# STAT 516 sp 2024 final exam

**150 minutes, no calculators or notes allowed**

## 1. Choosing the correct model

You will refer to this list of models in parts a) through d):

1. $Y_{ijk} = \mu + \tau_i + \gamma_j + (\tau\gamma)_{ij} + C_k + (\tau C)_{ik} + \varepsilon_{ijk},\ i = 1, \ldots, a,\ j = 1, \ldots, b,\ k = 1, \ldots, c,$ $C_k \overset{\text{ind}}{\sim} \text{Normal}(0, \sigma^2_C),\ (\tau C)_{ik} \overset{\text{ind}}{\sim} \text{Normal}(0, \sigma^2_{AC}),\ \varepsilon_{ijk} \overset{\text{ind}}{\sim} \text{Normal}(0, \sigma^2_\varepsilon)$.
2. $Y_{ij} = \mu + A_i + \varepsilon_{ij},\ i = 1, \ldots, a,\ j = 1, \ldots, n_i,\ A_i \overset{\text{ind}}{\sim} \text{Normal}(0, \sigma^2_A),\ \varepsilon_{ij} \overset{\text{ind}}{\sim} \text{Normal}(0, \sigma^2_\varepsilon)$.
3. $Y_{ijk} = \mu + \tau_i + \gamma_j + (\tau\gamma)_{ij} + \varepsilon_{ijk},\ i = 1, \ldots, a,\ j = 1, \ldots, b,\ k = 1, \ldots, c,\ \varepsilon_{ijk} \overset{\text{ind}}{\sim} \text{Normal}(0, \sigma^2_\varepsilon)$.
4. $Y_{ij} = \mu + \tau_i + B_j + \varepsilon_{ij},\ i = 1, \ldots, a,\ j = 1, \ldots, b,\ B_j \overset{\text{ind}}{\sim} \text{Normal}(0, \sigma^2_B),\ \varepsilon_{ij} \overset{\text{ind}}{\sim} \text{Normal}(0, \sigma^2_\varepsilon)$.
5. $Y_{ij} = \mu + \tau_i + \beta_i x_{ij} + \varepsilon_{ij},\ i = 1, \ldots, a,\ j = 1, \ldots, n_i,\ \varepsilon_{ij} \overset{\text{ind}}{\sim} \text{Normal}(0, \sigma^2_\varepsilon)$.
6. $Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + \varepsilon_{ijk},\ i = 1, \ldots, a,\ j = 1, \ldots, b,\ k = 1, \ldots, c,\ A_i \overset{\text{ind}}{\sim} \text{Normal}(0, \sigma^2_A),\ B_j \overset{\text{ind}}{\sim} \text{Normal}(0, \sigma^2_B),\ (AB)_{ij} \overset{\text{ind}}{\sim} \text{Normal}(0, \sigma^2_{AB}),\ \varepsilon_{ijk} \overset{\text{ind}}{\sim} \text{Normal}(0, \sigma^2_\varepsilon)$.

You wish to understand under what conditions the hellebores plant will produce lots of flowers. Consider the following experiments:

### a)

You first wish to understand if there is a significant genetic component to differences in the amount of flowers a hellebores plant will produce. You [redacted] sample six hellebores plants and from each plant you grow three clones, so that you have three plants for each of six randomly sampled genotypes. You then raise the plants under more or less identical conditions in a greenhouse. On each plant, you obtain, at a certain age, a measure of the total volume of flowers produced. At the end of the experiment you have eighteen response values, three for each of the six unique genotypes.

i. Select the appropriate model from models 1–6 and describe in detail the role of each term in the model (By *each term in the model* I mean if there is a $\mu$ tell me what $\mu$ is; [redacted], etc.).

This is a one-way random effects model:

(2) $Y_{ij} = \mu + A_i + \varepsilon_{ij}$

$\mu$ is an overall or baseline mean

$A_i$ is the random effect of genotype $i$

$\varepsilon_{ij}$ is an error term

ii. State the null and alternate hypotheses of interest.

$H_0: \sigma^2_A = 0$ vs $H_1: \sigma^2_A > 0$

iii. Give the numerator and denominator degrees of freedom of the F distribution used in testing your hypotheses.

