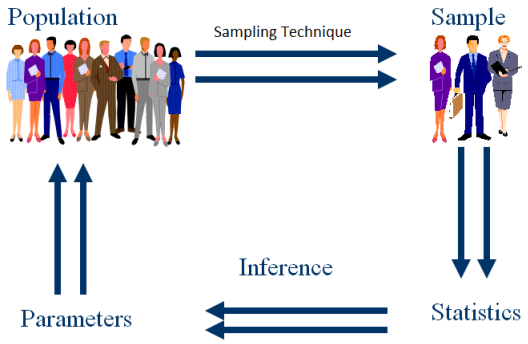
Chapter 1

ST 512 - Review

Readings: Chapters 1-8 as needed

Big ideas in stats:

- Population - all the values, items, or individuals of interest
- Parameter - a (usually) unknown summary value about the population
- Sample - a subset of the population we observe data on
- Statistic - a summary value calculated from the sample observations



Examples of paramters - (true) mean μ, (true) variance σ^2.
Examples of statistics - sample mean $\bar{y}$, sample variance $s^2 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}$
Inference - Making mathematically sound claims about the poulation using sample data.

Scales (Types) of Data:

- Qualitative or Categorical - A variable that is described by attributes or labels
  Subscales:
  Nominal - categories have no ordering (Male, Female)
  Ordinal - can order categories (Lickert scale data)

- Quantitative - A variable that is described by numerical measurements where arithmetic can be performed
  Subscales:
  Discrete - finite or countable finite number of values (# of flowers on a plant, 0, 1, 2, ...)
  Continuous - any value in an interval is possibel (Temperature, $(-459.67 \deg F, \infty)$

Random Variables and Things of Interest:

- Random Variable (RV) - Function that takes in outcomes from an experiment and outputs real numbers, or a numeric outcome to a random process

  Things of interest

  - Distribution - pattern and frequency of observable values
    For continuous RVs, visualized with a smooth curve.
  - Mean/Median - measures of center of the distribution

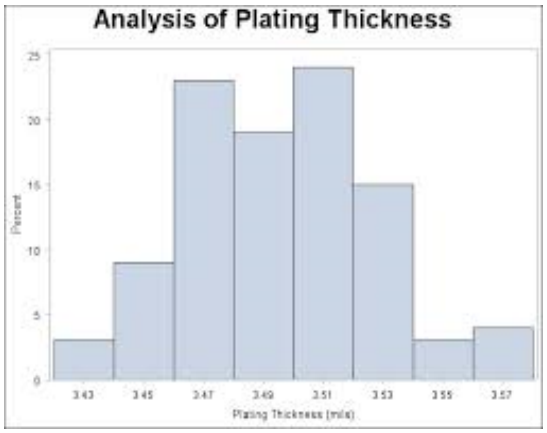
    Focus on mean: true mean μ, RV sample mean $\bar{Y}$, observed sample mean $\bar{y}$
  - Standard Deviation, Variance, IQR, Range - measures of spread for the distribution

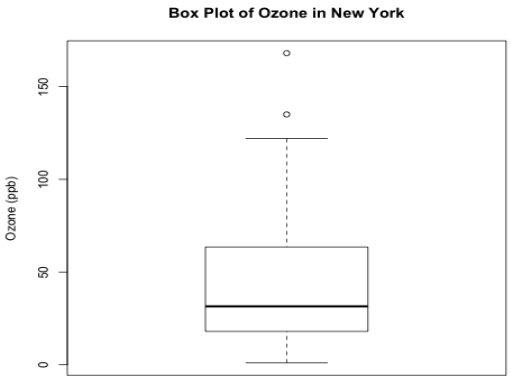
    Focus on SD and Variance: true variance σ^2, true SD σ, observed sample variance s^2, observed SD s

Graphical Descriptions of RV's:

- Histogram - Graphs the frequencies or relative frequencies of realizations of a RV

- Boxplot - Uses the Five Number Summary to display the realizations of a RV
  Five number summary: $min,\ Q_1,\ M,\ Q_3,\ max$





Statistics are also RVs. The distribution of a statistic is called a sampling distribution

Central Limit Theorem (CLT):

If a RV Y has a (true) mean μ and (true) variance σ^2, and a random sample is of size $n \geq 30$ is taken then

$$\bar{Y} \sim N(\mu, \sigma^2/n)$$

Note: If $Y \sim N(\mu, \sigma^2)$ then $\bar{Y} \sim N(\mu, \sigma^2/n)$ for any n.

2 main ways to make inference about a (true) mean, μ:

1. When the true SD, σ, is known we looked at the sampling distribution of the statistic

   $Z = \frac{\bar{Y}-\mu}{\sigma/n} \sim N(0,1)$ valid if $\bar{Y}$ has a normal distribution

   Allows us to form a CI: And a test statistic: Testing $H_0 : \mu = \mu_0$

   $\bar{y} \pm z_{\alpha/2}\sigma/\sqrt{n}$ $z_{obs} = \frac{\bar{y}-\mu_0}{\sigma/\sqrt{n}}$

2. When the true SD, σ, is unknown we looked at the sampling distribution of the statistic

   $T = \frac{\bar{Y}-\mu}{s/n} \sim t_{n-1}$ valid if $\bar{Y}$ has a normal distribution, allow for $n \geq 15$ or so in CLT

   Allows us to form a CI: And a test statistic: Testing $H_0 : \mu = \mu_0$

   $\bar{y} \pm t_{(n-1,\alpha/2)}s/\sqrt{n}$ $t_{obs} = \frac{\bar{y}-\mu_0}{s/\sqrt{n}}$

Inference about two (true) means, μ_1 and μ_2:

- From paired samples, $x_1, x_2, ..., x_n$ and $y_1, y_2, ..., y_n$ where difference is normally distributed

  CI: $(\bar{x} - \bar{y}) \pm t_{(n-1,\alpha/2)} s_{diff}/\sqrt{n}$

  HT: $H_0 : \mu_1 = \mu_2$, i.e. $\mu_1 - \mu_2 = 0 \quad t_{obs} = \frac{(\bar{x}-\bar{y})-0}{s_{diff}/\sqrt{n}}$

- Two separate samples from normal populations, $x_1, x_2, ..., x_n$ and $y_1, y_2, ..., y_n$

  CI: $(\bar{x} - \bar{y}) \pm t_{(\nu,\alpha/2)} \sqrt{s_X^2/n + s_Y^2/m}$ where ν is an estimate of df

  HT: $H_0 : \mu_1 = \mu_2$, i.e. $\mu_1 - \mu_2 = 0 \quad t_{obs} = \frac{(\bar{x}-\bar{y})-0}{\sqrt{s_X^2/n + s_Y^2/m}}$

Extension to inference about t (true) means, $\mu_1, \mu_2, ..., \mu_t$:

Balanced One-way ANOVA table (same number of replicates per group)

Source	DF	SS	MS	F-stat	P-value
Treatment	$t-1$	$n\sum_{i=1}^{t}(\bar{Y}_{i+} - \bar{Y}_{++})^2$	$\frac{SS(Trt)}{t-1}$	$\frac{MS(Trt)}{MS(E)}$	Use $F(t-1, t(n-1))$
Error	$t(n-1)$	$\sum_{i=1}^{t}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{i+})^2$	$\frac{SS(E)}{t(n-1)}$		
Total	$nt-1$	$\sum_{i=1}^{t}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{++})^2$			

Analysis used for a completely randomized design.

P-value tests $H_0 : \mu_1 = \mu_2 = ... = \mu_t$ vs H_A : at least 1 mean differs

One Way ANOVA model:

$$Y_{ij} = \mu + \alpha_i + E_{ij}$$

where $i = 1, 2, ...t$ and $j = 1, 2, ..., n$ (total sample size $= nt = N$)

μ = overall mean
α_i = effect from group i
E_{ij} = random error assumed to be $iid\ N(0, \sigma^2)$

For two quantitative variables measured on the same units, the linear relationship can be investigated:

Simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + E_i$ or use correlation.

For a hypothesis test, the p-value means

probability of obsevign a test statistic as extreme or more extreme than the one observed, assuming the null hypothesis is true.

For a given a null hypothesis, statistical significance implies

the observed value was unlikely to have occurred by random chance alone (assuming the null hypothesis is true).

For an observed confidence interval (cL, cU) we can say

We are ___% confident the true parameter value is contained in the interval. (***Do not say probability or chance!)

The idea of Confidence means

The procedure used to create the interval has a __% probability of producing an interval that contains the parameter.

i.e. If the experiment were done repeatedly and an interval made for each sample, __% of the intervals would contain the parameter value.

Chapter 2

ST 512 - Experiments

Readings: 7.2 and 7.3, pg 244-255

Example: An experiment was run to determine the effects of adding phosphorous ($0, 147, 294, 441$ kg/m^2) and nitrogen ($0, 45, 90, 135$ kg/m^2) to the soil of a certain type of grass (a Miscanthus species). The growth of the plant was of interest and at the end of the growing period the plant was dried and the weight recorded with the final measurement being recorded in megagram per hectare (0.1 kg/m^2). Four plots of grass were used in total. Within each plot, each combination of phosphorous and nitrogen was observed. A partial data table is given here:

Plot	P	N	Dry yield
1	0	135	1.95
1	0	45	3.51
1	0	90	2.87
1	0	0	2.88
1	294	45	2.37
1	294	0	3.5
1	294	135	3.55
1	294	90	4.4
...	...	...	...

Let's identify (if possible) the response, explanatory variable(s), factor(s), level(s), confounding factor(s), treatment(s), number of replicates, and experimental units.

Sources of Variation in the responses of an experiment:

1. **Treatment effect** - we hope there is an effect due to the variables we control
2. **Identified confounding variables** - We record some variables that are not of interest, but we think may have an effect on the response.
3. **Unidentified sources (Experimental Error or error variation)** -
   (a) Inherent variability in experimental units - Experimental units are different!
   Ex: No two people, paper towels, concrete blocks, or even lab rats are exactly the same.
   Consequence: Experimental units respond differently to the same treatment
   (b) Measurement error - Multiple measurements of a same experimental unit typically contain error.
   If the same experimental unit is measured more than once, will the value be the same?
   Ex: Blood Pressure, Quality Ratings of food, Break a water sample in two, measure each for bacteria
   (c) Variations in applying/creating treatments
   The treatment is not clearly defined, leaving room for interpretation.
   Ex: Two researchers mix concrete, will it come out exactly the same? Ovens don‘t heat exactly the same, etc.
   (d) Effects from any other extraneous (or lurking) variables - Extraneous variables are those variables that are not part of the treatment, but may influence the response.

Let‘s identify these in the previous example.

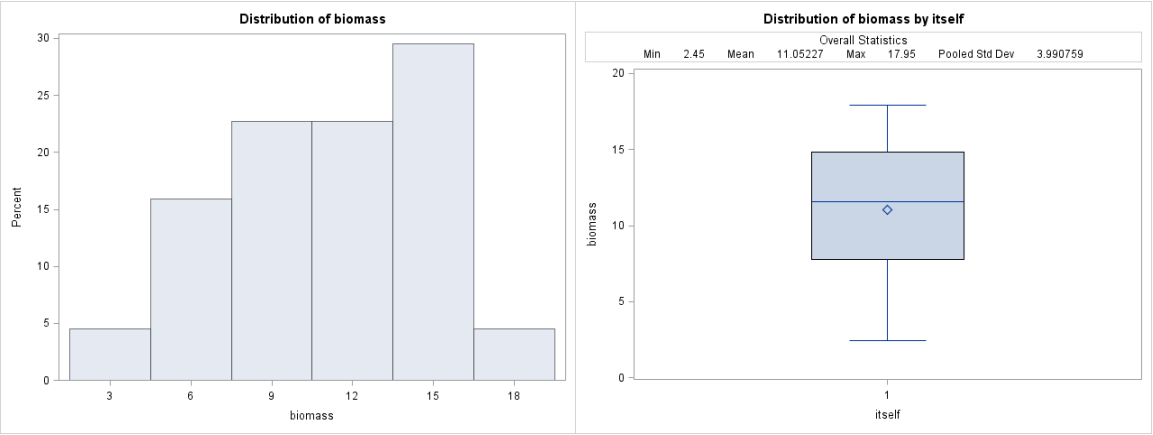
Chapter 3

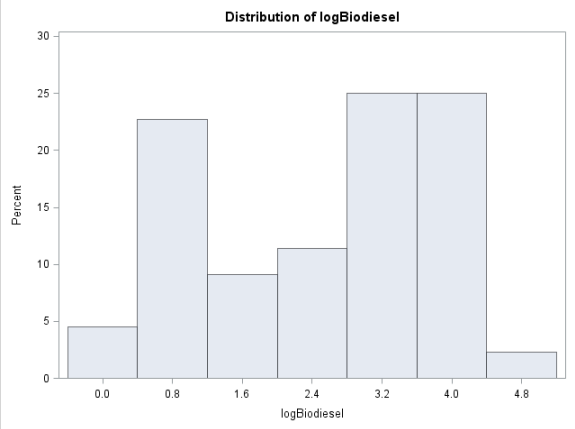
ST 512 - Correlation

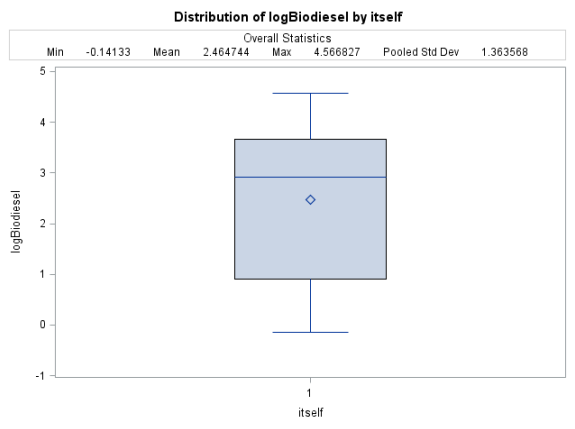
Readings for Correlation and SLR: 10.1-10.5 pg 378-420 and 10.7-10.8 pg 425-444 and 8.7 pg 305-311

Motivating example: One type of fuel is biodiesel, which comes from plants. An experiment was done to determine how much biodiesel could be generated from a certain type of plant grown in different medias. The final biomass was also recorded on 44 the plants from the experiment. Let's consider these two variables, the log of biodiesel and biomass.

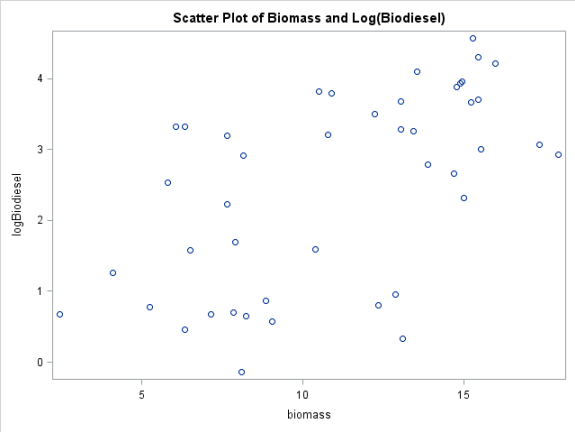
We can look at the distribution of each individually using our univariate methods (histogram, boxplot, mean/sd, etc.)

