Lab 7: Linear Models with Nominal Explanatory Variables

ENVIRON 710: Applied Statistical Modeling*

In this lab, we explore linear models with nominal explanatory variables, traditionally termed ANOVA. We start with a single nominal variable and then examine more complicated experimental designs, such as factorial, block, and repeated measures designs. There are many others, but this should get you comfortable with the basics so that you can learn others by yourself as needed.

We could analyze nominal variables to ask whether the quantity of fertilizer (none, low, medium and high) results in significantly different levels of plant growth. Note that our treatment, fertilizer, is a *factor*, with four levels (nominal values). The dependent variable (the variable that depends on the level of fertilization) is plant growth. The null hypothesis is that the mean plant growth is the same for each fertilizer level versus the alternative hypothesis that at least one of the means is different from one of the others.

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$
$$H_a : \text{Not } \mu_1 = \mu_2 = \mu_3 = \mu_4$$

Of course, the experiment should be replicated and randomized. So if plants were being grown in pots in a nursery, for example, we would apply each level of fertilization to 10 pots (N = 40 with 10 pots per treatment). The treatment applied to each pot should be randomly chosen to avoid confounding effects. For example, pots undergoing the same treatment should not be located together in the greenhouse and should not contain potting soil from the same bag, etc. For ANOVA models, we make a few assumptions:

1. the dependent variable should be normally or near-to-normally distributed for each level of the treatment (e.g. fertilization);
2. variances of each level of treatment are homogeneous; that is, each treatment level has approximately the same variance;
3. all observations are independent of each other, within and between groups;

Violations to the first two assumptions that are not extreme can be considered not serious. The sampling distribution of the test statistic is fairly robust, especially as sample size increases and more so if the sample sizes for all factor levels are equal. Try to keep the same sample sizes for each factor level.

A general rule of thumb for equal variances is to compare the smallest and largest sample standard deviations. If the ratio of these two sample standard deviations falls within 0.5 to 2, the assumption may not be violated.

The goals of the lab are to:

- Practice linear models with nominal variables in R
- Learn graphing methods to depict the results of ANOVA models
- Recognize different study designs, when and how to use them, and understand how to adapt models for each design
- Conduct *post-hoc* tests to evaluate statistical differences among levels of factors
- Practice correctly writing the results of your analyses.

Work through the lab, running the example code by typing it into R Studio (do not copy and paste from the pdf) - make sure you know what every line does. At the end of the lab, answer the problems using R

*Created by John Poulsen with edits from TAs.

Markdown to show your code, results and any requested graphs. *Submit your answers in R Markdown to the class Sakai site under the Assignments folder.*

More functions in R

In this lab, we introduce a few new R commands.

- `aov()` - fits an ANOVA model by a call to `lm` for each stratum
- `par()` - sets global graphics parameters
- `levels()` - displays the unique levels of a categorical (factor) variable
- `jitter()` - adds a small amount of noise to a numeric vector; good for graphing and offsetting potentially overlapping points
- `tapply()` - stands for *table apply*, and applies a function (3rd argument) to a variable (1st argument) separately for each group specified by the second argument
- `with()` - tells a function the data to use, possibly modifying the original data; avoids the need to use `attach()`
- `interaction.plot()` - plots the mean of the response for the two-way combinations of factors, thereby illustrating possible interactions
- `points()` - draws a sequence of points at user-specified coordinates
- `*` - symbol used to define a full first order model (all main effects and interactions): $y \sim \beta_0 + \beta_1 A + \beta_2 B + \beta_3 AB$
- `:` - symbol used to define an interaction in a model
- `mtext()` - writes text in one of the margins of a plot, where `side` specifies which side of the figure to write on and `line` specifies on which margin line to write
- `text()` - writes text within the plotting area of a figure, using x, y coordinates to specify the position
- `axis()` - adds an axis to the plot, allowing for custom axes (usually the axis on the original plot is first suppressed)
- `residuals()` - returns the residuals from a linear model.

Exploring the data

We are going to use the Africa plot data to test whether the three forest types (logged & hunted forest, logged only forest, pristine forest) differ in mean levels of aboveground biomass. What is the null hypothesis? What is the alternative hypothesis? This "natural experiment" is replicated in the sense that there are multiple plots in each forest type. The experiment is randomized in the sense that plot locations were chosen randomly within each forest type. However, the experiment is pseudoreplicated because the plots of each forest type are grouped spatially (i.e., treatments were not replicated). This was unavoidable as it was a constraint of the physical environment of the study site, but it does raise questions about the inferences that can be made from the study.

To keep things simple, we will analyze these data for the first census of the plots.

```
adat <- read.csv("Afrplots.csv", header = T)
adat$Site <- as.factor(rep(c(rep(1, 10), rep(2, 10), rep(3, 10)),
    2))
bdat <- adat[adat$CensusNo == 1, ]
bdat <- bdat[, !(names(bdat) %in% c("PlotCode", "CensusNo", "MeanGr",
    "Trees", "Dead", "Recruits"))]
```

To make sure `Site` is correctly classed as a factor, we evalute the levels of the factor.

```
levels(bdat$Site)
```

In the above code, we downloaded the data and added a column, attributing each forest type a code: 1 =