The ANOVA table will look like:

| Source | df | SS | MS | F |
|---|---|---|---|---|
| Genotype | 6-1 | | | |
| Error | 6(3-1)=12 | | | |
| Total | 18-1 | | | |

So the numerator and denominator dfs will be 5 and 12.

**b)**

You now wish to understand the effects of two different fertilizers on the volume of flowers a hellebores plant produces. You again sample six hellebores plants and from each plant you grow three clones, so that you have three plants for each of six randomly sampled genotypes. Within each set of clones, you randomly assign one to fertilizer A, one to fertilizer B, and one to receive no fertilizer. After a period of time, you obtain a measurement on each plant of the total volume of flowers it produced. This results in eighteen total response values.

i. Give the name of the experimental design.

This is a randomized complete block design with the genotypes as the blocks.

ii. Select the appropriate model from models 1–6 and describe in detail the role of each term in the model.

(4) $Y_{ij} = \mu + \tau_i + B_j + \varepsilon_{ij}$,

$\mu$ is an overall or baseline mean

$\tau_i$ is the effect of fertilizer treatment $i$

$B_j$ is the random effect of the genotype.

$\varepsilon_{ij}$ is an error term

iii. State the null and alternate hypotheses corresponding to the question of whether the fertilization treatments make any difference to the volume of flowers produced.

Let $\mu_i = \mu + \tau_i$ for $i = 1, 2, 3$.

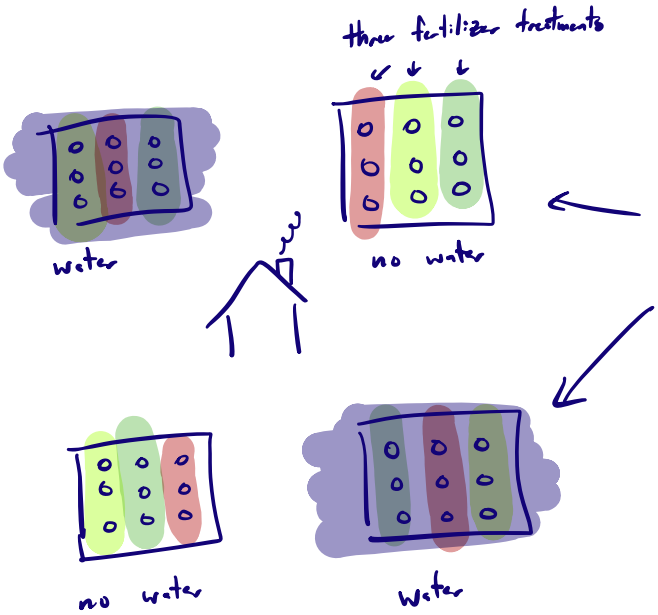
Then we wish to test $H_0: \mu_1 = \mu_2 = \mu_3$ vs

$H_1$: not all $\mu_i$ are equal.

**c)**

Your friend, who lacks access to a greenhouse, has similar research questions. She wishes to understand whether watering the hellebores plants, in addition to whether and how one fertilizes them, will make any difference to the total volume of flowers they produce. In each of four raised beds, two in her front yard and two in her back yard, she plants nine hellebores plants in three rows of three. She randomly assigns one of the raised beds in the front yard and one in the back yard to regular watering and the other to no watering. Within each bed, she assigns the three rows of three plants to different fertilization treatments at random such that one row receives fertilizer A, one receives fertilizer B, and one receives no fertilizer. After a period of time she measures the total volume of flowers produced by each plant in her study; she records in the end thirty-six response values.

i. Give the name of the experimental design.




This is a randomized complete block split plot design: The blocks are front yard / back yard, the whole plot EUs are the beds, and the split plot EUs are the rows.

ii. Select the appropriate model from models 1–6 and describe in detail the role of each term in the model.