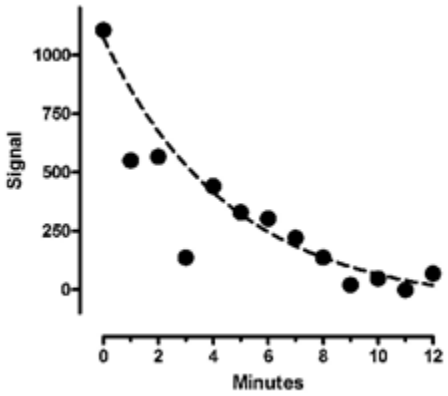






How can we visually inspect the association between the two? A **Scatter plot** gives a visual approximation of the "joint distribution" between two variables.





Properties of r_{XY}

- r_{XY} is an observed measure of the linear assn. between X and Y in a dataset.
- correlation coefficient is unitless and always between -1 and 1:

$$-1 \leq r_{XY} \leq 1$$

- The closer r_{XY} is to 1, the stronger the positive linear association
- The closer r_{XY} is to -1, the stronger the negative linear association
- The bigger $|r_{XY}|$, the stronger the linear association
- If $|r_{XY}| = 1$, then X and Y are said to be perfectly correlated (relationship is deterministic)

For the log(Biodiesel) (call this Y) and Biomass (call this X) example we can compute the sample correlation coefficient using summary statistics:

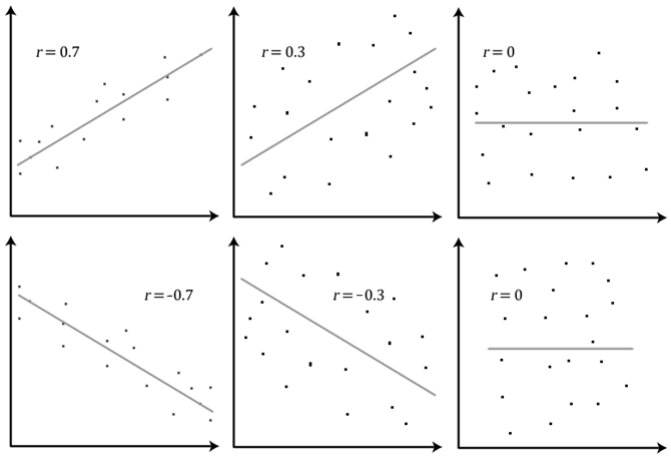
$$\bar{x} = 11.0523, \quad s_X = 3.9908, \quad \bar{y} = 2.4647, \quad s_Y = 1.3636$$

$$s_{XY} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1} = 3.1485$$

Applying the formula for r_{XY}, we get

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{3.1485}{\sqrt{3.9908 \times 1.3636}} = 0.5786$$

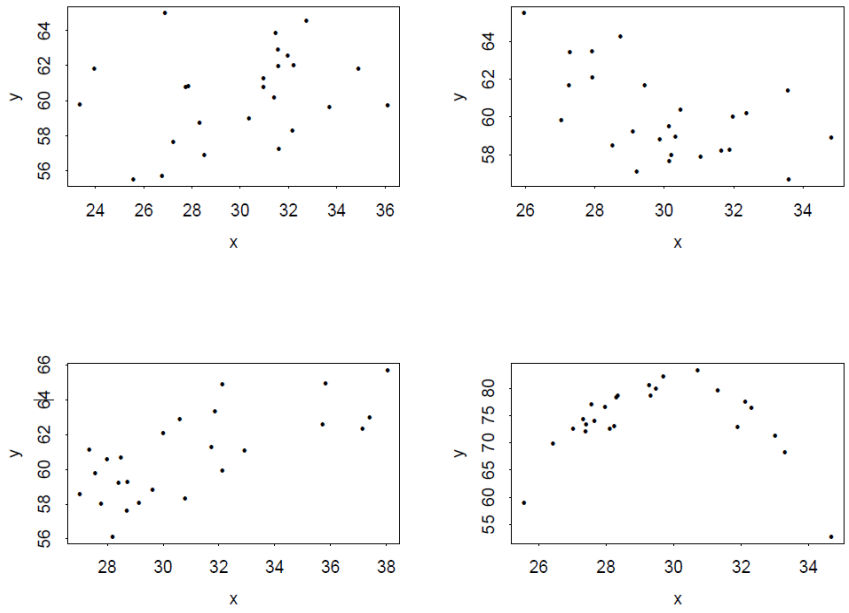
Some example scatter plots



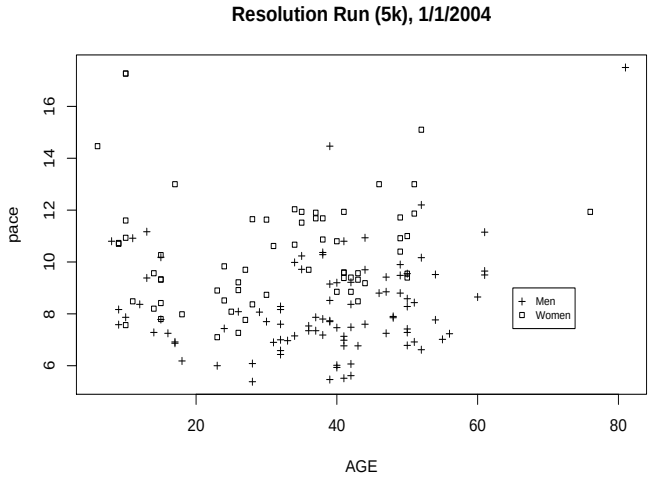
An exercise/activity:

Label the four plots below with the four sample correlation coefficients:

- $r = 0.3$ $\qquad$ $r = 0.7$
- $r = 0.1$ $\qquad$ $r = -0.6$



Would it be appropriate to use correlation to summarize the relationship between age and pace in the following scatter plot? Why or why not?



To perform a Hypothesis Test about ρ:
We often want to test the following hypotheses,

$$H_0 : \rho = 0 \qquad H_A : \rho \neq 0$$

Assuming H_0 is true, the test statistic is

$$z_{obs} = \left(\frac{1}{2}\sqrt{n-3}\right) \log \frac{1+r}{1-r}$$

and the reference distribution is the standard normal distribution, i.e. reject if $z_{obs} > z_{\alpha/2}$ or if $z_{obs} < z_{1-\alpha/2}$ where z_α satisfies $\alpha = \Pr(Z > z_\alpha)$ with $Z \sim N(0,1)$.

The p-value if found by finding $2P(Z > |z_{obs}|)$. Why do we multiply by 2?

To find a Confidence Interval for ρ:
An approximate $100(1-\alpha)\%$ confidence interval for ρ can be obtained by inverting the *Fisher transformation*:

$$\left(\frac{\frac{1+r}{1-r}e^{-2z_{\alpha/2}/\sqrt{n-3}} - 1}{\frac{1+r}{1-r}e^{-2z_{\alpha/2}/\sqrt{n-3}} + 1}, \frac{\frac{1+r}{1-r}e^{2z_{\alpha/2}/\sqrt{n-3}} - 1}{\frac{1+r}{1-r}e^{2z_{\alpha/2}/\sqrt{n-3}} + 1}\right).$$

For the log(Biodiesel) and Biomass example our hypothesis test is:

$$H_0 : \rho = 0 \qquad H_A : \rho \neq 0$$

$$\text{giving a test statistic of } z_{obs} = \frac{1}{2}\sqrt{44-3}\ log\left(\frac{1+0.5786}{1-0.5786}\right) = 4.228$$

Using an $\alpha = 0.05$ our rejection region is any z_{obs} outside of ± 1.96.

Our p-value $= 2P(Z > 4.228) = 2(0.00001) = 0.00002 < \alpha = 0.05$ so we reject our null hypothesis in favor of the alternative.

What is the interpretation of the p-value=0.00002?
The probability of getting a sample correlation (r) further (in magnitude) from 0 than 0.5786 assuming the true correlation (ρ) is 0 is 0.00002.

The corresponding 95% confidence interval is

$$\left(\frac{\frac{1+0.5786}{1-0.5786}e^{-2*1.96/\sqrt{44-3}}-1}{\frac{1+0.5786}{1-0.5786}e^{-2*1.96/\sqrt{44-3}}+1}, \frac{\frac{1+0.5786}{1-0.5786}e^{2*1.96/\sqrt{44-3}}-1}{\frac{1+0.5786}{1-0.5786}e^{2*1.96/\sqrt{44-3}}+1}\right) = (0.3401, 0.7471)$$

We can say that we are 95% confident that the true correlation (ρ) is between 0.3401 and 0.7471.

When we say confident, we mean that if we did this experiment repeatedly and made an interval for each experiment, the true correlation would fall in 95% of the intervals created.

How can we get SAS to do this for us?

```
proc corr data=bioexp FISHER(biasadj=NO);
var butterfat temp;
run;
```

Output From Proc Corr for Biomass and Log(Biodiesel) Example **1**

The CORR Procedure

2 Variables:	biomass	logBiodiesel

Covariance Matrix, DF = 43		
	biomass	logBiodiesel
biomass	15.92615751	3.14851427
logBiodiesel	3.14851427	1.85931767

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
biomass	44	11.05227	3.99076	486.30000	2.45000	17.95000
logBiodiesel	44	2.46474	1.36357	108.44873	-0.14133	4.56683

Pearson Correlation Coefficients, N = 44 Prob > |r| under H0: Rho=0		
	biomass	logBiodiesel
biomass	1.00000	0.57859 <.0001
logBiodiesel	0.57859 <.0001	1.00000

Pearson Correlation Statistics (Fisher's z Transformation)							
Variable	With Variable	N	Sample Correlation	Fisher's z	95% Confidence Limits		p Value for H0:Rho=0
biomass	logBiodiesel	44	0.57859	0.66035	0.340140	0.747136	<.0001

Note: Significant correlation does NOT imply causation

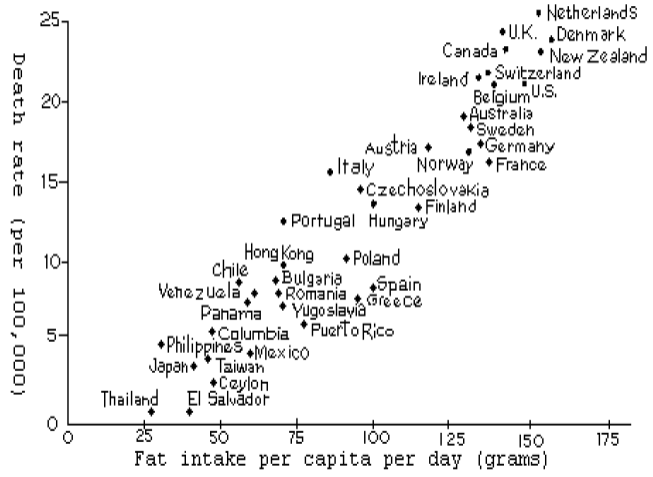
Famous examples of *spurious correlations*:

- A study finds a high positive correlation between coffee drinking and coronary heart disease. Newspaper reports say the fragrant essence of the roasted beans of *Coffea arabica* are a menace to public health.
- In a city, if you were to observe the amount of damage and the number of fire engines for enough recent fires, you would likely see a positive and significant correlation among these variables. Obviously, it would be erroneous to conclude that fire engines cause damage.
- *Lurking variable* - a third variable that is responsible for a correlation between two others. (A.k.a. confounding factor.)
  An example would be to assess the association between say the reading skills of children and other measurements taken on them, such as shoesize. There may be a statistically significant association between shoe size and reading skills, but that doesn't imply that one causes the other. Rather, both are positively associated with a third variable, *age*.
- Among 50 countries examined in a dietary study, high positive correlation among fat intake and cancer (see figure, next page). This example is taken from from *Statistics* by Freedman, Pisani and Purves.

> In countries where people eat lots of fat like the United States rates of breast cancer and colon cancer are high. This correlation is often used to argue that fat in the diet causes cancer. How good is the evidence?
>
> Discussion. If fat in the diet causes cancer, then the points in the diagram should slope up, other things being equal. So the diagram is some evidence for the theory. But the evidence is quite weak, because other things aren't equal. For example, the countries with lots of fat in the diet also have lots of sugar. A plot of colon cancer rates against sugar consumption would look just like figure 8, and nobody thinks that sugar causes colon cancer. As it turns out, fat and sugar are relatively expensive. In rich countries, people can afford to eat fat and sugar rather than starchier grain products. Some aspects of the diet in these countries, or other factors in the life-style, probably do cause certain kinds of cancer and protect against other kinds. So far, epidemiologists can identify only a few of these factors with any real confidence. Fat is not among them.

(p. 152, *Statistics* by Friedman, Pisani, Purves and Adhikari)

Figure 8. Cancer rates plotted against fat in the diet for a sample of countries



Source: K. Carroll. "Experimental evidence of dietary factors and hormone-dependent cancers Cancer Research vol. 35 (1975) p.3379. Copyright by Cancer Research. Reproduced by permission

Chapter 4

ST 512 - Simple Linear Regression

Readings for Correlation and SLR: 10.1-10.5 pg 378-420 and 10.7-10.8 pg 425-444 and 8.7 pg 305-311

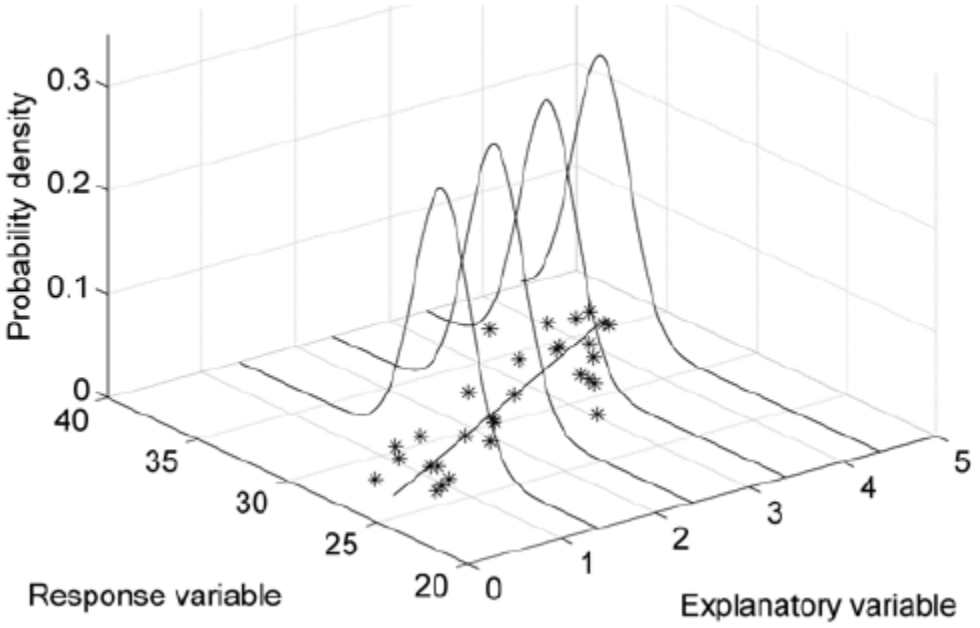
Fit a linear regression model - A probabilistic model for Y conditional on $X = x$:

$$Y_i = \beta_0 + \beta_1 x_i + E_i$$

Defintions:

- Y_i - response (also called dependent variable)
- x_i - explanatory variable (also called independent variable or predictor variable)
- E_i - random error for observation i
- $\beta_0 = E(Y|X = 0)$ - True population intercept (average value of response when $X = 0$
- β_1 - True population slope (average change in Y per unit increase in x)
- σ^2 - Error variance (variance due to experimental error)

Note: We make the assumption that $E_1, \ldots, E_n$ are independent and identically distributed normal random variables with mean 0 and variance σ^2. We write $E_i \overset{iid}{\sim} N(0, \sigma^2)$. This variance is assumed the same for all x, called assumption of **homoskedasticity**.



