight gain due to protein level in beef based diets was 18.2g than that in the cerial diet (95% CI [2.1, 34.3]).

All reported p-values and confidence intervals were corrected for multiple testing