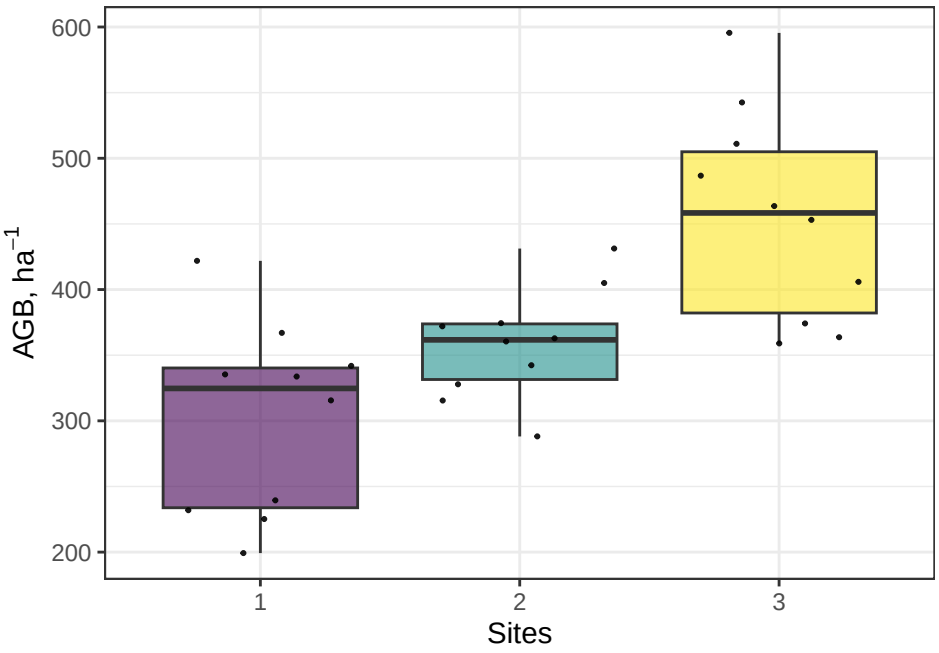
logged & hunted forest; 2 = logged only forest; 3 = pristine forest. Then, we created a new database `bdat` that only includes the first census.

We could explore the data with `ggpairs` (not shown).

```
ggpairs(bdat)
```

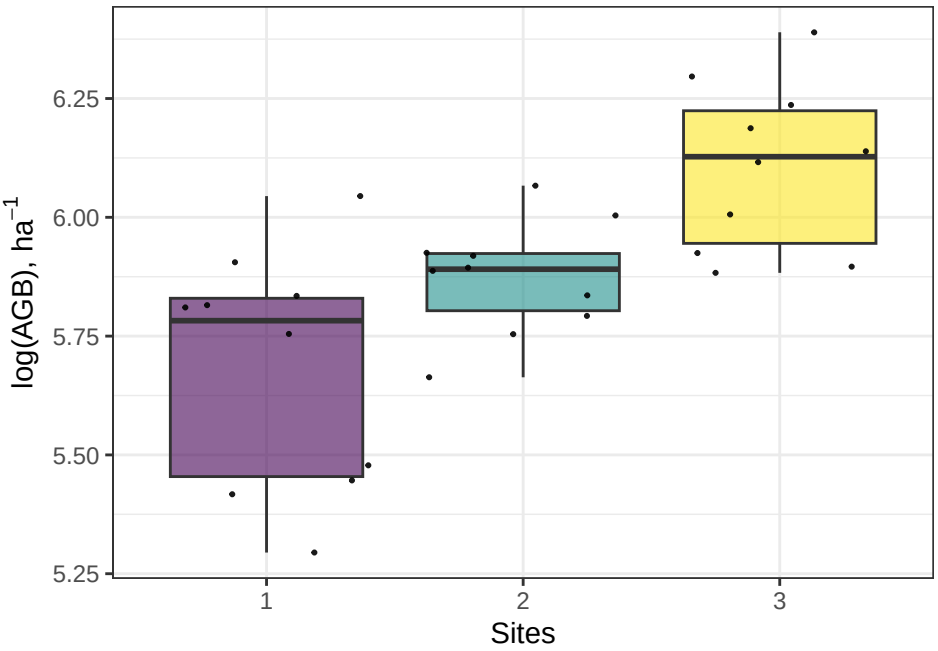
Or we could specifically try something different by plotting data points over boxplots of each level of `Site`. Loading the `viridis` package provides a different color scale for the plot.

```
require(viridis)
ggplot(bdat, aes(x = Site, y = ChaveMoist, fill = Site)) + geom_boxplot() +
    scale_fill_viridis(discrete = TRUE, alpha = 0.6) + geom_jitter(color = "black",
    size = 0.4, alpha = 0.9) + xlab("Sites") + ylab(expression(paste("AGB, ",
    ha^-1))) + theme_bw() + theme(legend.position = "none")
```



There could be some deviation from our assumption of normally distributed data for each group or level of site. Log transform the data and see if the boxplots look better.

```
ggplot(bdat, aes(x = Site, y = log(ChaveMoist), fill = Site)) +
    geom_boxplot() + scale_fill_viridis(discrete = TRUE, alpha = 0.6) +
    geom_jitter(color = "black", size = 0.4, alpha = 0.9) + xlab("Sites") +
    ylab(expression(paste("log(AGB), ", ha^-1))) + theme_bw() +
    theme(legend.position = "none")
```



A log-transformation does not seem to center the median, although it might reduce the skew a bit in the tails. What happens if we try qq plots? Here we plot a qqplot for the first site, logged & hunted forest.

```
ggplot(bdat[bdat$Site == 1, ], aes(sample = ChaveMoist)) + stat_qq() +
    stat_qq_line() + labs(y = "Sample quantiles", x = "Theoretical quantiles") +
    theme_bw()
```

To Do

Run qq plots for the other sites for the raw data and log-transformed data.