1. $E(Y|X = x) = \beta_0 + \beta_1 x = \mu(x)$ (The line describes the mean Y for a given X.)
2. $\text{Var}(Y|X = x) = \sigma^2$

For the log(Biodiesel) and Biomass example let's find our fitted line. Recall the summary stats on page 10.

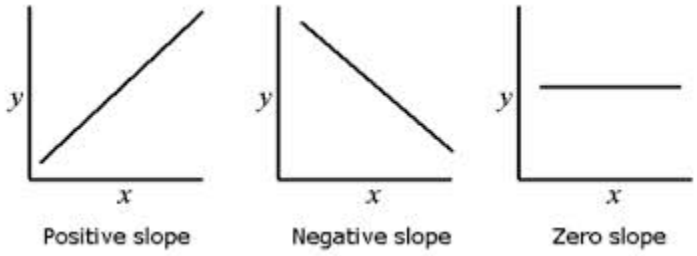
$$\hat{\beta}_1 = s_{XY}/s_X^2 = 3.1485/3.9908^2 = 0.1977$$
$$\hat{\beta}_0 = 2.4647 - 11.0523 * 0.1977 = 0.2797$$
$$\hat{y} = 0.2808 + 0.1977x$$

This line can now be used to make predictions for new X values by simply plugging in the x!

Again, we have now have point estimates for our true parameters. How can we make inference (claims about the true values)? Do we have a *significant linear relationship*?
Under the normal distribution assumption on the errors, the RV's $\hat{\beta}_0$ and $\hat{\beta}_1$ follow normal distributions. Thus, we can use this as a basis for inference.

What value of the slope do we test?

- If a linear relationship, Y will tend to change with X (i.e. $\beta_1 \neq 0$)
- If no linear relationship, Y won't tend to change with X (i.e. $\beta_1 = 0$).



Any hypothetical slope, like $\boxed{H_0 : \beta_1 = \text{slope}_0}$ may be tested using the T-statistic below with $df = n - 2$:

$$T = \frac{\hat{\beta}_1 - \text{slope}_0}{\widehat{SE}(\hat{\beta}_1)}$$

and any hypothetical intercept, like $\boxed{H_0 : \beta_0 = \text{intercept}_0}$ may be tested using the T-statistic below with $df = n - 2$:

$$T = \frac{\hat{\beta}_0 - \text{intercept}_0}{\widehat{SE}(\hat{\beta}_0)}$$

Confidence intervals for β_0, β_1

$100(1-\alpha)\%$ confidence intervals for β_0 and β_1 are given by

$$\hat{\beta}_0 \pm t(n-2, \alpha/2)\sqrt{MS[E]\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}.$$

$$\hat{\beta}_1 \pm t(n-2, \alpha/2)\sqrt{\frac{MS[E]}{S_{xx}}}.$$

Often we will only care about the test and CI for the slope. The hypothesis test is equivalent to checking if 0 is in the confidence interval. It will depend on the context of the question if testing β_0=0 makes sense.

Confidence interval for $\mu(x_0) = E(Y|X = x_0)$
The point estimate for $\mu(x_0)$ is simply $\hat{\beta}_0 + \hat{\beta}_1 x_0$. We need to know about the variability of this estimate and we can again use the t-distribution for inference.

$$\text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0 | X = x_0) \quad =$$

This yields a confidence interval of the form

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t(n-2, \alpha/2)\sqrt{MS[E]\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

Note: We are attempting to capture the true *mean* at x_0 in this interval.

Prediction interval for a new observation x_0
The point estimate for at x_0 is still $\hat{Y}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$. However, the variability will change.

$$\text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0 + E_{new} | X = x_0) \quad =$$

Thus we can form a PI using

$$\hat{Y}(x_0) \pm t(n-2, \alpha/2)\sqrt{MS[E]\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}.$$

Note: In this interval we are attempting to capture the next Y value that takes on x_0. As this is a much more difficult task, PI's are wider than CI's.

The ANOVA table from simple linear regression

The full ANOVA table for SLR is given below:

Source	Sum of squares	df	Mean Square	F-Ratio
Regression	$SS(R)$	1	$MS(R)$	$MS(R)/MS(E)$
Error	$SS(E)$	$n-2$	$MS(E)$	
Total	$SS(Tot)$	$n-1$		

The mean squares represent standardized measures of variation due to the different sources and are given by $SS(source)/dfsource$. Ratios of mean squares often follow an F-distribution and are appropriate for testing different hypotheses of interest.

In this case, to test

$$H_0 : \beta_1 = 0 \qquad \text{vs} \qquad H_1 : \beta_1 \neq 0$$

$F = MS(R)/MS(E) \sim F(1, n-2)$.

That is, the F statistic follows an F-distribution with 1 numerator df and $n-2$ denominator df. In SLR, this F test is equivalent to the T test we already looked at. The relationship is that $T^2 = F$.

Note: The mean square for error, $MS[E]$, is an unbiased estimator for σ^2. It is an estimate of the variability due left over once we account for our explanatory variable.

How to get tests in SAS?
For our Biodiesel and Biomass example we can get much of our output from SAS using the following commands:

```
proc reg data=bioexp ;
model logbiodiesel=biomass/clb;
run;
```

Output From Proc Reg for Biomass and Log(Biodiesel) Example 1

The REG Procedure
Model: MODEL1
Dependent Variable: logBiodiesel

Number of Observations Read	44
Number of Observations Used	44

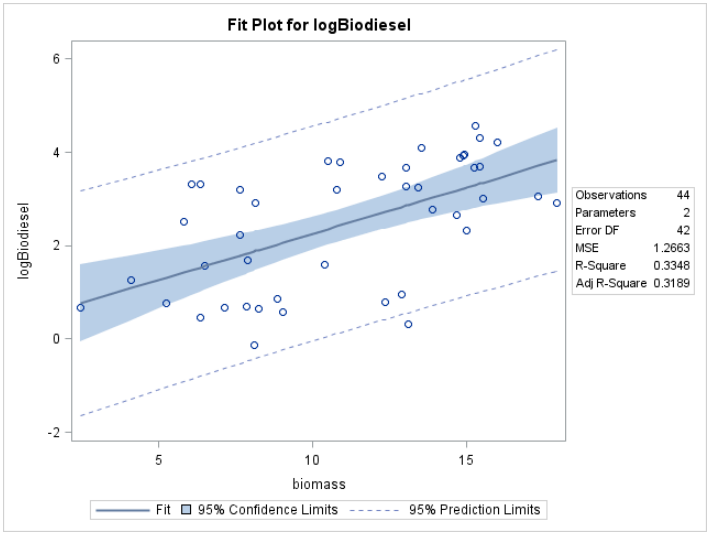
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	26.76509	26.76509	21.14	<.0001
Error	42	53.18557	1.26632		
Corrected Total	43	79.95066			

Root MSE	1.12531	R-Square	0.3348
Dependent Mean	2.46474	Adj R-Sq	0.3189
Coeff Var	45.65627		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	0.27977	0.50463	0.55	0.5822	-0.73862	1.29816
biomass	1	0.19769	0.04300	4.60	<.0001	0.11091	0.28447

Using $\alpha = 0.05$, (1) let's find the CI for the slope by hand, (2) form a CI for the mean log of biodiesel when biomass is 12, and (3) form a PI for a future log biodiesel measurement for a biomass of 12.

SAS will also produce a very nice plot that includes *pointwise* confidence and prediction bands at all points. Notice that the bands get wider the further x_0 is from $\bar{x}$. Why?

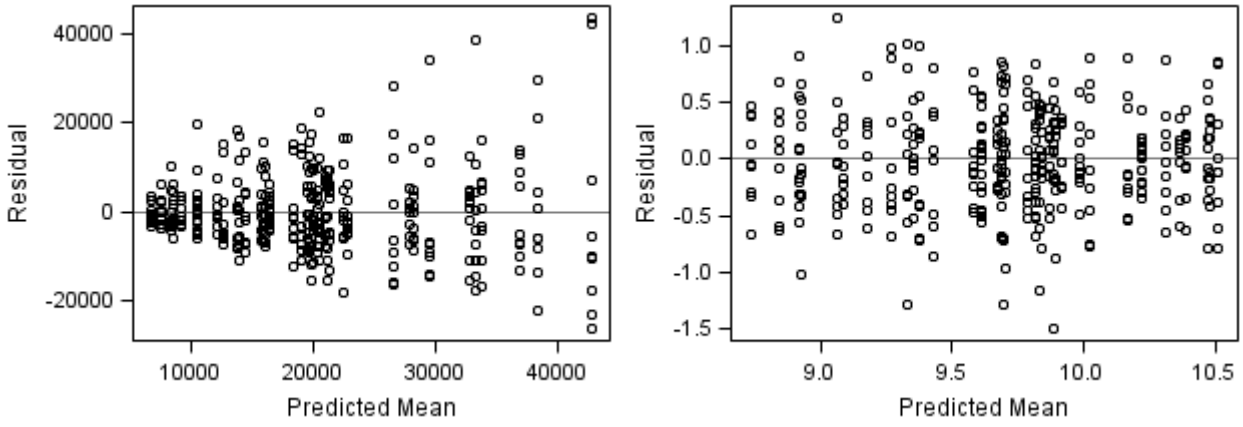


Checking assumptions

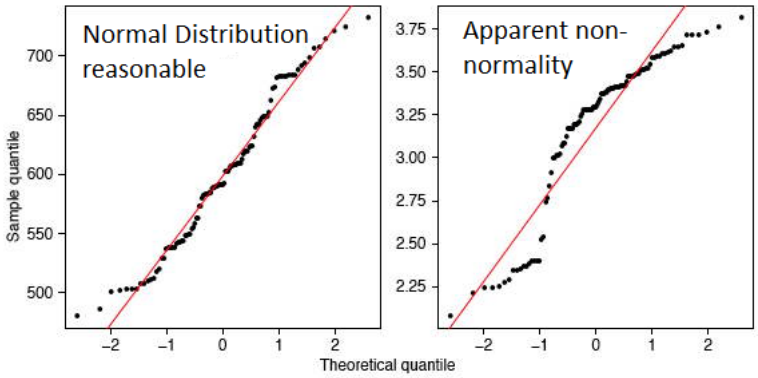
Firstly, we should always inspect a scatter plot to determine if the linear relationship we are assuming in our model is appropriate.

Secondly we can check our assumption of $iidN(0, \sigma^2)$ errors.

- Independence - There is not a check for independence of errors, we simply need to consider whether or not our EUs can be considered independent.
- Constant variance - A residuals vs fitted (predicted) values plot or a residual vs independent variable plot are tools for detecting heteroskedasticity (non-constant variance).



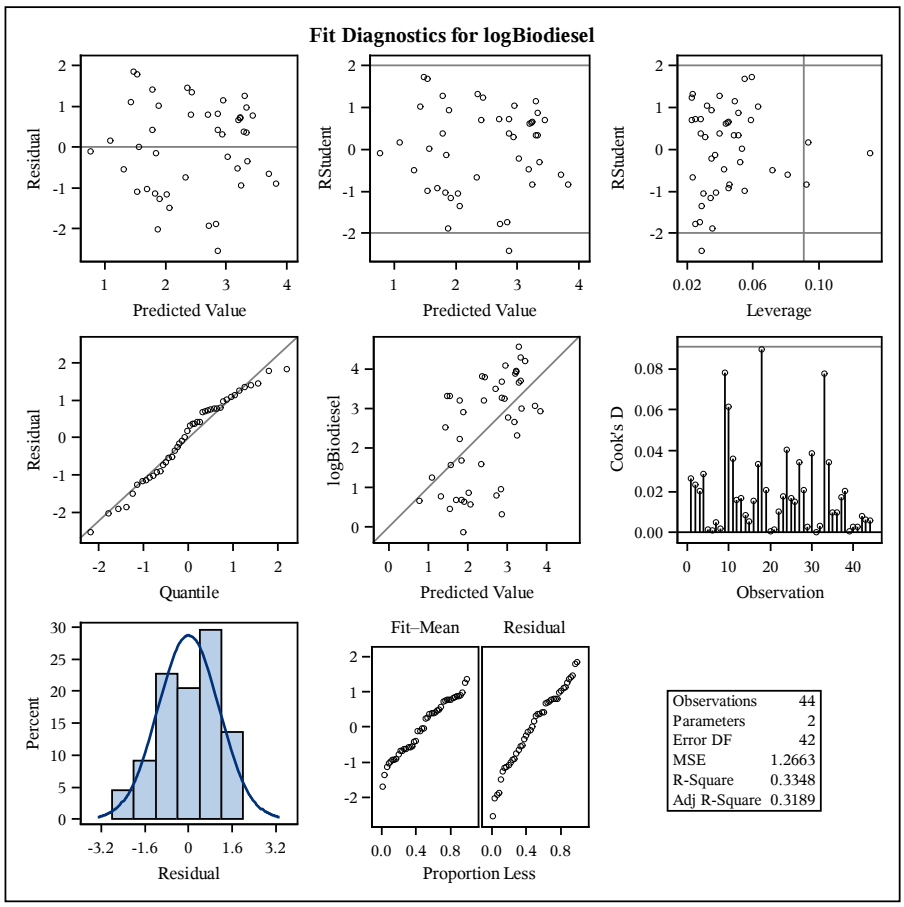
- Normality of errors - A quantile-quantile plot (or qq-plot for short) can be inspected to see if normality is reasonable.