(1) $Y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + C_k + (\tau C)_{ik} + \varepsilon_{ijk}$

$\mu$ = overall/baseline mean

$\tau_i$ = effect of watering / not watering

$\beta_j$ = effect of fertilizer treatment

$(\tau\beta)_{ij}$ = interaction between watering / fertilizer

$C_k$ = block effect (front/back yard)

$(\tau C)_{ik}$ = interaction between watering and front/back yard

$\varepsilon_{ijk}$ = an error term.

iii. State the null and alternate hypotheses corresponding to the question of whether the position of a bed in the front yard versus the back yard plays any role in the total volume of flowers produced by the hellebores plants.

$H_0: \sigma_C^2 = 0$ vs $H_1: \sigma_C^2 > 0$.

**d)**

Another friend of yours, not equipped with a greenhouse or with any raised beds, but who has already several hellebores plants growing in his yard, wishes also to do a study. He decides to assign each of the 20 hellebores plants in his yard at random to fertilizer treatments such that seven receive fertilizer A, seven receive fertilizer B, and six receive no fertilizer. He notices that not all the plants receive the same amount of sunlight, and decides to record, for each plant, the average number of hours per day of sunlight it receives during the course of the study. He records along with these values the total volume of flowers produced by each plant.

i. Give the name of the experimental design.

This is an analysis of covariance design.

ii. Select the appropriate model from models 1–6 and describe in detail the role of each term in the model.

(5) $Y_{ij} = \mu + \tau_i + \beta x_{ij} + \varepsilon_{ij}$

$\mu$ = overall/baseline mean
$\tau_i$ = effect of fertilizer treatment
$\beta$ = effect of hrs of sunlight
$x_{ij}$ = sunlight for each plant
$\varepsilon_{ij}$ = an error term.

iii. What is the purpose of recording the sunlight information for each plant? How will this be used in the analysis?

The EUs (the plants) are not homogeneous, which introduces additional variability into the study. Incorporating the sunlight information will allow us to capture this variability. Moreover, it will allow us to compute means for each group which are adjusted for the differences in sunlight among the EUs.

## 2. Logistic regression

Organizers of next year's Save The Plankton 50k Ultramarathon (STP50k) wish to predict each registrant's probability of completing the event. For a random sample of 100 participants in the most recent STP50k, the organizers recorded whether or not the participant finished as well as the number of long-distance running events each of these participants had completed prior to their participation in the STP50k.

The data are summarized in the table below, where `y` is `1` if the participant completed the event and `0` otherwise and `x` is the number of long-distance running events completed by the participant prior to their participation in the STP50k.

```
table(y,x)
```

```
   x
y    2  3  4  5  6  7  8  9 10 11 12 13 14 15 16
  0  1  1  1  5  2 11  7  5  7  6  2  0  0  1  0
  1  0  0  0  0  0  1  3  6  3  8 12  9  5  3  1
```