Before building a model, let's test the assumption of homogeneity of variances by evaluating the ratios of the sample standard deviations.

```
with(bdat, sd(ChaveMoist[Site == 1])/sd(ChaveMoist[Site == 2]))
with(bdat, sd(ChaveMoist[Site == 2])/sd(ChaveMoist[Site == 3]))
with(bdat, sd(ChaveMoist[Site == 1])/sd(ChaveMoist[Site == 3]))
```

The ratios range from 0.52 to 1.73. This is less than our criterion of the largest standard deviation being two times bigger than the smallest.

Conducting a linear model with one nominal variable

Now we will conduct the one-way ANOVA. Recall that we are testing for differences among means of the levels of the factor, *Site*. In other words, do logged & hunted, logged only, and pristine forests all have the same mean biomass?

```
mod1 <- lm(ChaveMoist ~ factor(Site), data = bdat)
summary.aov(mod1)
summary(mod1)
```

Note the syntax used to define the model: dependent variable ~ independent variable. Here, the `summary.aov()` function produces the traditional ANOVA table, providing the sum-of-squares (`Sum Sq`), mean squares (`Mean Sq`) and F-statistic (`F value`). The summary of the results of `mod1` provides a lot more information, which we will discuss below, but notice that the last line of the results gives the same F-statistic and p-value as the ANOVA table.

Interpreting results of our test

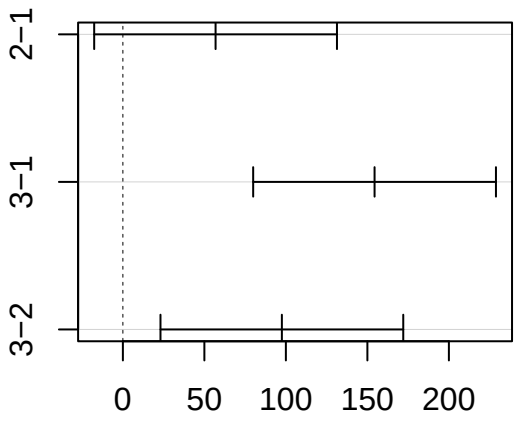
Study the results from `summary()` and note the large F-statistic and very small p-value. These tell us that we should reject the null hypothesis that the three forest types have the same mean biomass values in *favor of the alternative hypothesis that there is a difference in biomass between at least two of the sites*. We do not yet know which forest types are significantly different from each other.

To determine which of the means are significantly different from one another, let's conduct a *post-hoc* test – Tukey's Honest Significant Difference.

```
TukeyHSD(mod1)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = x)
##
## $`factor(Site)`
##          diff       lwr      upr     p adj
## 2-1  56.86767 -17.60653 131.3419 0.1599999
## 3-1 154.40412  79.92992 228.8783 0.0000605
## 3-2  97.53645  23.06225 172.0107 0.0084513
```

```
plot(TukeyHSD(mod1))
```



The output shows the difference, `diff`, in mean biomass in pairwise comparisons of the sites, confidence intervals, `lwr` and `upr`, of the difference, and the probability, `p adj` of the sites having the same mean biomass. The plot presents the 95% CI's for the differences between the pairwise comparisons.

In this example, there is not a statistically significant difference between sites 2 and 1 (logged only and hunted & logged forest) as illustrated by the fact that the CI overlaps 0 and the p-value is greater than 0.05. By contrast, there appears to be significant differences in biomass between sites 1 and 3 and sites 2 and 3.