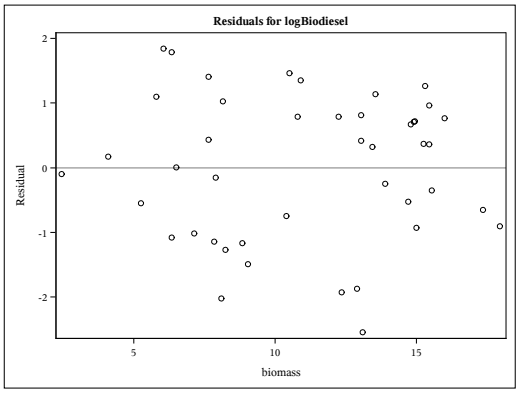
We can inspect the diagnostic plots that SAS produces when the reg procedure is used:

Output From Proc Reg for Biomass and Log(Biodiesel) Example **2**

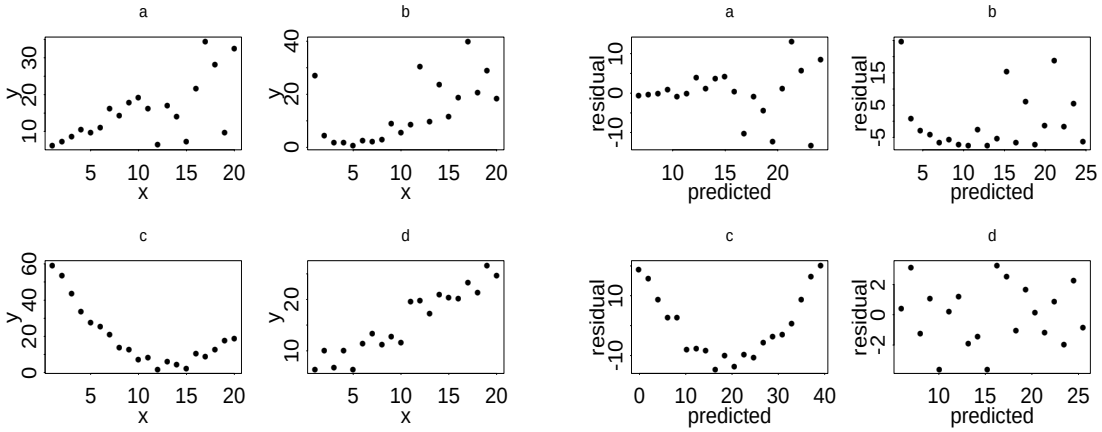
The REG Procedure
Model: MODEL1
Dependent Variable: logBiodiesel

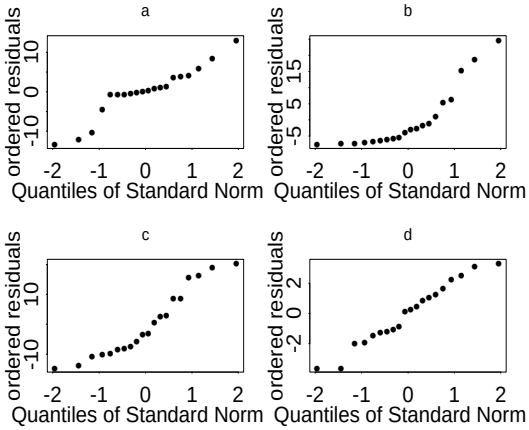


The REG Procedure
Model: MODEL1
Dependent Variable: logBiodiesel



An exercise: Match up letters a,b,c,d with the model violation - Heteroscedasticity, Nonlinearity, Nonnormality, Model fits





Chapter 5

ST 512 - Multiple Linear Regression

Readings: 11.1-11.6 and 11.9-11.11 pg 463 - 515 and 529 - 539

Motivating Example
(Taken from Probabilitiy and Statistics, Devore) Soil and sediment adsorption, the extent to which chemicals collect in a condensed form on the surface, is an important characteristic influencing the effectiveness of pesticides and various agricultural chemicals. A study was done on 13 soil samples that measured Y = phosphate adsorption index, X_1 = amount of extractable aluminum, and X_2 = amount of extractable iron. The data are given below:

Adsorption	Aluminum	Iron
4	13	61
18	21	175
14	24	111
18	23	124
26	64	130
26	38	173
21	33	169
30	61	169
28	39	160
36	71	244
65	112	257
62	88	333
40	54	199

MLR model for two quantitative explantory variables:
For observation i we can use the model

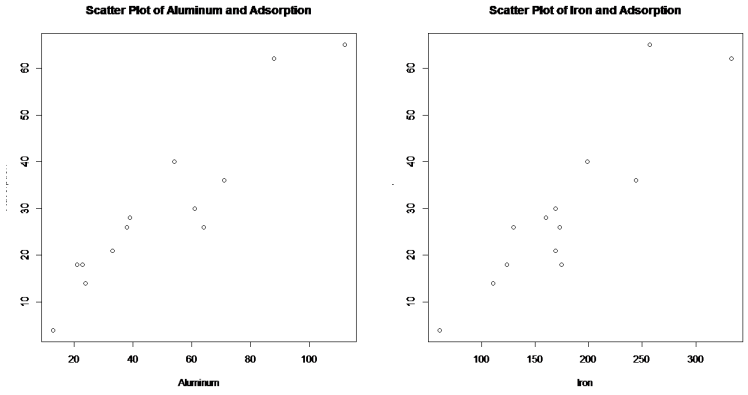
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + E_i$$

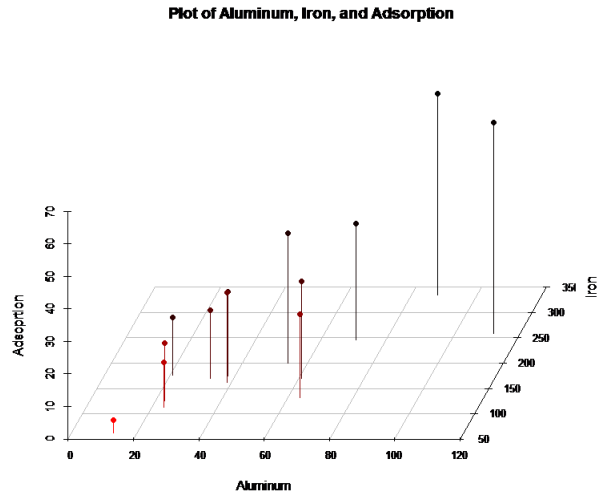
For clarity we can write the model for each subject

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + E_1 \\ Y_2 &= \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + E_2 \\ \vdots &= \vdots \\ Y_{13} &= \beta_0 + \beta_1 X_{13,1} + \beta_2 X_{13,2} + E_{13} \end{aligned}$$

Generally our model is

$$Adsorption = \beta_0 + \beta_1 Aluminum + \beta_2 Iron + Experimental\ Error$$

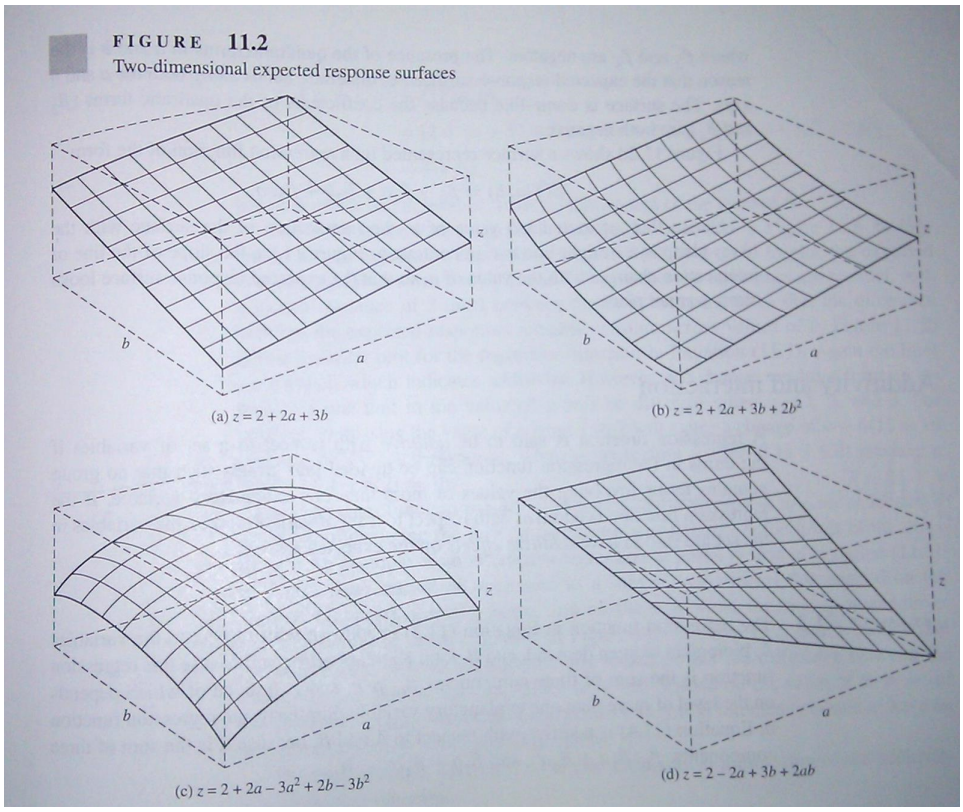




In this case, we don't want to find the best fitting line, but rather the best fitting *plane* (the one the minimizes the squared distances between the plane and the data points). Our hypothesis of interest is that at least one of our variables is useful (i.e. at least one partial slope is truly non-zero). We can then test

$$H_0 : \beta_1 = \beta_2 = 0 \quad vs \quad H_A : \text{ at least one is non-zero}$$

When we fit an MLR model with p different predictors we are really attempting to find the best 'response surface' of degree p in a $p + 1$ dimensional space. For instance, with one predictor, we are fitting the best line in a 2-d space. The plots below give a number of surfaces that can be fit using two predictors when quadratic or interaction terms are included.



A link to visualizing different surfaces:

`http://www.ats.ucla.edu/stat/sas/teach/reg_int/reg_int_cont.htm`

Very brief matrix review:
Note: Capital boldface letters are usually used for matrices and boldface lower case letters are usually used for vectors (matrices where the number of rows or the number of columns is 1).

Matrices - rectangular arrays of numbers that have a great many uses. Some matrices, (with *dimension* in parentheses):

$$\begin{aligned}
\mathbf{A} &= \begin{pmatrix} 7 & 5 \\ 5 & 2 \\ 3 & 2 \end{pmatrix} \quad (3 \times 2) \\
\mathbf{B} &= \begin{pmatrix} 4 & 2 & 1 \\ 3 & 1 & 1 \end{pmatrix} \quad (2 \times 3) \\
\mathbf{C} &= \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \quad (2 \times 2) \\
\mathbf{I}_2 &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad (2 \times 2)
\end{aligned}$$

Matrix operations

1. Transposition - swap rows for columns, columns for rows:

$$t(\mathbf{A}) = \mathbf{A}' = \begin{pmatrix} 7 & 5 & 3 \\ 5 & 2 & 2 \end{pmatrix} \quad \text{“transpose of } \mathbf{A}\text{”}$$

2. Addition is elementwise, matrices must have same *dimension*

$$\mathbf{C} + \mathbf{I}_2 = \begin{pmatrix} 1+1 & 1+0 \\ -1+0 & 1+1 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ -1 & 2 \end{pmatrix} \quad (2 \times 2)$$

   Subtraction, same deal

$$\mathbf{C} - \mathbf{I}_2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \quad (2 \times 2)$$

3. Multiplication requires *conformability*. Element in i^{th} row, j^{th} column of $\mathbf{AB}$ is dot-product of i^{th} row of $\mathbf{A}$, j^{th} column of $\mathbf{B}$:

$$\begin{aligned}
\mathbf{AB} &= \begin{pmatrix} 7 & 5 \\ 5 & 2 \\ 3 & 2 \end{pmatrix} \begin{pmatrix} 4 & 2 & 1 \\ 3 & 1 & 1 \end{pmatrix} \\
\mathbf{AB} &= \begin{pmatrix} 7\cdot 4 + 5\cdot 3, & 7\cdot 2 + 5\cdot 1, & 7\cdot 1 + 5\cdot 1 \\ 5\cdot 4 + 2\cdot 3, & 5\cdot 2 + 2\cdot 1, & 5\cdot 1 + 2\cdot 1 \\ 3\cdot 4 + 2\cdot 3, & 3\cdot 2 + 2\cdot 1, & 3\cdot 1 + 2\cdot 1 \end{pmatrix} \\
&= \begin{pmatrix} 43 & 19 & 12 \\ 26 & 12 & 7 \\ 18 & 8 & 5 \end{pmatrix}
\end{aligned}$$

(The product **DE** is not necessarily equal to **ED**). The matrices **D** and **E** are conformable for the product **DE** if **D** has the same number of columns as **E** has rows. Note that in the product of **AB** of the matrices given above, $\mathbf{A}(3 \times 2)$ and $\mathbf{B}(2 \times 3)$ are conformable.

I is reserved for the *identity* matrix, which is *square*, *symmetric*, *diagonal* with $1's$ along the diagonal and $0's$ elsewhere:

$$\mathbf{I}_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Multiplication of any (conformable) matrix **M** by **I** gives **M**: $\mathbf{AI}_3 = \mathbf{A} = \mathbf{I}_2\mathbf{A}$

4. Inversion. The *inverse* $\mathbf{M}^{-1}$ of a *square* $(r \times r)$ matrix **M**, if it exists, satisfies $\mathbf{MM}^{-1} = \mathbf{I}_r$ (similar to the reciprocal of real number). A square matrix with an inverse is called *non-singular*.
   Inversion can be computationally challenging, but not for (2×2) case:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

   Find $\mathbf{C}^{-1}$.

$$\mathbf{C}^{-1} = \frac{1}{2} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$$

   The *rank* of a matrix is equal to the number of *linearly independent* rows or columns of the matrix. Vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are linearly independent if $\sum_i a_i \mathbf{x}_i = 0$ implies $a_1 = a_2 = \cdots = a_n = 0$.

   Matrix uses: model statements, systems of linear equations, covariance matrices of random vectors,....

   Consider two lines $y_1 = 5 - x, y_2 = 3 + x$. Do these lines intersect? Where? Write this system of two equations in two unknowns using a matrix:

$$\mathbf{C} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 5 \\ 3 \end{pmatrix}$$

   (left-multiply both sides by $\mathbf{C}^{-1}$)

$$\begin{pmatrix} x \\ y \end{pmatrix} = \mathbf{C}^{-1} \begin{pmatrix} 5 \\ 3 \end{pmatrix} = \begin{pmatrix} 1 \\ 4 \end{pmatrix}$$

   The lines intersect at the solution, $(x = 1, y = 4)$.

Matrices are cool and very useful!

If we have a random vector (just like a random variable but in vector form, i.e. components yield numeric answers that are random), call it $\mathbf{Y}$, and a constant vector, call it $\mathbf{a}$, then

$$E(\mathbf{a}'\mathbf{Y}) = \mathbf{a}'E(\mathbf{Y})$$

$$\text{Var}(\mathbf{a}'\mathbf{Y}) = \mathbf{a}'\text{Var}(\mathbf{Y})\mathbf{a}$$

Understanding matrices if very important as this is how we will look at our models for much of the rest of the class. Also, SAS and other statistical programs use matrices in their calculations and in their output.