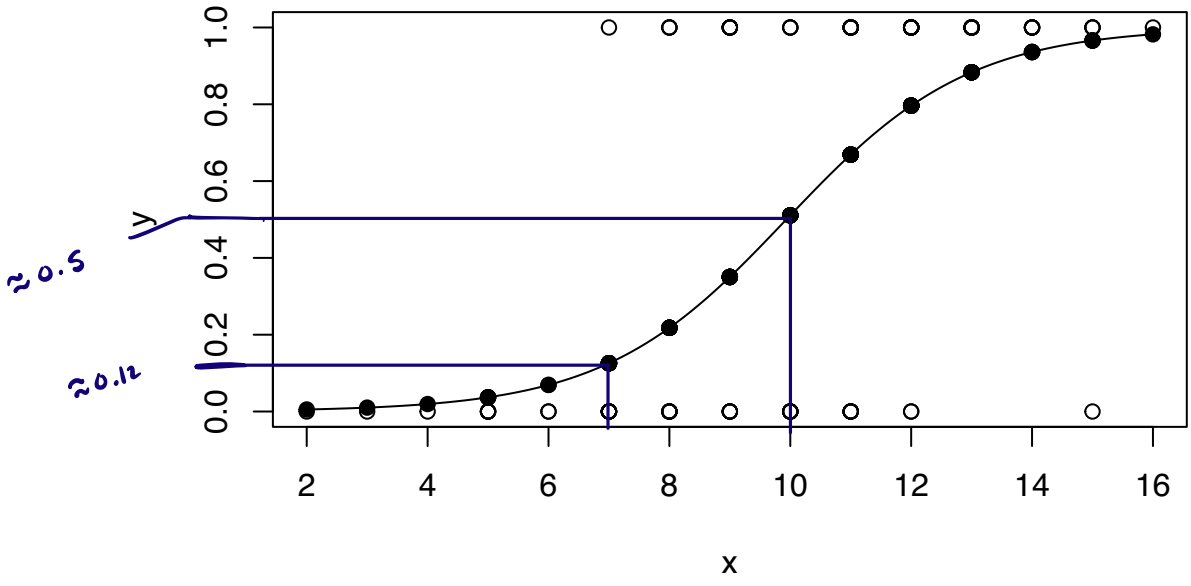
Here is some additional R output:

```
glm_out <- glm(y~x,family="binomial")
summary(glm_out)
```

```
Call:
glm(formula = y ~ x, family = "binomial")

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.5619     1.3410  -4.893 9.91e-07 ***
x             0.6605     0.1308   5.048 4.47e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 138.589  on 99  degrees of freedom
Residual deviance:  92.785  on 98  degrees of freedom
AIC: 96.785

Number of Fisher Scoring iterations: 5
```

```
xseq <- seq(min(x),max(x),length = 200)
b0_hat <- coef(glm_out)[1]
b1_hat <- coef(glm_out)[2]
pi_xseq <- 1/(1 + exp( - (b0_hat + b1_hat * xseq)))
pi_hat <- 1/(1 + exp( - (b0_hat + b1_hat * x)))
```

```
plot(y~x)
lines(pi_xseq ~ xseq)
points(pi_hat ~ x,pch = 19)
```




**a)**

Does it look like the number of long-distance running events completed prior to the STP50k is a significant predictor of whether a participant will complete the STP50k? Justify your answer.

Since the p-value is $4.47 \times 10^{-7}$, we conclude that it is a significant predictor.

**b)**

Give an expression for the estimated probability that a participant who has in the past completed 10 long-distance running events will complete the STP50k (you do not have to evaluate your expression). In addition, use the plot to provide an approximate answer.

For $x = 10$, we have

$$\hat{\pi} = \frac{e^{-6.56 + 0.66 * 10}}{1 + e^{-6.56 + 0.66 * 10}}.$$



From the plot, this looks to be approximately 0.5

**c)**

Use the plot to give the approximate odds of a participant completing the STP50k who has completed 7 long-distance running events in the past.

For $x = 7$, we have $\hat{\pi} \approx 0.12$. So odds $\approx \frac{0.12}{1-0.12}$.

**d)**

Give the estimate of the factor by which the odds of completing the STP50k increase with each additional long-distance running event completed in the past.

This is the odds ratio, which is given by $e^{\hat{\beta}_1} = e^{0.66}$.

**e)**

Give careful interpretations of these two confidence intervals. Write a couple of sentences.

What is the relationship between completion of long-distance running events in the past and completing the STP50k?

```
confint.default(glm_out, parm = "x")
```

```
      2.5 %    97.5 %
x 0.4040312 0.9169441
```

```
exp(confint.default(glm_out, parm = "x"))
```

```
     2.5 %   97.5 %
x 1.497851 2.501634
```

The first is the C.I. for $\beta_1$, which is the change in the log-odds due to a unit increase in $x$.

The second is the C.I. for $e^{\beta_1}$, which is the factor by which the odds of completing the STP50k increase with each additional completed long-distance running event.

Completion of more long-distance running events increases the probability that a participant will finish the STP50k.