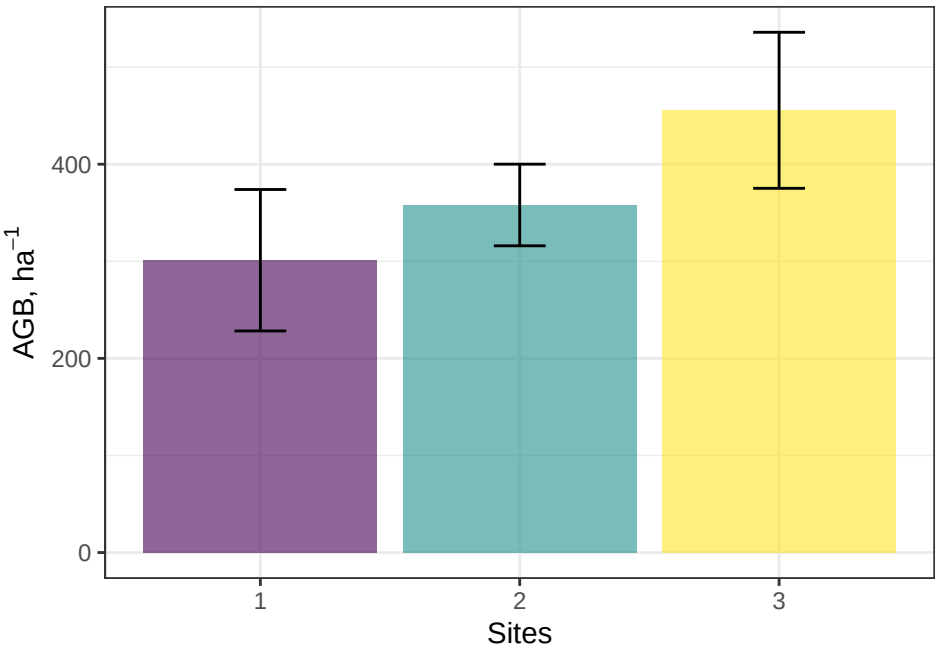
Graphing the results

As you have seen above, a boxplot is a good way to demonstrate results. Barplots are another commonly used method for visualizing results. R, inconveniently, does not have an automatic function to do error bars. Therefore, we need to write a function to summarize the data and then graph it.The function is below. Run it to produce a barplot of the biomass data per forest type.

```
data_summary <- function(data, varname, groupnames){
  require(plyr)
  summary_func <- function(x, col){
    c(mean = mean(x[[col]], na.rm=TRUE),
      sd = sd(x[[col]], na.rm=TRUE))
  }
  data_sum<-ddply(data, groupnames, .fun=summary_func,
                  varname)
  data_sum <- rename(data_sum, c("mean" = varname))
 return(data_sum)
}

smry <- data_summary(data = bdat, varname="ChaveMoist",
                     groupnames=c("Site"))

ggplot(smry, aes(x=Site, y=ChaveMoist, fill=Site)) +
    geom_bar(position=position_dodge(), stat="identity") +
    scale_fill_viridis(discrete = TRUE, alpha=0.6) +
    labs(x = "Sites", y = expression(paste("AGB, ", ha^-1))) +
    geom_errorbar(aes(ymin=ChaveMoist-sd, ymax=ChaveMoist+sd),
                  width=.2,                    # Width of the error bars
                  position=position_dodge(.9)) +
    theme_bw() + theme(legend.position = "none")
```



There is a growing concensus that boxplots or plots of the group means and their 95% confidence intervals represent data better than barplots.

Now let's move onto to different experimental designs for nominal variables.

Factorial Design

A factorial design has two or more factors, each with two or more levels. The test subjects for a factorial design are assigned to treatment levels of every factor combination at random. This means that we can investigate statistical interactions, in which the response to one factor depends on the level of another factor. Let's use an example of animal diets from Crawley (2005). In this dataset (download `Growth.csv`), the response variable is weight gain of domestic animals after 6 weeks. There are two factors, diet and supplement, and each level of diet is crossed with each level of supplement making it a factorial design.

```
feed <- read.csv("Growth.csv", header = T)
```

How many levels are there of each factor? Use `levels()` to check it out. Before we get started, also check to make sure diet and supplement are defined as factors.

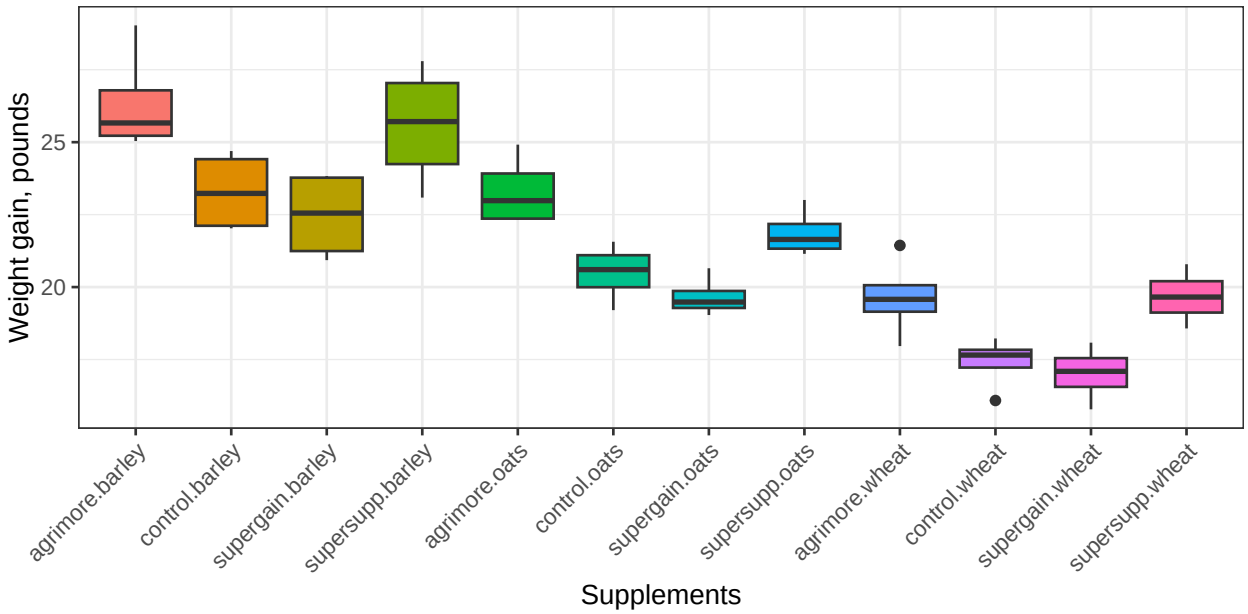
```
feed <- read.csv("Growth.csv", header = T, stringsAsFactors = T)
```

Getting a feel for the data

We first want to get a picture of the data. We will again use `ggplot` but because we want to graph two nominal variables we need to use the `interaction` function to compute a factor which represents the interaction of both factors. Note also the use of the `axis.text.x` argument to adjust the angle of the x-axis tick labels.

```
feed$suppdiet <- interaction(feed$supplement, feed$diet)

ggplot(data = feed, aes(y = gain, x = suppdiet)) + geom_boxplot(aes(fill = suppdiet)) +
    labs(y = "Weight gain, pounds", x = "Supplements") + theme_bw() +
    theme(legend.position = "none", axis.text.x = element_text(angle = 45,
        vjust = 1, hjust = 1))
```



Running the model

There are a couple ways to run a factorial ANOVA in R - both lines here do the same thing. In the first line we specify that we want to estimate the two main effects, diet and supplement, and the interaction (designated by :) between the variables. The * in the second line is a shortcut for the same model