Matrix formulation of MLR

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}$$

All of the response RVs are placed into the **response vector**:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

For observation i we can group all of the explanatory variables into a vector

$$\mathbf{x}_i = (1, x_{i1}, x_{i2}, x_{i3}, \ldots, x_{ip}).$$

The 1 in the first spot of the vector is for the intercept. If we 'stack' these row vectors on top of each other we can make a matrix called the **design matrix**:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$$

We also form a column vector corresponding to the regression parameters, called the **'beta vector'**:

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

and a column vector for the error terms, called the **error vector**:

$$\mathbf{E} = \begin{pmatrix} E_0 \\ E_1 \\ \vdots \\ E_n \end{pmatrix}$$

Now we can see that our MLR model (a system of n equations with $p+1$ unknowns)

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + E_1 \\ Y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + E_2 \\ \vdots &= \vdots \\ Y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} + E_n \end{aligned}$$

can be easily rewritten as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}$$

Our assumptions on the errors can now be specified as $\mathbf{E} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$ (multivariate normal distribution). $\sigma^2\mathbf{I}_n$ is called the variance-covariance matrix:

$$Var(\mathbf{E}) = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix}$$

The diagonals of the matrix give the variances for the E_i's $(Var(E_1), Var(E_2), \dots, Var(E_n))$ and the off-diagonals (say row i column j) give the covariances between E_i's and the E_j's $(Cov(E_i, E_j))$. As the off-diagonals are all 0, our errors are uncorrelated (which for the multivariate normal distribution implies independence).

Let's look at some of these quantities for our adsorption example. We have $n = 13$ and $p = 2$.

$$\mathbf{y} = \begin{pmatrix} 4 \\ 18 \\ 14 \\ 18 \\ 26 \\ 26 \\ 21 \\ 30 \\ 28 \\ 36 \\ 65 \\ 62 \\ 40 \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & 13 & 61 \\ 1 & 21 & 175 \\ 1 & 24 & 111 \\ 1 & 23 & 124 \\ 1 & 64 & 130 \\ 1 & 38 & 173 \\ 1 & 33 & 169 \\ 1 & 61 & 169 \\ 1 & 39 & 160 \\ 1 & 71 & 244 \\ 1 & 112 & 257 \\ 1 & 88 & 333 \\ 1 & 54 & 199 \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

The predicted values can be written as

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

and residuals as

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

- $\hat{\mathbf{y}}$ is called the vector of fitted or predicted values
- $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}$ is called the hat matrix as it 'places' the hat on $\mathbf{y}$
- $\mathbf{e}$ is the vector of residuals

We will still use least squares to select the parameters, which can be written as the minimum of:

$$SS(E) = \sum_{i=1}^{n}(obs_i - pred_i)^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2 = \mathbf{e}'\mathbf{e}$$

For the adsorption example, many of these matrices are given below:

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 13 & 641 & 2305 \\ 641 & 41831 & 133162 \\ 2305 & 133162 & 467669 \end{pmatrix} \quad (\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 0.633138 & 0.002477 & -0.003826 \\ 0.002477 & 0.000265 & -0.000088 \\ -0.003826 & -0.000088 & 0.000046 \end{pmatrix}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{pmatrix} -7.3507 \\ 0.3490 \\ 0.1127 \end{pmatrix} \quad \hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \begin{pmatrix} 4.0610 \\ 19.7008 \\ 13.5350 \\ 14.6511 \\ 29.6363 \\ 25.4084 \\ 23.2126 \\ 32.9846 \\ 24.2923 \\ 44.9271 \\ 60.7012 \\ 60.8904 \\ 33.9226 \end{pmatrix}$$

$$SS(E) = \mathbf{e}'\mathbf{e} = 191.7897 \qquad \hat{\sigma}^2 = MS(E) = SS(E)/(n-p-1) = 191.7897/10 = 19.17897$$

$$\hat{\boldsymbol{\Sigma}} = MS(E)(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 12.14294 & 0.04750 & -0.07337 \\ 0.04750 & 0.00508 & -0.00168 \\ -0.07337 & -0.00168 & 0.00088 \end{pmatrix}$$

The parameter estimates and the variance-covariance matrix are very useful for making inference about our intercept and partial slope parameters (done very similary to SLR). Let's use the above to find the following

1. What is the estimate for β_2? What is the interpretation?
2. What is the standard error of $\hat{\beta}_2$?
3. Conduct a test to determine if $\beta_2 = 0$ plausible (technically, after accounting for the linear association between extractable aluminum and adsorption index). Hint: $t(0.025, 10) = 2.228$
4. Estimate the mean adsorption index among the population of ALL soil with extractable aluminum = 100 and extractable iron = 150. Report a standard error for this estimate and a 95% confidence interval and a 95% prediction interval.

Recall that the overall hypotheses we want to test are

$$H_0 : \beta_1 = \beta_2 = 0 \quad vs \quad H_A : \text{ at least one is non-zero}$$

This is the test done in the ANOVA table given in the output from a MLR model. This is called the **global F-test** as it tests whether at least one of the terms in the model is important for predicting the response.

The ANOVA table for MLR follows the same ideas as in SLR. We are taking the total amount of variation in the response (SS(Tot)) and partitioning it into a part due to the model (SS(R)) and a part due to experimental error (SS(E)). In fact, the formulas for the sums of squares remain the same, only the degrees of freedom and the F-distribution used for finding the p-value change.

The full ANOVA table for MLR is given below:

Source	Sum of squares	df	Mean Square	F-Ratio
Regression	$SS(R)$	p	$MS(R)$	$MS(R)/MS(E)$
Error	$SS(E)$	$n-p-1$	$MS(E)$	
Total	$SS(Tot)$	$n-1$		

How to do MLR in SAS?

The following code will produce output appropriate for analysis:

```
proc reg data=adexp ;
model adsorp=aluminum iron/clb;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: adsorp

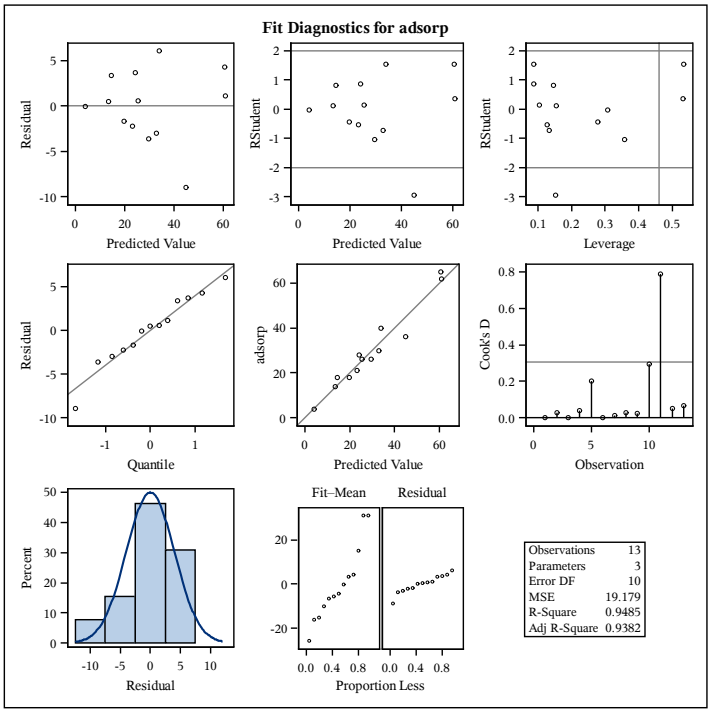
Number of Observations Read	14
Number of Observations Used	13
Number of Observations with Missing Values	1

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	3529.90308	1764.95154	92.03	<.0001
Error	10	191.78922	19.17892		
Corrected Total	12	3721.69231			

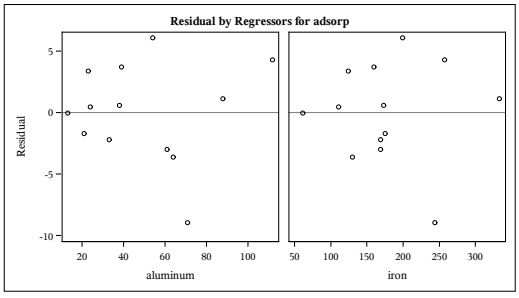
Root MSE	4.37937	**R-Square**	0.9485
Dependent Mean	29.84615	**Adj R-Sq**	0.9382
Coeff Var	14.67316		

Parameter Estimates							
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**95% Confidence Limits**	
Intercept	1	-7.35066	3.48467	-2.11	0.0611	-15.11498	0.41366
aluminum	1	0.34900	0.07131	4.89	0.0006	0.19012	0.50788
iron	1	0.11273	0.02969	3.80	0.0035	0.04658	0.17889

The REG Procedure
Model: MODEL1
Dependent Variable: adsorp



The REG Procedure
Model: MODEL1
Dependent Variable: adsorp



A non-additive model example:
A random sample of students taking the same exam:

IQ	Study TIME	GRADE
105	10	75
110	12	79
120	6	68
116	13	85
122	16	91
130	8	79
114	20	98
102	15	76

Consider regressing GRADE on IQ (X_1), TIME(X_2), and TI ($X_1 * X_2$), where TI = TIME*IQ. That is, we fit the model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + E$$

```
proc reg;
model Grade = IQ Time TI;
run;
```

```
                                The SAS System                                1
                              The REG Procedure

                           Analysis of Variance

                                    Sum of           Mean
Source                      DF     Squares         Square    F Value    Pr > F

Model                        3      610.81033      203.60344      26.22    0.0043
Error                        4       31.06467        7.76617
Corrected Total              7      641.87500

                           Parameter Estimates

                     Parameter     Standard
Variable     DF       Estimate        Error  t Value  Pr > |t|
Intercept     1      72.20608     54.07278     1.34    0.2527
IQ            1      -0.13117      0.45530    -0.29    0.7876
Time          1      -4.11107      4.52430    -0.91    0.4149
TI            1       0.05307      0.03858     1.38    0.2410
```

Discussion of the interaction model:
We call the product TI = Time*IQ an "interaction" term. That is, our explanatory variables do not have an independent effect on the response.

$$\widehat{MeanGrade} = 72.21 - 0.13 * IQ - 4.11 * Time + 0.0531 * TI$$

Now if IQ = 100 we get

$$\widehat{MeanGrade} = (72.21 - 13.1) + (-4.11 + 5.31) * Time$$

and if IQ 120 we get

$$\widehat{MeanGrade} = (72.21 - 15.7) + (-4.11 + 6.37) * Time.$$

Thus we expect an extra hour of study to increase the grade by 1.20 points for someone with IQ = 100 and by 2.26 points for someone with IQ = 120 if we use this interaction model.

Generally, we can interpret the (true) β parameters in the model as:

- β_0 - Average value of Grade when IQ and Study Time are 0
- β_1 - Average change in Grade for a unit increase in IQ when Study Time is 0
- β_2 - Average change in Grade for a unit increase in Study Time when IQ is 0
- β_3 - Average change in the slope for IQ (or Study Time) for a given value of Study Time (or IQ).

The interpretation of the interaction 'slope' can be seen by looking at the following:

$$\mu(x_1+1, x_2) - \mu(x_1, x_2) = \beta_0 + \beta_1(x_1+1) + \beta_2 x_2 + \beta_3(x_1+1)(x_2) - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \beta_3 x_1(x_2)$$

$$= \beta_1 + \beta_3 x_2$$

So β_3 is the amount the slope for x_1 changes per unit change in x_1 while x_2 is held constant.

Note: The global p-value is significant, but none of our individual terms are. This gives evidence that our model is over-fit. we may want to go back to the simpler "main effects" model.

Model Selection:

x_1, x_2, x_3 denote p independent variables. Consider several models:

1. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_1 x_1$
2. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_2 x_2$
3. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_3 x_3$
4. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
5. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_3 x_3$
6. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
7. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_2 x_2 + \beta_3 x_3$

A is nested in B means model A can be obtained by restricting (e.g. setting to 0) parameter values in model B.
True or false:

- Model 1 nested in Model 4
- Model 1 nested in Model 5
- Model 2 nested in Model 4
- Model 4 nested in Model 1
- Model 3 nested in Model 4
- Model 5 nested in Model 4
- Model 3 nested in Model 7
- Model 1 nested in Model 7

A nested in $B \longrightarrow A$ called *reduced model*, B called *full model*.

p - number of regression parameters in full model
q - number of regression parameters in reduced model
$p - q$ - number of regression parameters being tested.

Let's get a handle on this notation. Give the extra regression SS terms for comparing some of the nested models on preceding page:

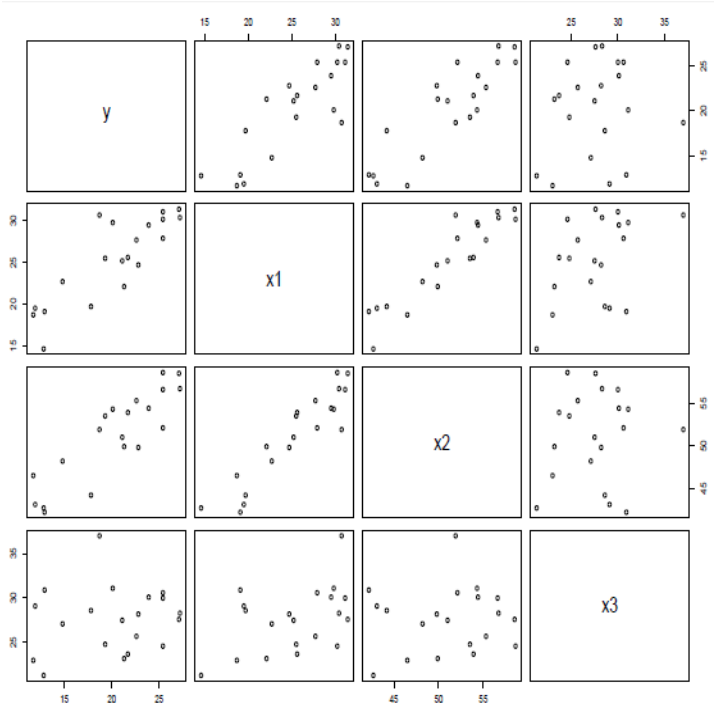
- Model 1 in model 4: $R(\beta_2, \beta_3|\beta_1)$
- Model 2 in model 4:
- Model 3 in model 4:
- Model 1 in model 5: $R(\beta_3|\beta_1)$
- Model 5 in model 4:

An example: How to measure body fat?

For each of $n = 20$ healthy individuals, the following measurements were made: bodyfat percentage y_i, triceps skinfold thickness, x_1, thigh circumference x_2, midarm circumference x_3.

x1	x2	x3	y
19.5	43.1	29.1	11.9
24.7	49.8	28.2	22.8
30.7	51.9	37.0	18.7
29.8	54.3	31.1	20.1
19.1	42.2	30.9	12.9
25.6	53.9	23.7	21.7
31.4	58.5	27.6	27.1
27.9	52.1	30.6	25.4
22.1	49.9	23.2	21.3
25.5	53.5	24.8	19.3
31.1	56.6	30.0	25.4
30.4	56.7	28.3	27.2
18.7	46.5	23.0	11.7
19.7	44.2	28.6	17.8
14.6	42.7	21.3	12.8
29.5	54.4	30.1	23.9
27.7	55.3	25.7	22.6
30.2	58.6	24.6	25.4
22.7	48.2	27.1	14.8
25.2	51.0	27.5	21.1

```
ods graphics on;
proc corr plots=matrix;
var y x1 x2 x3;
run;
```



Pearson Correlation Coefficients, N = 20
Prob > |r| under H0: Rho=0

	y	x1	x2	x3
y	1.00000	0.84327	0.87809	0.14244
		<.0001	<.0001	0.5491
x1	0.84327	1.00000	0.92384	0.45778
	<.0001		<.0001	0.0424
x2	0.87809	0.92384	1.00000	0.08467
	<.0001	<.0001		0.7227
x3	0.14244	0.45778	0.08467	1.00000
	0.5491	0.0424	0.7227	

Looking at the scatter plots and the correlation output, marginal associations between y and x_1 and between y and x_2 are highly significant, providing evidence of a strong $r \approx 0.85$ linear association between average bodyfat and triceps skinfold and between average bodyfat and thigh circumference.

Notice the scatter plot between x_1 and x_2, there is a strong linear relationship. This means that triceps skinfold and thigh circumference are giving some of the same information. This can lead to issues when fitting a model.

Multicollinearity: linear associations among the independent variables; causes problems such as inflated sampling variances for $\hat{\boldsymbol{\beta}}$.