## 3. Analysis of covariance

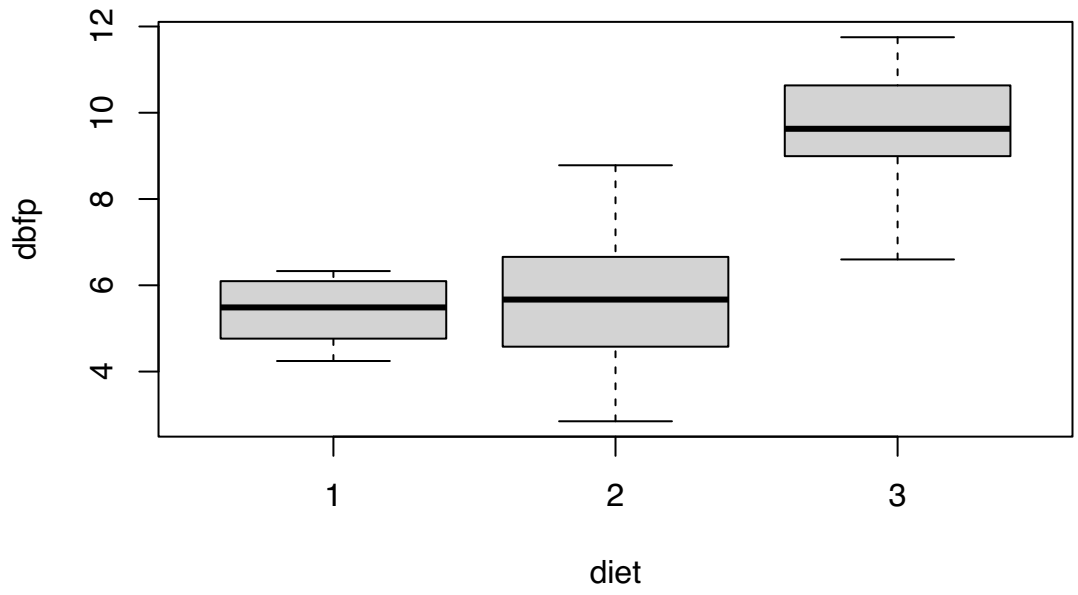
A number of mice are assigned randomly to three different diets (`diet`). After a period of time on the diet, the change in the body fat percentage (`dbfp`) of each mouse is recorded. In addition, the weight (`wt`) of each mouse is recorded at the start of the experiment. It is of interest to see how the diet effects the body fat percentage of mice.

Some R output follows; note that `lm()` is run with three different model specifications.

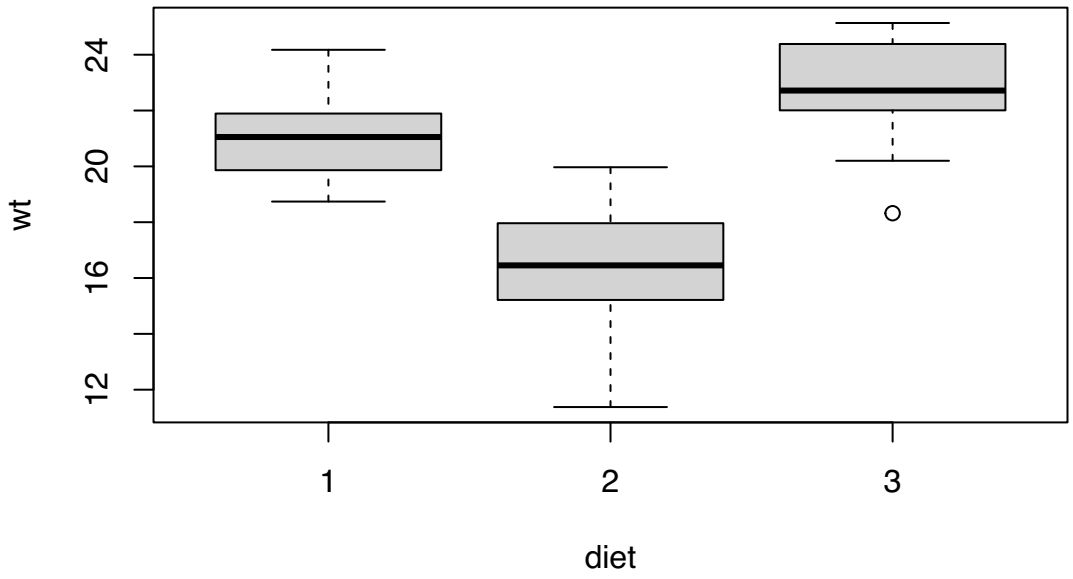
```
head(mice)
```

```
      dbfp diet       wt
1 4.244676    1 19.20617
2 6.066143    1 21.36970
3 6.124775    1 24.17569
4 5.486788    1 18.73925
5 5.486860    1 20.83950
6 6.329132    1 21.26484
```

```
boxplot(dbfp ~ diet, data = mice)
```




```
boxplot(wt ~ diet, data = mice)
```




```
library(car)
```

```
Loading required package: carData
```

```
lm_out1 <- lm(dbfp ~ diet + wt + diet:wt, data = mice)
summary(lm_out1)
```