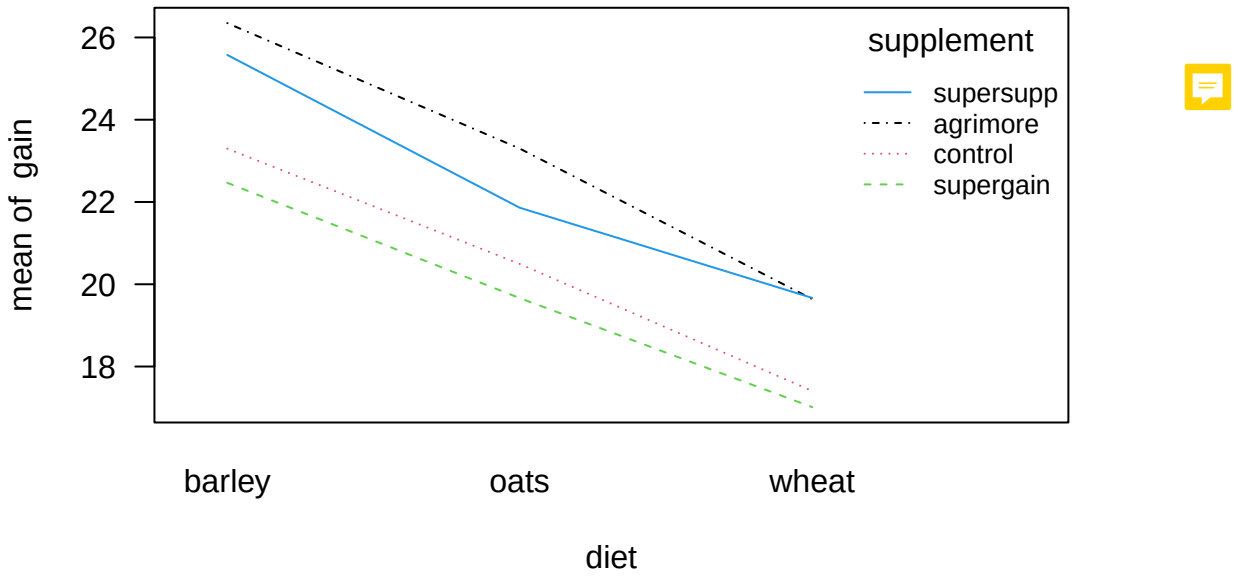
```
mod.fac <- lm(gain ~ diet + supplement + diet:supplement, data = feed)
mod.fac <- lm(gain ~ diet * supplement, data = feed)
```

Use `summary.aov()` to examine the ANOVA table and `summary()` to view all the effects.

The results provide no support of a significant interaction between diet and supplement. We can use an interaction plot to get a better idea of what is going on. An interaction plot displays the dependent variable on the y-axis, the levels of one factor on the x-axis, and separate lines for the means of each level of the second variable.

These types of plots can be used to determine whether an interaction term should be included in our model. The `interaction.plot` function is a bit annoying because it isn't very flexible (e.g., there is no way to change the position of the legend without hacking the code).

```
with(feed, interaction.plot(diet, supplement, gain, col = c(1,
    2, 3, 4), las = 1, cex = 0.9))
```



The consistent decline in weight gain from barley to oats to wheat demonstrates the significance of the main effect of diet. Similarly, the separation of the lines for supplement also suggests a difference along that main effect. The lines are parallel to each other, indicating no interaction.

Refining the model

In many cases, the goal of analysis is to look for the simplest (most parsimonious) model for our data. Therefore, we can refine the model taking out non-important - here interpreted as non-significant - variables. Let's simplify our ANOVA model by taking out the interaction term.

```
mod.fac <- lm(gain ~ diet + supplement, data = feed)
summary(mod.fac)
```

Like the previous model, the ANOVA results tell us that diet and supplement are statistically significant main effects, but where do the differences in weight gain lie?

```
require(graphics)
TukeyHSD(aov(mod.fac), "diet", ordered = TRUE)
plot(TukeyHSD(aov(mod.fac), "diet"))
```

The Tukey test tells us that there are significant differences between the means of each of the diets. In other words, the mean weight gain is greatest on barley, then oats, and then wheat, with significant differences in mean weight gain between each pair of diets at the 0.05 level.