```
proc reg data=bodyfat;
   model y=x1/covb;
   model y=x2/covb;
   model y=x3/covb;
   model y=x1 x2/covb;
   model y=x1 x2 x3/covb;
run;
```

Yields the following output:

Output From Proc Reg for Bodyfat Example

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	21
Number of Observations Used	20
Number of Observations with Missing Values	1

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	352.26980	352.26980	44.30	<.0001
Error	18	143.11970	7.95109		
Corrected Total	19	495.38950			

Root MSE	2.81977	R-Square	0.7111
Dependent Mean	20.19500	Adj R-Sq	0.6950
Coeff Var	13.96271		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1.49610	3.31923	-0.45	0.6576
x1	1	0.85719	0.12878	6.66	<.0001

Covariance of Estimates		
Variable	Intercept	x1
Intercept	11.01731839	-0.419670565
x1	-0.419670565	0.0165844918

Output From Proc Reg for Bodyfat Example

The REG Procedure
Model: MODEL2
Dependent Variable: y

Number of Observations Read	21
Number of Observations Used	20
Number of Observations with Missing Values	1

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	381.96582	381.96582	60.62	<.0001
Error	18	113.42368	6.30132		
Corrected Total	19	495.38950			

Root MSE	2.51024	R-Square	0.7710
Dependent Mean	20.19500	Adj R-Sq	0.7583
Coeff Var	12.43002		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-23.63449	5.65741	-4.18	0.0006
x2	1	0.85655	0.11002	7.79	<.0001

Covariance of Estimates		
Variable	Intercept	x2
Intercept	32.006329324	-0.619332881
x2	-0.619332881	0.0121034372

Output From Proc Reg for Bodyfat Example

The REG Procedure
Model: MODEL3
Dependent Variable: y

Number of Observations Read	21
Number of Observations Used	20
Number of Observations with Missing Values	1

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	10.05160	10.05160	0.37	0.5491
Error	18	485.33790	26.96322		
Corrected Total	19	495.38950			

Root MSE	5.19261	R-Square	0.0203
Dependent Mean	20.19500	Adj R-Sq	-0.0341
Coeff Var	25.71236		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	14.68678	9.09593	1.61	0.1238
x3	1	0.19943	0.32663	0.61	0.5491

Covariance of Estimates		
Variable	Intercept	x3
Intercept	82.735867956	-2.946694682
x3	-2.946694682	0.1066869907

Output From Proc Reg for Bodyfat Example

The REG Procedure
Model: MODEL4
Dependent Variable: y

Number of Observations Read	21
Number of Observations Used	20
Number of Observations with Missing Values	1

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	385.43871	192.71935	29.80	<.0001
Error	17	109.95079	6.46769		
Corrected Total	19	495.38950			

Root MSE	2.54317	R-Square	0.7781
Dependent Mean	20.19500	Adj R-Sq	0.7519
Coeff Var	12.59305		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-19.17425	8.36064	-2.29	0.0348
x1	1	0.22235	0.30344	0.73	0.4737
x2	1	0.65942	0.29119	2.26	0.0369

Covariance of Estimates			
Variable	Intercept	x1	x2
Intercept	69.900312587	1.8469661215	-2.273097628
x1	1.8469661215	0.0920751757	-0.081628463
x2	-2.273097628	-0.081628463	0.0847900309

The REG Procedure
Model: MODEL5
Dependent Variable: y

Number of Observations Read	21
Number of Observations Used	20
Number of Observations with Missing Values	1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	396.98461	132.32820	21.52	<.0001
Error	16	98.40489	6.15031		
Corrected Total	19	495.38950			

Root MSE	2.47998	R-Square	0.8014
Dependent Mean	20.19500	Adj R-Sq	0.7641
Coeff Var	12.28017		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	117.08469	99.78240	1.17	0.2578
x1	1	4.33409	3.01551	1.44	0.1699
x2	1	-2.85685	2.58202	-1.11	0.2849
x3	1	-2.18606	1.59550	-1.37	0.1896

Covariance of Estimates

Variable	Intercept	x1	x2	x3
Intercept	9956.5279384	300.1979628	-257.3823153	-158.6704127
x1	300.1979628	9.0933087788	-7.779145105	-4.7880263
x2	-257.3823153	-7.779145105	6.6668028532	4.0946155019
x3	-158.6704127	-4.7880263	4.0946155019	2.545617053

Question: Why is the global p-value in the last model significant, i.e. at least one predictor is useful, but the individual tests are all nonsignificant?

In the bodyfat data, consider comparing the simple model that Y depends only on x_1 (triceps) versus the full model that it depends on all three.

$$\begin{aligned} \text{Model } A: \mu(x_1, x_2, x_3) &= \beta_0 + \beta_1 x_1 \\ \text{Model } B: \mu(x_1, x_2, x_3) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \end{aligned}$$

or the null hypothesis

$$H_0: \beta_2 = \beta_3 = 0 \quad \text{vs} \quad H_1: \beta_2, \beta_3 \text{ not both } 0$$

after accounting for x_1. Our F statistic can be used

$$F = \frac{(396.9 - 352.3)/2}{6.15} = \frac{22.3}{6.15} = 3.64$$

How many *df* for numerator and denominator?

The 95^{th} percentile is $F(0.05, \quad , \quad) = 3.63$.

Our conclusion about the hypotheses?

That is, after accounting for the linear dependence between triceps and bodyfat, there is still some linear association between mean bodyfat and at least one of x_2, x_3 (thigh,midarm).

To get the nested model F-ratio in SAS:

```
proc reg data=bodyfat;
     model y=x1 x2 x3;
     test x2=0,x3=0;
run;
```

Full mode vs only Triceps 4

The REG Procedure
Model: MODEL1

Test 1 Results for Dependent Variable y

Source	DF	Mean Square	F Value	Pr > F
Numerator	2	22.35741	3.64	0.0500
Denominator	16	6.15031		

However, we saw in the previous output that a model with all three variables is no good. This is due to the multicollinearity. We will now very briefly look at a few automated model selection techniques.

Using proc reg to perform variable selection:
We'll discuss three hypothesis testing methods for selecting variables (there are many other ways to accomplish this we won't discuss).

1. **Forward Selection** - Start with nothing and work forward.
   (a) Begin with a model with only β_0
   (b) Calculate $R(\beta_i|\beta_0)$ for all possible predictors and find p-values for each
   (c) Take most significant p-value less than a cutoff (say 0.3), add predictor into model.
   (d) Say β_j was added in the last step, repeat above process with added predictor. That is, calculate $R(\beta_i|\beta_0, \beta_j)$ for all other predictors, etc.
   (e) Stop when no predictors are below the cutoff or if the full model is selected.
2. **Backward Selection** - Start with everything and work backward.
   (a) Start with full model.
   (b) Locate variable with largest p-value greater than a cutoff (say 0.1), remove that variable.
   (c) Repeat until all p-values are less than the cut off or the null model (intercept only model) is chosen.
3. **Subset Selection** - Compute all possible models, pick best.
   (a) Compare each of the models using a criterion.
   (b) Choose model that minimizes that criterion. Possible criteria include:
      - *Adjusted* $R^2 = 1 - \frac{n-1}{n-p-1}(1 - R^2)$ (takes into account the addition of more predictors)
      - Mallow's C_P, AIC, AICc, or BIC (all take into account the model complexity, not just how well the model fits the data)

How to do these model selection methods in SAS?

```
proc reg data=bodyfat plots=none;
    model y=x1 x2 x3/selection=cp ;
    model y=x1 x2 x3/selection=forward SLentry=0.3;
    model y=x1 x2 x3/selection=backward SLstay=0.1;
   model y=x1 x2 x3/selection=adjrsq;
run;
```

Variable Selection Methods on Bodyfat Example

The REG Procedure
Model: MODEL1
Dependent Variable: y

C(p) Selection Method

Number of Observations Read	21
Number of Observations Used	20
Number of Observations with Missing Values	1

Number in Model	C(p)	R-Square	Variables in Model
1	2.4420	0.7710	x2
2	3.2242	0.7862	x1 x3
2	3.8773	0.7781	x1 x2
3	4.0000	0.8014	x1 x2 x3
2	4.0657	0.7757	x2 x3
1	7.2703	0.7111	x1
1	62.9128	0.0203	x3

Variable Selection Methods on Bodyfat Example

The REG Procedure
Model: MODEL2
Dependent Variable: y

Number of Observations Read	21
Number of Observations Used	20
Number of Observations with Missing Values	1

Forward Selection: Step 1

Variable x2 Entered: R-Square = 0.7710 and C(p) = 2.4420

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	381.96582	381.96582	60.62	<.0001
Error	18	113.42368	6.30132		
Corrected Total	19	495.38950			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-23.63449	5.65741	109.97344	17.45	0.0006
x2	0.85655	0.11002	381.96582	60.62	<.0001

Bounds on condition number: 1, 1

No other variable met the 0.3000 significance level for entry into the model.

Summary of Forward Selection							
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x2	1	0.7710	0.7710	2.4420	60.62	<.0001

Variable Selection Methods on Bodyfat Example

The REG Procedure
Model: MODEL3
Dependent Variable: y

Number of Observations Read	21
Number of Observations Used	20
Number of Observations with Missing Values	1

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.8014 and C(p) = 4.0000

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	396.98461	132.32820	21.52	<.0001
Error	16	98.40489	6.15031		
Corrected Total	19	495.38950			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	117.08469	99.78240	8.46816	1.38	0.2578
x1	4.33409	3.01551	12.70489	2.07	0.1699
x2	-2.85685	2.58202	7.52928	1.22	0.2849
x3	-2.18606	1.59550	11.54590	1.88	0.1896

Bounds on condition number: 708.84, 4133.4

Backward Elimination: Step 1

Variable x2 Removed: R-Square = 0.7862 and C(p) = 3.2242

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	389.45533	194.72767	31.25	<.0001
Error	17	105.93417	6.23142		
Corrected Total	19	495.38950			

Variable Selection Methods on Bodyfat Example

The REG Procedure
Model: MODEL3
Dependent Variable: y

Backward Elimination: Step 1

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	6.79163	4.48829	14.26834	2.29	0.1486
x1	1.00058	0.12823	379.40373	60.89	<.0001
x3	-0.43144	0.17662	37.18554	5.97	0.0258

Bounds on condition number: 1.2651, 5.0605

All variables left in the model are significant at the 0.1000 level.

Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x2	2	0.0152	0.7862	3.2242	1.22	0.2849

The REG Procedure
Model: MODEL4
Dependent Variable: y

Adjusted R-Square Selection Method

Number of Observations Read	21
Number of Observations Used	20
Number of Observations with Missing Values	1

Number in Model	Adjusted R-Square	R-Square	Variables in Model
3	0.7641	0.8014	x1 x2 x3
2	0.7610	0.7862	x1 x3
1	0.7583	0.7710	x2
2	0.7519	0.7781	x1 x2
2	0.7493	0.7757	x2 x3
1	0.6950	0.7111	x1
1	-.0341	0.0203	x3

Types of Sums of Squares

Given that we have 4 predictors, $X_1 - X_4$ we really can have a number of tests based on nested models for $\beta_4 = 0$ (or for any other β for that matter). Let's write them down in terms of extra regression sums of squares:
$R(\beta_4|\beta_0)$ (SLR test)
$R(\beta_4|\beta_0, \beta_1)$ (test after accounting for X_1)
$R(\beta_4|\beta_0, \beta_2)$ (test after accounting for X_2)
$R(\beta_4|\beta_0, \beta_3)$ (test after accounting for X_3)
$R(\beta_4|\beta_0, \beta_1, \beta_2)$ (test after accounting for X_1 and X_2)
$R(\beta_4|\beta_0, \beta_1, \beta_3)$ (test after accounting for X_1 and X_3)
$R(\beta_4|\beta_0, \beta_2, \beta_3)$ (test after accounting for X_2 and X_3)
$R(\beta_4|\beta_0, \beta_1, \beta_2, \beta_3)$ (test after accounting for X_1, X_2, and X_3)

Some of these tests can be easily found using different types of sums of squares.

The tests given for the parameter estimates are all type III tests and this is the test usually done to determine if a slope term has significance. However, type I tests are very useful for model building. For example, if we wanted to look at building a model for the bodyfat example and we thought the order of importance for the variables was X_1 (triceps), X_3 (midarm), and X_2 (thigh), we could get sequential tests for these models using type I sums of squares.

In SAS proc reg use the following code:

```
proc reg data=bodyfat;
   model y=x1 x3 x2/ss1; *Note the order of variables is important for Type I;
run;
```

Sequential tests for bodyfat example **1**

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	21
Number of Observations Used	20
Number of Observations with Missing Values	1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	396.98461	132.32820	21.52	<.0001
Error	16	98.40489	6.15031		
Corrected Total	19	495.38950			

Root MSE	2.47998	**R-Square**	0.8014
Dependent Mean	20.19500	**Adj R-Sq**	0.7641
Coeff Var	12.28017		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS
Intercept	1	117.08469	99.78240	1.17	0.2578	8156.76050
x1	1	4.33409	3.01551	1.44	0.1699	352.26980
x3	1	-2.18606	1.59550	-1.37	0.1896	37.18554
x2	1	-2.85685	2.58202	-1.11	0.2849	7.52928

Let‘s label the Type I SS in terms of extra regression sums of squares (R notation).

Note: we will soon use proc glm for our model analysis and this gives even better output for type I sums of squares. (The tests given for type I sums of squares use the *full model* MS(E) rather than the full model MS(E) up to that point. This test still works because MS(E) from each model is an unbiased estimate of σ^2. The tests using the different MS(E) terms could give different results, but will usually agree.