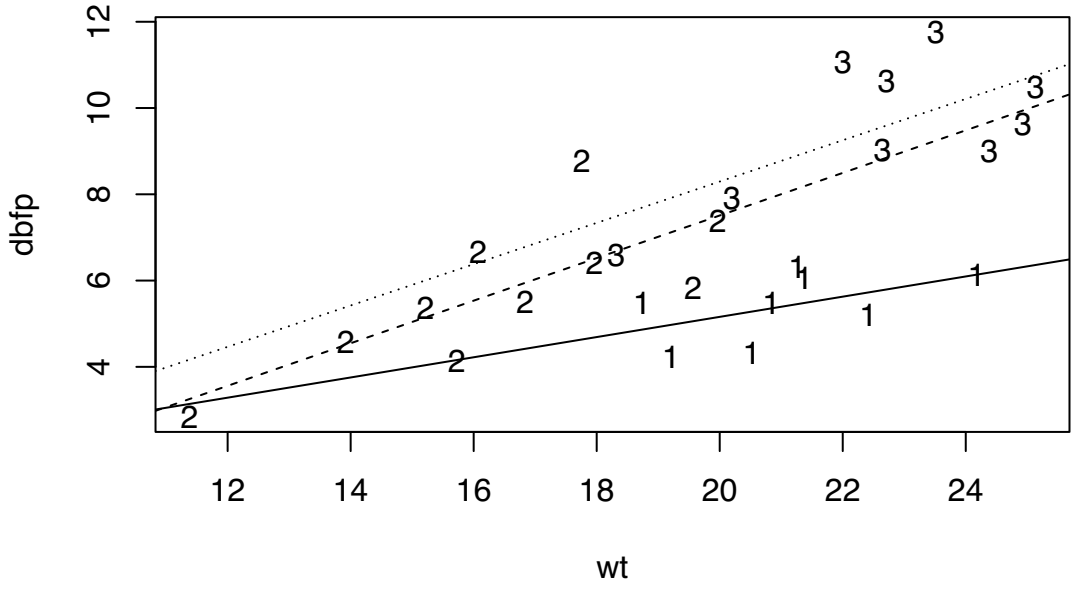
```
Call:
lm(formula = dbfp ~ diet + wt + diet:wt, data = mice)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4703 -0.6363 -0.1043  0.6054  2.3808

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.4748     5.1664   0.092    0.928
diet2        -2.8384     5.6902  -0.499    0.623
diet3        -1.7576     6.5345  -0.269    0.791
wt            0.2342     0.2445   0.958    0.349
diet2:wt      0.2594     0.2835   0.915    0.371
diet3:wt      0.2447     0.3012   0.812    0.426
```

```
Residual standard error: 1.116 on 21 degrees of freedom
Multiple R-squared:  0.821, Adjusted R-squared:  0.7784
F-statistic: 19.27 on 5 and 21 DF,  p-value: 3.309e-07
```

```r
Anova(lm_out1,type = "III")
```

```
Anova Table (Type III tests)

Response: dbfp
             Sum Sq Df F value Pr(>F)
(Intercept)  0.0105  1  0.0084 0.9276
diet         0.3305  2  0.1326 0.8765
wt           1.1432  1  0.9172 0.3491
diet:wt      1.1158  2  0.4476 0.6451
Residuals   26.1745 21
```