To Do

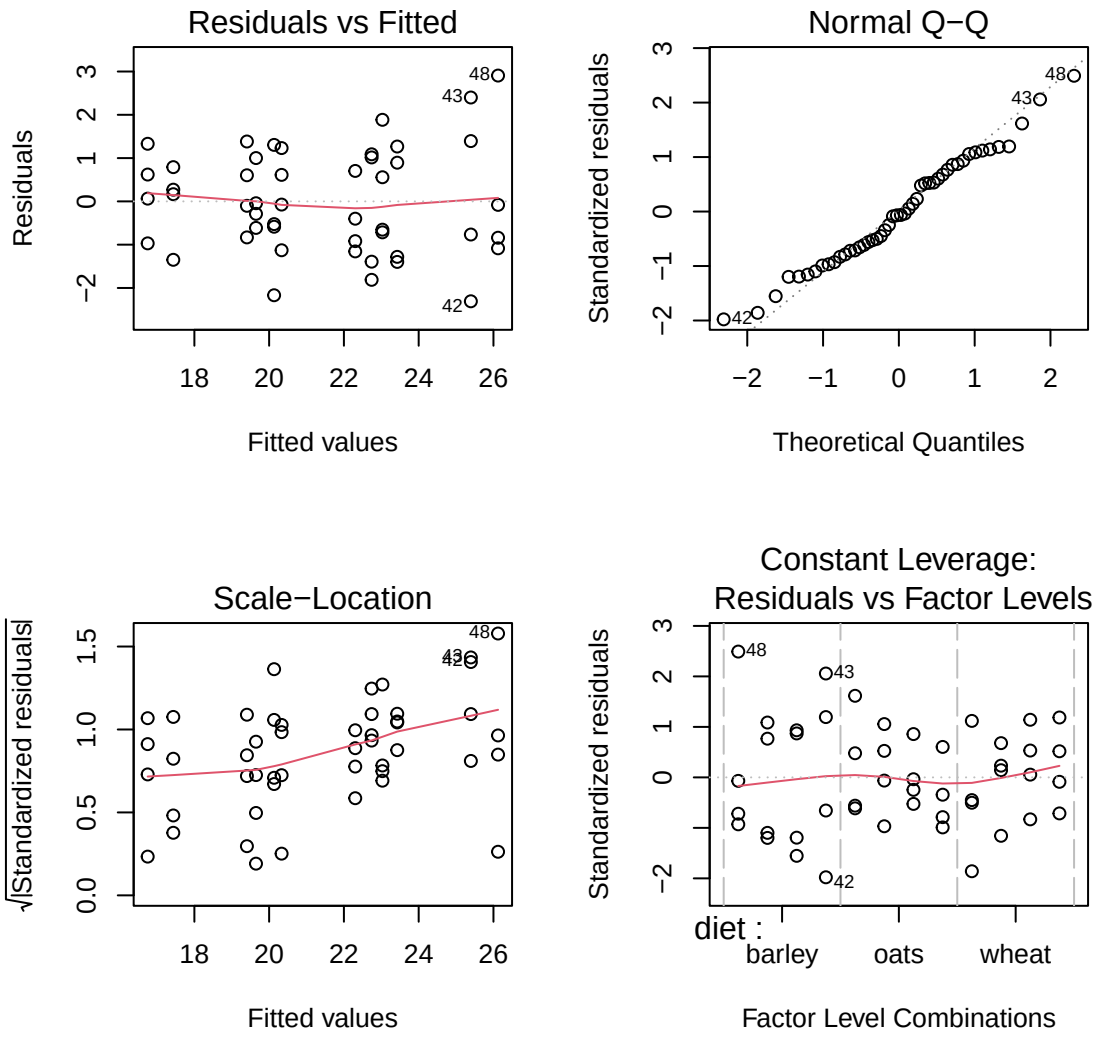
Modify the above code to examine where the *post-hoc* differences lie for different levels of supplement.

Checking ANOVA assumptions

When conducting any statistical analysis it is important to evaluate how well the model fits the data and that the data meet the assumptions of the model. There are numerous ways to do this and a variety of

statistical tests to evaluate deviations from model assumptions. Generally statisticians examine multiple diagnostic plots after running regression models.

```
par(mfrow = c(2, 2))
plot(mod.fac)
```



Plotting of the model provides four diagnostic plots.

Residual plot

The first plot is a plot of the residuals (distance of the data points from the expected values (treatment means)) versus the fitted data. Residuals can be thought of as elements of variation unexplained by the fitted model. We expect them to be (roughly) normally and (approximately) independently distributed with a mean of 0 and constant variance. In the plot, points should be randomly scattered around the centerline. Any pattern indicates a violation of linearity. Note that because we are dealing with categorical independent variables the residuals form swarms around their treatment means.

```
mod.fac1 <- lm(log(gain) ~ diet + supplement, data = feed)
summary(mod.fac1)
```

QQ plot

The second plot is a qq plot, which we have already used to evaluate the normality of data. Significant departures from the line suggest violations of normality. If the pattern were S-shaped or banana shaped, we would need a different model.

Or, we could look at the distribution of residuals.

```
hist(residuals(mod.fac))
```

Standardized residual plot

The third plot is a plot of standardized residuals versus the fitted values. It repeats the first plot, but on a different scale. It shows the square root of the standardized residuals (where all the residuals are positive). This plot shows if residuals are spread equally along the ranges of predictors. This is how you can check the assumption of equal variance (homoscedasticity). It's good if you see a horizontal line with equally (randomly) spread points. In this plot there might be some evidence of heteroscedasticity because the variance seems to increase with the treatment means. Remedies to this would be to square root or log transform the dependent variable or add a higher order term to the model (e.g. an interaction).

To Do

Log-transform the dependent variable, re-run the model, and evaluate the residual plot. Does the log-transformation improve the residual plots?

Leverage plot

The fourth plot is a residuals-leverage plot that shows Cook's distance for each of the observed values of the factor levels. Cook's distance measures relative change in the coefficients as each replicate is deleted. So the point is to highlight those y (response) values that have the biggest effect on parameter estimates. The idea is to verify that no single data point is so influential that leaving it out changes the structure of the model. In this plot, there do not seem to be any outliers, although observation 48 has more leverage than other observations.

Block Design

The basic idea behind blocking is to group the experimental units into blocks of similar units and carry out the assignment of treatments separately within each block. With every treatment included at least once in every block, the design is called a complete block design. For example, say we want to compare the effectiveness of three fertilizers for increasing tomato growth. We would apply each of the fertilizers (treatments) to randomly chosen tomato plants in several different gardens (block). We are interested in which fertilizer has the greatest effect on growth, but take into account variation across gardens because there may be unmeasured factors (soil fertility, exposure to the sun) that lead plants in some gardens to have more or less growth than other gardens.

Blocking removes as much variability as possible from the random error so that the differences among the groups are more evident. The focus of the analysis is on the difference among groups (or treatments), not the blocks.