```
proc glm data=bodyfat;
   model y=x1 x3 x2;
run;
```

Sequential tests for bodyfat example using GLM 2

The GLM Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	396.9846118	132.3282039	21.52	<.0001
Error	16	98.4048882	6.1503055		
Corrected Total	19	495.3895000			

R-Square	Coeff Var	Root MSE	y Mean
0.801359	12.28017	2.479981	20.19500

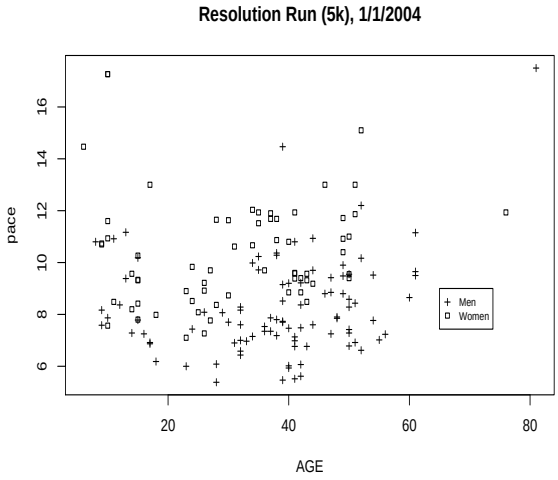
Source	DF	Type I SS	Mean Square	F Value	Pr > F
x1	1	352.2697968	352.2697968	57.28	<.0001
x3	1	37.1855371	37.1855371	6.05	0.0257
x2	1	7.5292779	7.5292779	1.22	0.2849

Source	DF	Type III SS	Mean Square	F Value	Pr > F
x1	1	12.70489278	12.70489278	2.07	0.1699
x3	1	11.54590217	11.54590217	1.88	0.1896
x2	1	7.52927788	7.52927788	1.22	0.2849

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	117.0846948	99.78240295	1.17	0.2578
x1	4.3340920	3.01551136	1.44	0.1699
x3	-2.1860603	1.59549900	-1.37	0.1896
x2	-2.8568479	2.58201527	-1.11	0.2849

A linear regression example with a quadratic explanatory variable:
Data was collected on 5 kilometer run times. The variables collected were age, sex, and pace.

	Obs	age	sex	race	pace
	1	28	M	16.6833	5.38333
	2	39	M	16.9500	5.46667
	3	41	M	17.1333	5.51667
	4	42	M	17.4000	5.61667
(abbreviated)	...	...	...	...	...
	157	52	F	46.8833	15.1000
	158	10	F	53.6000	17.2667
	159	10	F	53.6167	17.2667
	160	81	M	54.3167	17.5000



Quadratic model for pace (Y) as a function of age (x):

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + E_i \quad \text{for } i = 1, \ldots, 160$$

where $E_i \overset{iid}{\sim} N(0, \sigma^2)$.

Question: What does σ^2 represent in the model?

Question: What do the parameters mean, i.e. what is their interpretation?

We may want to compare this model with a SLR model

$$Y_i = \beta_0 + \beta_1 x_i + E_i \text{ for } i = 1, \ldots, 160$$

Question: How can we compare the two models?

```
/* age2 defined in data step as age*age */
PROC REG;
   MODEL pace=age;
   MODEL pace=age age2/ss1 covb;
RUN;
```

Model: MODEL1
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.09650	1.09650	0.22	0.6396
Error	158	786.99821	4.98100		
Corrected Total	159	788.09472			

Root MSE	2.23182	R-Square	0.0014
Dependent Mean	9.12063	Adj R-Sq	-0.0049

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	8.92271	0.45724	19.51	<.0001
age	1	0.00564	0.01203	0.47	0.6396

Model: MODEL2
Analysis of Variance

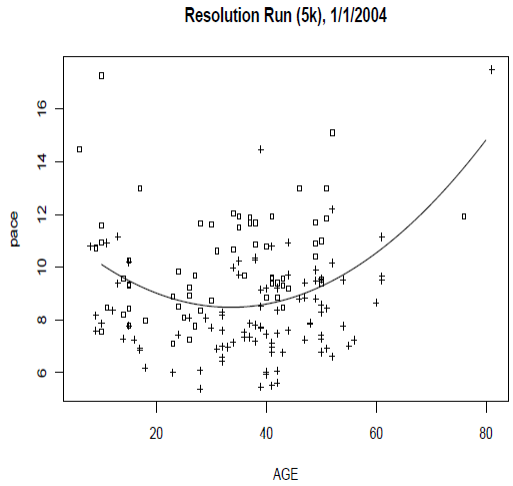
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	113.64500	56.82250	13.23	<.0001
Error	157	674.44972	4.29586		
Corrected Total	159	788.09472			

Root MSE	2.07265	R-Square	0.1442
Dependent Mean	9.12063	Adj R-Sq	0.1333

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS
Intercept	1	11.78503	0.70216	16.78	<.0001	13310
age	1	-0.19699	0.04113	-4.79	<.0001	1.09650
age2	1	0.00294	0.00057380	5.12	<.0001	112.54850

Covariance of Estimates

Variable	Intercept	age	age2
Intercept	0.4930258	-0.0265145	0.0003209
age	-0.0265145	0.0016921	-0.0000227
age2	0.0003209	-0.0000227	0.0000003



Fitted models are:

$$\text{Model 1: } \hat{\mu}(x) = 8.923 + 0.0056age$$

$$\text{Model 2: } \hat{\mu}(age) = 11.785 - 0.197age + 0.00294age^2$$

$$
\begin{aligned}
F &= \frac{R(\beta_2|\beta_0,\beta_1)}{MS(E)_{full}} \\
&= \frac{(SS(R)_{full} - SS(R)_{red})/1}{MS(E)_{full}} \\
&= \frac{(113.6 - 1.1)/1}{4.3} \\
&= \frac{(SS(E)_{red} - SS(E)_{full})/1}{MS(E)_{full}} \\
&= \frac{(787.0 - 674.4)/1}{4.3} = 26.2 \\
&= \left(\frac{\hat{\beta}_2}{SE}\right)^2 = (5.12)^2
\end{aligned}
$$

with $F(0.05, 1, 157) = 3.90$. Since $26.2 >> 3.9$, we reject that the linear model is appropriate when compared to the quadratic model. This is the same test as the t-test for age2!

Chapter 6

ST 512 - Extra Correlation and Regression Questions

Problems with worked out solutions.

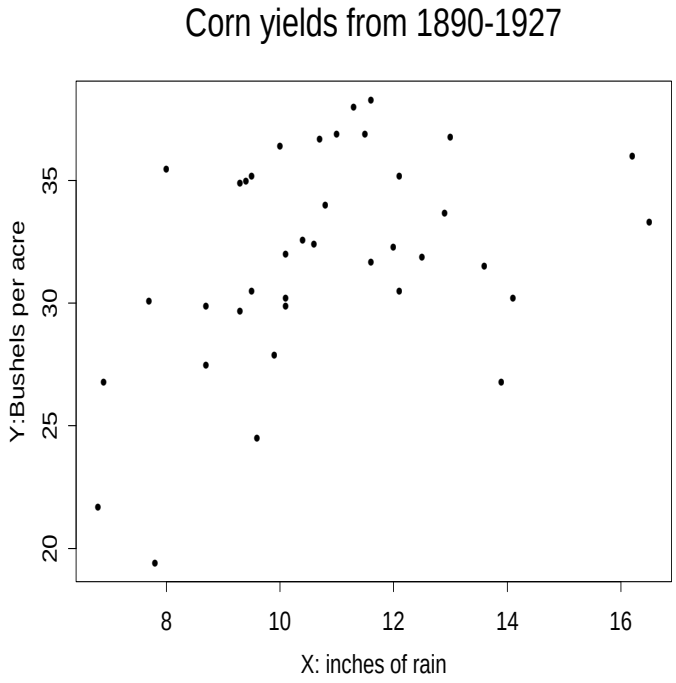
An example: The association between corn yield and rainfall:

Yields y (in bushels/acre) on corn raised in six midwestern states from 1890 to 1927 recorded with rainfall x (inches/yr).

$y_1, \ldots, y_{38}$ and $x_1, \ldots, x_{38}$.

Year	1890	1891	1892	1893	1894	1895	1896	1897	1898	1899
Yield	24.5	33.7	27.9	27.5	21.7	31.9	36.8	29.9	30.2	32
Rainfall	9.6	12.9	9.9	8.7	6.8	12.5	13	10.1	10.1	10.1
Year	1900	1901	1902	1903	1904	1905	1906	1907	1908	1909
Yield	34	19.4	36	30.2	32.4	36.4	36.9	31.5	30.5	32.3
Rainfall	10.8	7.8	16.2	14.1	10.6	10	11.5	13.6	12.1	12
Year	1910	1911	1912	1913	1914	1915	1916	1917	1918	1919
Yield	34.9	30.1	36.9	26.8	30.5	33.3	29.7	35	29.9	35.2
Rainfall	9.3	7.7	11	6.9	9.5	16.5	9.3	9.4	8.7	9.5
Year	1920	1921	1922	1923	1924	1925	1926	1927		
Yield	38.3	35.2	35.5	36.7	26.8	38	31.7	32.6		
Rainfall	11.6	12.1	8	10.7	13.9	11.3	11.6	10.4		

A *scatter plot* provides a visual for inspecting the association between Y and X.



From the scatter plot, the form of the association appears to be linear or slightly quadratic, the strength is weak to moderate, and the direction is positive.

A correlation analysis was done and a SLR model was fit using SAS yielding the following (partial) output:

```
proc corr data=corn cov;
var yield rain;
run;
```

```
proc reg data=corn;
model yield=rain;
run;
```

2 Variables:	yield rain

Covariance Matrix, DF = 37		
	yield	rain
yield	19.04190612	3.98025605
rain	3.98025605	5.13217639

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
yield	38	31.91579	4.36370	1213	19.40000	38.30000
rain	38	10.78421	2.26543	409.80000	6.80000	16.50000

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	114.21474	114.21474	6.97	0.0122
Error	36	590.33578	16.39822		
Corrected Total	37	704.55053			

Use the output to

1. Find the value of the correlation coefficient.
2. Find the 95% confidence interval for the population correlation, ρ. Interpret the interval you found.
3. Without conducting a hypothesis test, use the confidence interval to make a conclusion about the hypotheses $H_0 : \rho = 0$ vs $H_A : \rho \neq 0$.
4. Find the fitted line for the SLR with yield as the response and rainfall as the predictor.
5. Use the summary statistics to create 95% confidence intervals for β_1 and β_0. Note: $t(38, 0.025) = 2.024$.
6. Find a 95% confidence interval for the true mean yield for corn from the six states for a rainfall of 14inches.
7. Suppose the rainfall collected for a year was 14 inches but the yield of the corn from the six states was not recorded. Find a 95% prediction interval for the yield of the corn from that year.

Solutions:

1. From the output we have

$$\bar{x} = 10.78, \quad s_X^2 = 5.13 \quad s_X = 2.27$$
$$\bar{y} = 31.92, \quad s_Y^2 = 19.04 \quad s_Y = 4.36$$
$$s_{XY} = 3.98$$

   Applying the formula for r, we get

$$r = \frac{s_{XY}}{s_X s_Y} = \frac{3.98}{\sqrt{5.13 \times 19.04}} = 0.40$$

2. With $r = 0.40$, $n = 38$, and $z_{\alpha/2} = 1.96$, a 95% interval is given by

$$\left(\frac{\frac{1+0.4}{1-0.4} e^{-2*1.96/\sqrt{38-3}} - 1}{\frac{1+0.4}{1-0.4} e^{-2*1.96/\sqrt{38-3}} + 1}, \frac{\frac{1+0.4}{1-0.4} e^{2*1.96/\sqrt{38-3}} - 1}{\frac{1+0.4}{1-0.4} e^{2*1.96/\sqrt{38-3}} + 1} \right) = (0.09, 0.64).$$

   We are 95% confident that the true correlation between corn yield and rainfall is between 0.09 and 0.64.

3. Since there is a one-to-one correspondence between a two-sided HT at the 0.05 level and a $100(1-\alpha)\%$ CI, we would reject $H_0 : \rho = 0$ as 0 is not in the interval.

4. To find the fitted least squares line:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{s_{xy}}{s_x^2} \\
&= \frac{3.98}{5.13} = 0.776(\text{ bushels per acre } \div \text{ inches per year}) \\
or &= r_{xy}\frac{s_y}{s_x} \\
&= (0.40)\sqrt{\frac{19.04}{5.13}} \\
&= 0.771(offduetorounding) \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1\bar{x} \\
&= 31.92 - 0.776(10.78) \\
&= 23.555 \text{ bushels per acre}
\end{aligned}$$

yielding the least squares line of

$$\hat{y} = 23.555 + (0.776)x.$$

5. To find confidence intervals for the regression parameters:
For β_1, note that

$$S_{xx} = (n-1)s_x^2 = 5.13(38-1) = 189.81$$

and we can estimate σ^2 using the $MS(E) = 16.40$. Thus, a 95% CI for β_1 is given by

$$0.776 \pm 2.024\sqrt{\frac{16.40}{189.81}} = (0.181, 1.371)$$

We are 95% confident the true value of the slope lies in this interval. For $\hat{\beta}_0$,

$$23.555 \pm 2.024\sqrt{16.40\left(\frac{1}{38} + \frac{(10.78)^2}{189.81}\right)} = (17.052, 30.058)$$

6. First we can find the point estimate

$$\hat{\mu}(14) = 23.555 + 0.776 * 14 = 34.419$$

The standard error of this mean estimate is

$$\sqrt{16.40\left(\frac{1}{38} + \frac{(14-10.78)^2}{189.81}\right)} = 1.125$$

Thus, a 95% CI for the mean corn yield for 14 inches of rain is

$$34.419 \pm 2.024 * 1.125 = (32.142, 36.696)$$

We are 95% confident that the true mean corn yield for all years with 14 inches of rain is between 32.142 adn 36.696 bushels per acre.

7. A 95% prediction interval is then

$$34.419 \pm 2.024\sqrt{16.40\left(1+\frac{1}{38}+\frac{(14-10.78)^2}{189.81}\right)} = (25.898, 42.940)$$

We are 95% confident that a year that has 14 inches of rain will have a yield between 2.898 and 42.940 bushels per acre.

Looking at the scatter plot, there may be a quadratic relationship. We can fit a linear regression model with rainfall and rainfall squared as predictors to investigate this. Use the following SAS output to conduct a LOF test for the SLR model.