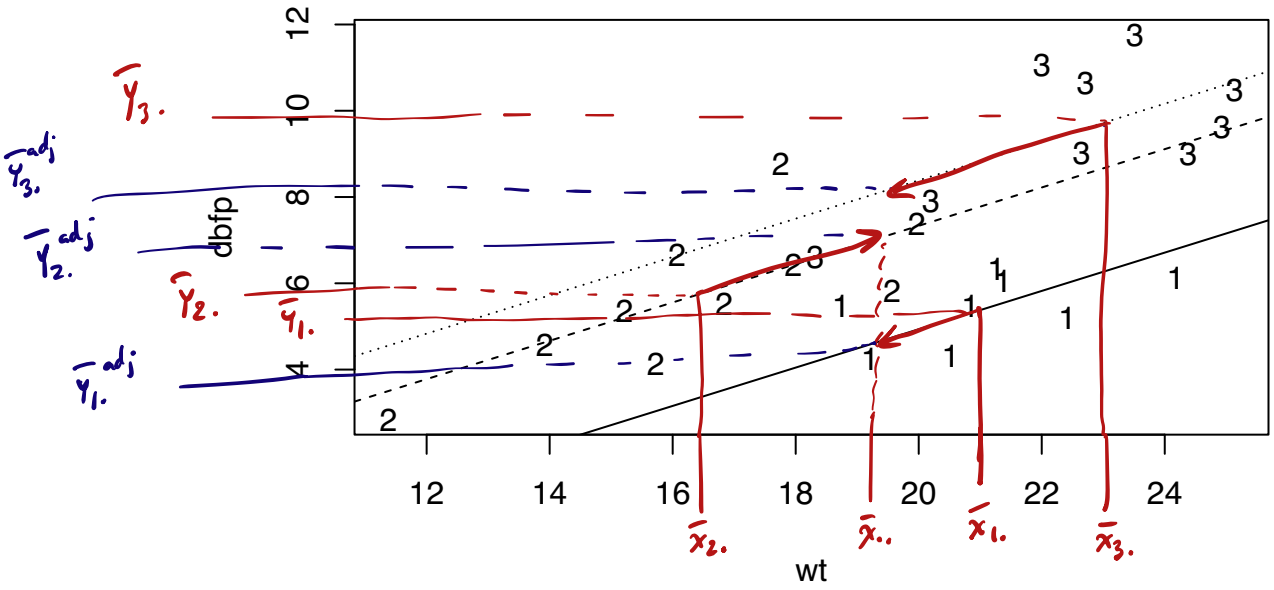
```r
plot(dbfp ~ wt, pch = as.character(diet), data = mice)
parms1 <- coef(lm_out1)
abline(parms1[1],parms1[4])
abline(parms1[1] + parms1[2],parms1[4] + parms1[5], lty = 2)
abline(parms1[1] + parms1[3],parms1[4] + parms1[6], lty = 3)
```



```
lm_out2 <- lm(dbfp ~ diet + wt, data = mice)
Anova(lm_out2, type = "III")
```

```
Anova Table (Type III tests)

Response: dbfp
            Sum Sq Df F value    Pr(>F)
(Intercept)  4.140  1  3.4893  0.074559 .
diet        52.500  2 22.1233 4.382e-06 ***
wt          24.023  1 20.2465  0.000162 ***
Residuals   27.290 23
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(dbfp ~ wt, pch = as.character(diet), data = mice)
parms2 <- coef(lm_out2)
abline(parms2[1],parms2[4])
abline(parms2[1] + parms2[2],parms2[4], lty = 2)
abline(parms2[1] + parms2[3],parms2[4], lty = 3)
```




```
lm_out3 <- lm(dbfp ~ diet, data = mice)
anova(lm_out3)
```

```
Analysis of Variance Table

Response: dbfp
          Df Sum Sq Mean Sq F value    Pr(>F)
diet       2 94.922  47.461  22.198 3.484e-06 ***
Residuals 24 51.313   2.138
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(aov(dbfp ~ diet, data = mice))
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = dbfp ~ diet, data = mice)

$diet
        diff       lwr      upr     p adj
2-1 0.342909 -1.389178 2.074996 0.8746452
3-1 4.157028  2.382689 5.931368 0.0000143
3-2 3.814119  2.136343 5.491896 0.0000220
```

**a)**

What is the point of looking at the boxplots of the weights across the diet groups?