In the completely randomized design (e.g. one-way ANOVA), the total variation (SS_T) is subdivided into variation due to differences among the treatment groups (SS_A) and variation within the groups (SS_E). Within-group variation is considered to be random variation and among-group variation is due to differences between groups and random variation. To remove the effects of the blocking from the random variation

component in the randomized block design, the within-group variation (SS_E) is subdivided into variation due to differences among the blocks (SS_{BL}) and random variation (SS_ϵ).

In this example, we want to test whether four different antibiotics result in different levels of antibodies in the blood. Sixteen people are randomly assigned one of the four antibiotics and samples of their blood were taken for analysis. To process the blood samples as quickly as possible, the samples were sent to four different laboratories – each lab receives one blood sample treated with one of the four antibiotics.

Let's create the dataframe.

```
lab <- c(rep(1:4, each = 4))
antibiotic <- rep(c(1:4), 4)
results <- (c(9.3, 9.4, 9.6, 10, 9.4, 9.3, 9.8, 9.9, 9.2, 9.4,
    9.5, 9.7, 9.7, 9.6, 10, 10.2))
dlabs <- data.frame(lab = factor(lab), antibiotic = factor(antibiotic),
    results)
```

Each laboratory has its own instruments and personnel that might cause variation in the results across laboratories. Our variable of interest is the level of antibodies in the blood samples; laboratories are blocks whose uncontrolled effects we want to separate from the main effect.

First, let's do the analysis without the block effect. What is your interpretation?

```
mod.no.blk <- lm(results ~ antibiotic, data = dlabs)
```

Now, let's include the block effect. Does this change your interpretation?

```
mod.blk <- lm(results ~ factor(antibiotic) + factor(lab), data = dlabs)
```

The `TukeyHSD` function we have been using for *post hoc* tests only works for `aov` models. If you want to conduct the *post hoc* test on an `lm` model, first install and load the `mosaic` package. Then you can use `TukeyHSD` for the test.

Hopefully, up to now this all seems pretty straightforward. `mod.blk` looks pretty much like a factorial ANOVA without the interaction, and we are treating the effect of lab as a nuisance. If it were only so easy...

Recall that there are two types of ANOVA. Fixed effects ANOVA applies when the treatments have been specifically chosen. For example, when you are interested in the effects of the particular antibiotics above. In other words, you want to know how antibiotics 1, 2, 3 and 4 specifically impact the results.

H_0: There is no difference in level of antibodies among antibiotics 1, 2, 3, and 4

Random effects ANOVA, applies to hypotheses that are more general. Instead of examining the effects of four specific antibiotics, your null hypothesis might be:

H_0: There is no difference in level of antibodies among all antibiotics.

Therefore, the antibiotics chosen are merely representatives of a wider range of antibiotics, even though your random selection might be antibiotics 1, 2, 3, and 4.

If we treat the blocking variable as a fixed effect, then the inference will only apply to those particular blocks (or samples). If we treat the blocking variable as a random effect, then inference can be made to the population of all possible blocks. The second option is what we are after; however, the rule of thumb is that you need at least six subjects (six samples) to estimate a random effect (i.e. variance) or the precision on the estimate cannot be estimated. Note that this also depends upon the assumption that the blocks are chosen randomly from a normal distribution of blocks.

The take-away message is that it is preferable to treat the blocking factor as a random effect, but that may not always be possible. Let's see what happens when we set up the model for the Model II ANOVA. Note that the term `(1|lab)` below classifies `lab` as a random effect, rather than a fixed effect. The package `lme4` runs the random effects model but `lmerTest` is necessary for p-values.

```
require(lme4)
require(lmerTest)
mod.re <- lmer(results ~ antibiotic + (1 | lab), data = dlabs)
summary(mod.re)
```

The other advantage of `lmer` is that it allows us to run *post hoc* tests. Install and load the packages `emmeans` and `multcomp`. The `glht()` function provides general linear hypotheses and multiple comparisons for parametric models. The `linfct()` argument specifies the linear hypothesis to be tested.

```
require(emmeans)
require(multcomp)

summary(glht(mod.re, linfct = mcp(antibiotic = "Tukey")))
```

Incidentally, the above model could be used for situations where you have a fixed effect and a random effect, where the random effect is not necessarily a block, but some other nuisance variable.

Repeated Measures Design

A repeated measure design is used when all members of a random sample are measured under a number of different conditions. As the sample is exposed to each condition in turn, the measurement of the dependent variable is repeated. Using a standard model would not be appropriate because it fails to model the correlation between the repeated measures: the data violate the assumption of independence.

Let's look at an example of two advertising campaigns in ten different cities. For each city, the sales before, during and after the advertising campaign are measured. We want to know if the campaign has a significant effect on sales.