This is to see if the weights of the mice differ across the treatment groups. If the weight of a mouse affects the change in body fat percentage, then differences in weights across treatment groups may affect our perception of the treatment effects.

**b)**

Should we conclude that the effect of the mouse weight on the change in body fat percentage is different across the three treatment groups? Explain why or why not.

Since the interaction term diet x weight is not significant (p-value = 0.6451), there is insufficient evidence to conclude that the weight of a mouse has an effect depending on the diet. That is, there is insufficient evidence to conclude that the slope coefficients are different across the treatment groups.

**c)**

Suppose we did *not* take the weights of the mice into account. What would we conclude about the three diets? Look closely into the R output.

We would reject the null hypothesis of no difference in means with p-value $3.484 \times 10^{-6}$. In addition, Tukey's pairwise comparison of means would yield the conclusion that diets 1 and 2 are not different, but that diet 3 lead to a higher change in body fat % - age than diets 1 & 2.

**d)**

The mean change in body fat percentage among all the mice on diet 1 was 5.408 and that among all the mice on diet 3 was 9.565. The difference between these means is 4.157. If we adjusted these means by taking into account the weights of the mice, would the difference between the adjusted means be greater than 4.157 or less than 4.157? How can you tell? Explain your answer in detail. Recall that the treatment group means are given by $\bar{Y}_{i\cdot} = \hat{\mu} + \hat{\tau}_i + \hat{\beta}\bar{x}_{i\cdot}$, while the adjusted treatment group means are given by $\bar{Y}_{i\cdot}^{\text{adj}} = \hat{\mu} + \hat{\tau}_i + \hat{\beta}\bar{x}_{\cdot\cdot}$.

The difference between the adjusted group means would be smaller than 4.157. This can be seen in the figure: the adjusted mean for each group is the height of the line at $x = \bar{x}_{\cdot\cdot}$, the mean of all the covariate values. The unadjusted means are the heights of the lines at $\bar{x}_{1\cdot}$ and $\bar{x}_{3\cdot}$. Both are adjusted down, but $\bar{Y}_{3\cdot}$ is adjusted further down.

**e)**

Which diet will have the lowest adjusted mean? Does your answer contradict the side-by-side boxplots of the change in body fat percentages across the diets? Explain your answer.

Diet 1 will have the lowest adjusted mean, since the line for diet 1 is the lowest.

We could not have concluded this from the boxplots alone, because the boxplots, which ignore the weights, make it appear as though Diet 1 and Diet 2 have might have an equal effect on the change in body fat percentage.

## 4. One-way ANOVA

Assume $Y_{ij} = \mu_i + \varepsilon_{ij}$, $i = 1, \ldots, a$ and $j = 1, \ldots, n$ with $\varepsilon_{ij} \overset{\text{ind}}{\sim} \text{Normal}(0, \sigma^2_\varepsilon)$ and consider the three quantities

1. $n \sum_{i=1}^{a} (\bar{Y}_{i.} - \bar{Y}_{..})^2$
2. $\sum_{i=1}^{a} \sum_{j=1}^{n} (Y_{ij} - \bar{Y}_{i.})^2$
3. $\sum_{i=1}^{a} \sum_{j=1}^{n} (Y_{ij} - \bar{Y}_{..})^2$

**a)**

Give the name of and the degrees of freedom associated with each of the above quantities.

(1) Treatment sum of squares ($SS_{Treatment}$): $a-1$

(2) Error sum of squares ($SS_{Error}$): $a(n-1)$

(3) Total sum of squares ($SS_{Total}$): $an-1$

**b)**

State which of the above quantities describes between-treatment variation (in the response values) and which describes within-treatment variation.

$SS_{Treatment}$ describes between-treatment variation;

$SS_{Error}$ describes within-treatment variation.

**d)**

Show how we may construct from the above quantities a test statistic for testing $H_0$: $\mu_1 = \cdots = \mu_a$ which has an F distribution when $H_0$ is true.

Define

$$F_{stat} = \frac{SS_{Treatment} / (a-1)}{SS_{Error} / [a(n-1)]}.$$



Reject $H_0: \mu_1 = \cdots = \mu_a$ if $F_{stat} > F_{a-1, n(a-1), \alpha}$.