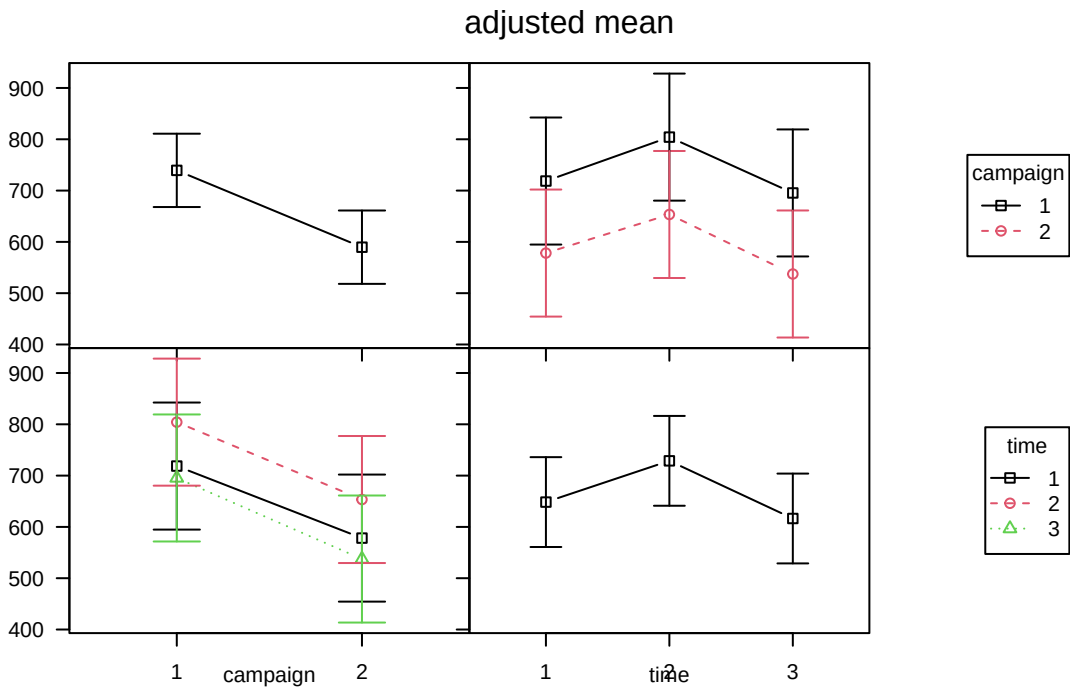
```
ad <- read.csv("Sales.csv", header = T)
ad <- with(ad, data.frame(sales, city = factor(city), campaign = factor(campaign),
    time = factor(time)))
```

Let's first proceed without including repeated measures. It looks like there is not a significant effect of campaign, time, or their interaction.

```
mod.sales <- lm(sales ~ campaign * time, data = ad)
summary(mod.sales)
```

The package `phia` offers another way to plot interactions. Download it and add it to your working library.

```
require(phia)
plot(interactionMeans(mod.sales), las = 1)
with(ad, tapply(sales, list(time), mean))
with(ad, tapply(sales, list(campaign), mean))
```



Now let's add in the repeated measures of city (i.e. the sales measurements were taken at the same city before, during, and after the ad campaign). In the random effect term we are indicating that campaign is nested within city. Don't worry about the results of the random effects yet, just examin the fixed effects. Does this change your interpretation?

```
mod.rep <- lmer(sales ~ campaign * time + (1 | city/campaign),
    data = ad)
summary(mod.rep)
```

To make it simpler, let's take out the interaction in the model

```
mod.rep <- lmer(sales ~ campaign + time + (1 | city/campaign),
    data = ad)
summary(mod.rep)
```

There doesn't seem to be a significant difference between the two marketing campaigns on sales, but sales during the campaign were significantly higher than before the campaign. Sales after the campaign were significantly lower than before the campaign.

Reporting Results

When reporting the results of a linear model with nominal variables, report the F-statistic, degrees of freedom, and p-value of the entire model. For example, for `mod.blk`: There is a significant difference in effects of different antiobiotics on levels of antibodies in the blood taking into account the effects of different laboratories ($F_{6,9} = 22.69$, $p < 0.001$).

For models with random effects (linear mixed effects models), we can't present the F-statistic and degrees-of-freedom because the df's are difficult to estimate accurately. We will talk more about this in the future.

Problems

Analyze the data for the three problems below. Make sure to use the appropriate ANOVA design for each study. For each problem, write a 1-page description of your analysis and the results. Please include your code at the bottom of your write-up in R Markdown. Each write-up should include the following information:

- Null and alternative hypotheses of your tests
- Justification for your choice of ANOVA model
- Results of your statistical test, interpreting your test in 2-3 sentences that include the appropriate reporting of the statistics
- An interpretation of any necessary *post-hoc* tests
- A description of how you checked the assumptions of your statistical test

Problem 1

Determine if the average weight of confiscated elephant tusks has decreased over time. Elephants are poached for their ivory, and USFWS authorities confiscate ivory when they find it entering the country. The data in *TuskData.csv* are the average weights of elephant tusks from 20 different seizure sites in 1970, 1990, and 2010. Include at least one graph that shows the means and standard errors or confidence intervals of the weight of tusks over the three years. It is not necessary to include a barplot if you prefer to graph your results in a different way.

Problem 2

Evaluate the effect of salt on plant biomass growth in 24 experimental vegetation plots (download `salt.csv`). The assigned treatment (one of 6 levels of salt addition: 10, 15, 20, 25, 30, 35 g m^{-2}) is applied to the soil of a plot and at the end of the experiment the biomass of plants in each plot is measured. The experimental units are grouped into four blocks of six plots each, based on geographic proximity, and the treatments are assigned completely at random within each block. Thus, each treatment occurs exactly once in each block.

Problem 3

Pangolins are scaly anteaters that inhabit the tropical forests of Asia and Africa and are hunted for their meat and their scales, which are made of hair. A researcher wants to evaluate the effect of diet on the thickness of pangolin scales, with the idea that thicker scales would better protect pangolins from predators. She rears pangolins and provides them with identical diets, but different doses (0.5, 1, 2 milligrams) of supplements (Vitamin B and Zinc). Download the `ScaleThickness.csv` file from Sakai and analyze the data to determine the potential effects of `doses` and `supplements` and whether there is an interaction between the two of them.