```
data corn;
set corn;
rain2=rain*rain;
run;
```

```
proc reg data=corn;
model yield= rain rain2/ss1;
run;
```

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	209.02175	104.51087	7.38	0.0021
Error	35	495.52878	14.15797		
Corrected Total	37	704.55053			

Root MSE	3.76271	R-Square	0.2967
Dependent Mean	31.91579	Adj R-Sq	0.2565
Coeff Var	11.78948		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS
Intercept	1	-5.01467	11.44158	-0.44	0.6639	38707
rain	1	6.00428	2.03895	2.94	0.0057	114.21474
rain2	1	-0.22936	0.08864	-2.59	0.0140	94.80700

The F statistic for the LOF test can be written as

$$F = \frac{\frac{SS(R)_f - SS(R)_r}{p-q}}{MS(E)_f} = \frac{\frac{209.02-114.21}{2-1}}{14.16} = 6.70$$

We compare this to the critical value from the appropriate F distribution: $F(1, 35, 0.05) = 4.12$

Therefore, we reject $H_0 : \beta_2 = 0$ in favor of $H_A : \beta_2 \neq 0$.

Note: This test statistic of 6.70 is equivalent to the t-test squared $(-2.59)^2$ since we only have 1 numerator degree of freedom.

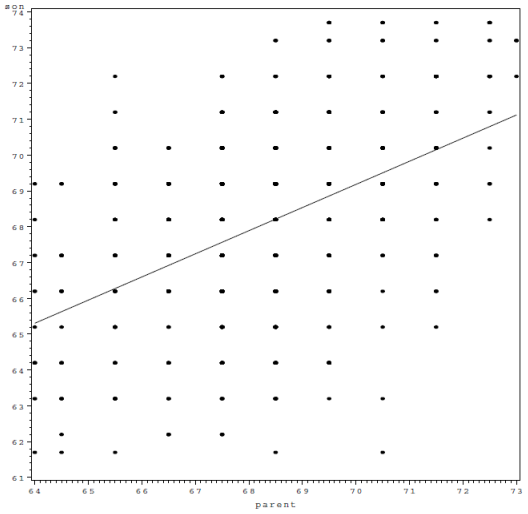
A couple random questions:

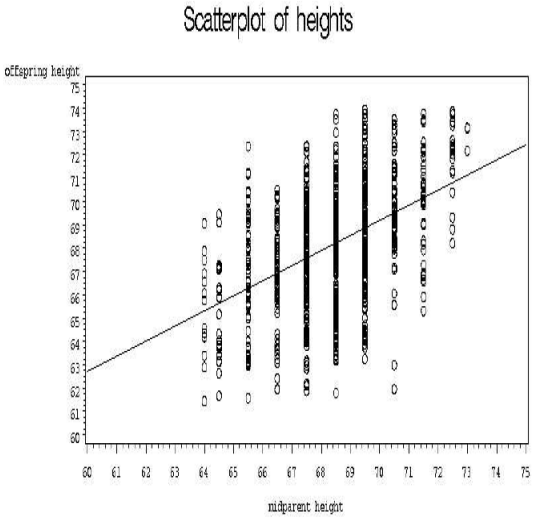
1. What type of data is needed to run a MLR model?
2. An industrial quality control expert takes 200 hourly measurements on an industrial furnace which is under control and finds that a 95% confidence interval for the mean temperature is (500.35, 531.36). As a result he tells management that the process should be declared out of control whenever hourly measurements fall outside this interval and, of course, is later fired for incompetence. (Why and what should he have done?)

Solutions:

1. The data situation needed for a MLR model is p quantitative predictors and 1 quantitative response measured on the same individuals (i.e. units).
2. The interval found is for the **mean** temperature. The interval is not trying to capture a single new observation. The interval that should have been found is a prediction interval. This interval will be much wider as it is much more difficult to predict a new value as opposed to predicting the true mean.

The classical regression example - The association between height of adults and their parents






```
/* ------------------------------------------------------------------
|  Stigler , History of Statistics pg. 285 gives Galton's famous data |
| on heights of sons (columns,Y) and average parents' height (rows,X) |
| scaled to represent a male height (essentially sons' heights versus |
| fathers' heights).   Taken from Dickey's website.                   |
\ ------------------------------------------------------------------*/
```

	61.7	62.2	63.2	64.2	65.2	66.2	67.2	68.2	69.2	70.2	71.2	72.2	73.2	73.7
73.0	0	0	0	0	0	0	0	0	0	0	1	3	0	
72.5	0	0	0	0	0	0	0	1	2	1	2	7	2	4
71.5	0	0	0	0	1	3	4	3	5	10	4	9	2	2
70.5	1	0	1	0	1	1	3	12	18	14	7	4	3	3
69.5	0	0	1	16	4	17	27	20	33	25	20	11	4	5
68.5	1	0	7	11	16	25	31	34	48	21	18	4	3	0
67.5	0	3	5	14	15	36	38	28	38	19	11	4	0	0
66.5	0	3	3	5	2	17	17	14	13	4	0	0	0	0
65.5	1	0	9	5	7	11	11	7	7	5	2	1	0	0
64.5	1	1	4	4	1	5	5	0	2	0	0	0	0	0
64.0	1	0	2	4	1	2	2	1	1	0	0	0	0	0

Many questions we could answer using this dataset: Note: $t(927, 0.025) \approx t(926, 0.025) = 1.963$

1. Suppose we ignore midparent height x. Consider estimating the mean $\mu_Y = E(Y)$. Recall a method for obtaining a confidence interval for the mean height of the sons in the population from which these data were randomly sampled. Use summary statistics in the output that follows to complete this naive analysis.

For the rest of the problems, consider a linear regression between the sons' heights and the midparent height. Let $Y_1, \ldots, Y_n$ denote the sons' heights. Given their average parent height, $X = x_i$,

$$Y_i = \beta_0 + \beta_1 x_i + E_i \quad \text{for } i = 1, \ldots, n (n = 928).$$

where $E_1, \ldots, E_n$ are *iid* Normal with mean 0 and variance σ^2.

Output is given on the following pages, use it to answer the following:

2. What is the meaning, in words, of β_1?
3. True/false: (a) β_1 is a statistic (b) β_1 is a parameter (c) β_1 is unknown.
4. What is the observed value of $\hat{\beta}_1$?
5. True/false: (a) $\hat{\beta}_1$ is a statistic (b) $\hat{\beta}_1$ is a parameter (c) $\hat{\beta}_1$ is unknown.
6. Is $\hat{\beta}_1 = \beta_1$?
7. How much does $\hat{\beta}_1$ vary about β_1 from sample to sample? (Provide an estimate of the standard error, as well as an expression indicating how it was computed.)
8. What is a region of plausible values for β_1 suggested by the data? (i.e. a CI)
9. What is the line that best fits these data, using the criterion that smallest sum of squared residuals is "best?"
10. How much of the observed variation in the heights of sons (the y-axis) is explained by this "best" line?
11. Give an expression in terms of the parameters of the model for the true average height of sons with midparent height $x = 68$.
12. What is the estimated average height of sons whose midparent height is $x = 68$?
13. Is this the true average height in the whole population of sons whose midparent height is $x = 68$?
14. What is the estimated standard deviation among the population of sons whose parents have midparent height $x = 68$?
15. What is the estimated standard deviation among the population of sons whose parents have midparent height $x = 72$? Bigger, smaller, or the same as that for $x = 68$?
16. What is the estimated standard error of the estimated average height for sons with midparent height $x = 68$, i.e. $\hat{\mu}(68) = \hat{\beta}_0 + 68\hat{\beta}_1$? Provide an expression for this standard error.
17. Is the estimated standard error of $\hat{\mu}(72)$ bigger, smaller, or the same as that for $\hat{\mu}(68)$?

18. What quantity can you use to describe or characterize the linear association between height and midparent height in the whole population? Is this a parameter or a statistic?

19. Is the observed linear association between son's height and midparent height strong? To answer, report the value of r and the p-value from the appropriate test.

20. Define $\mu_Y, \sigma_Y, \mu_X, \sigma_X, \rho$. Parameters or statistics?

21. What are plausible values for ρ suggested by the data? (i.e. form a CI)

22. Is $\boxed{E_1, \ldots, E_{928} \overset{iid}{\sim} N(0, \sigma^2)}$ a reasonable assumption?

23. Based **solely** on this study, can we conclude that larger parents cause larger sons?

Code that produced this output available on web.

The CORR Procedure

2 Variables:	son	parent

Covariance Matrix, DF = 927		
	son	**parent**
son	6.340028724	2.064614487
parent	2.064614487	3.194560689

Simple Statistics						
Variable	**N**	**Mean**	**Std Dev**	**Sum**	**Minimum**	**Maximum**
son	928	68.08847	2.51794	63186	61.70000	73.70000
parent	928	68.30819	1.78733	63390	64.00000	73.00000

Pearson Correlation Coefficients, N = 928 Prob > |r| under H0: Rho=0		
	son	**parent**
son	1.00000	0.45876 <.0001
parent	0.45876 <.0001	1.00000

Pearson Correlation Statistics (Fisher's z Transformation)							
Variable	**With Variable**	**N**	**Sample Correlation**	**Fisher's z**	**95% Confidence Limits**		**p Value for H0:Rho=0**
son	**parent**	928	0.45876	0.49574	0.406407	0.508115	<.0001

The REG Procedure
Model: MODEL1
Dependent Variable: son

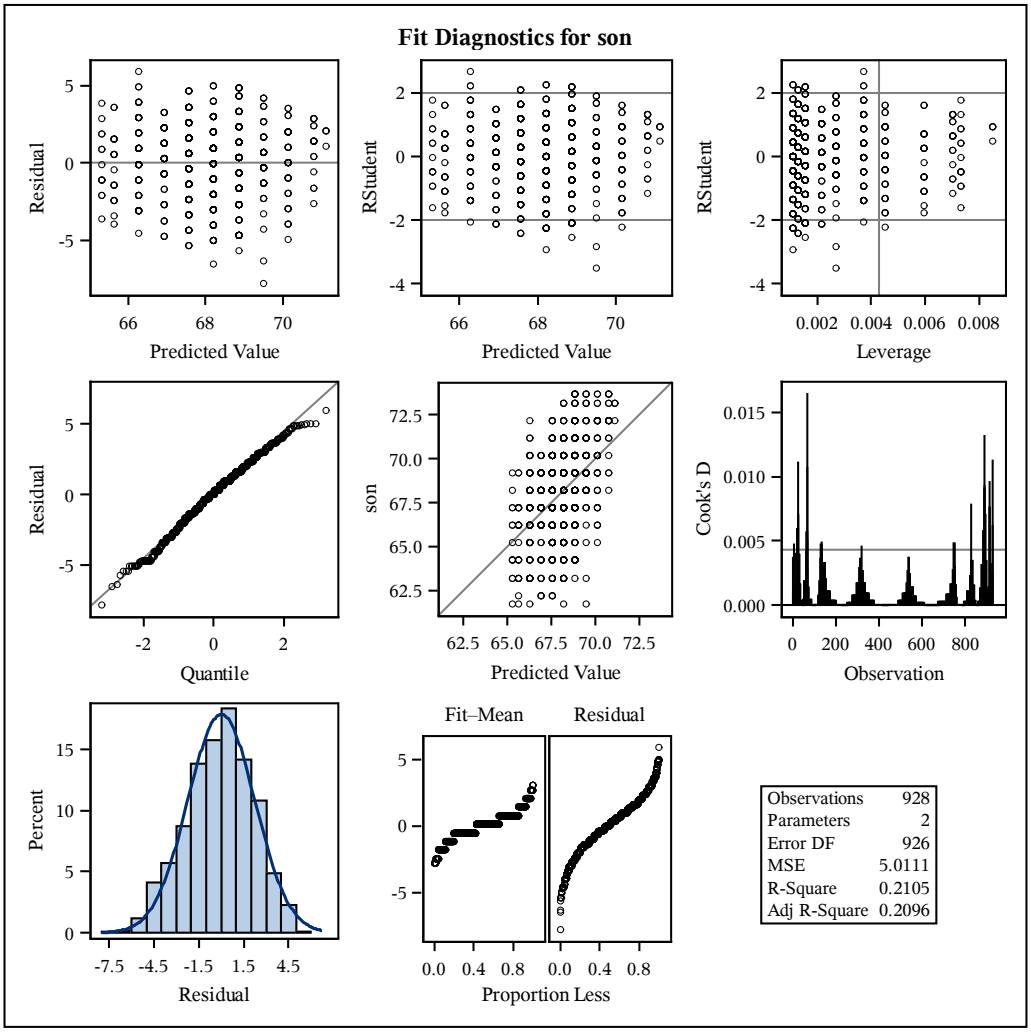
Number of Observations Read	930
Number of Observations Used	928
Number of Observations with Missing Values	2

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	1236.93401	1236.93401	246.84	<.0001
Error	926	4640.27261	5.01109		
Corrected Total	927	5877.20663			

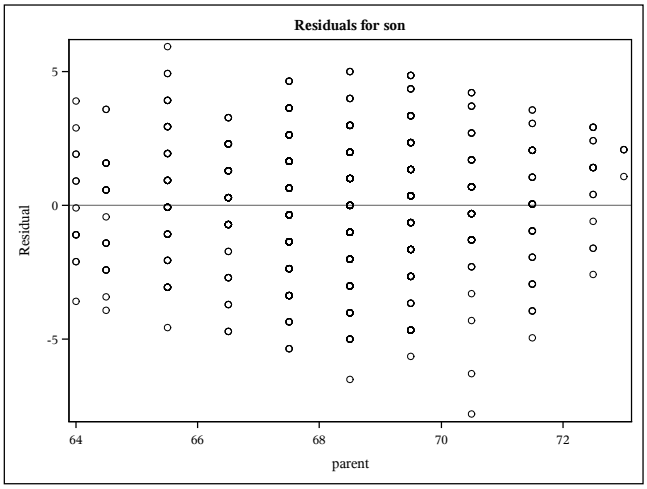
Root MSE	2.23855	**R-Square**	0.2105
Dependent Mean	68.08847	**Adj R-Sq**	0.2096
Coeff Var	3.28770		

Parameter Estimates							
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**95% Confidence Limits**	
Intercept	1	23.94153	2.81088	8.52	<.0001	18.42510	29.45796
parent	1	0.64629	0.04114	15.71	<.0001	0.56556	0.72702

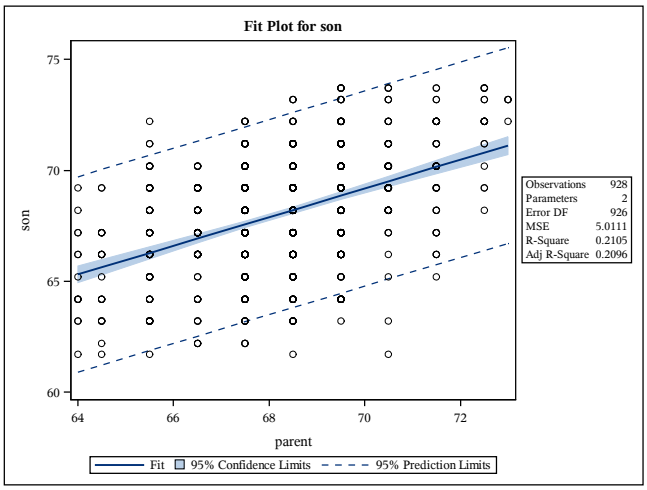
The REG Procedure
Model: MODEL1
Dependent Variable: son



The REG Procedure
Model: MODEL1
Dependent Variable: son



The REG Procedure
Model: MODEL1
Dependent Variable: son



Galton Height Data Output 6

Obs	parent	son	yhat	stdmean	cilow	cihigh	pilow	pihigh	r
1	68	.	67.8893	0.07457	67.7429	68.0356	63.4936	72.2849	.
2	72	.	70.4745	0.16871	70.1434	70.8056	66.0688	74.8801	.

Answers to questions from simple linear regression:

1. A CI for a mean when the true standard deviation is unknown is

$$\bar{y} \pm t(n-1, \alpha/2)s/\sqrt{n} = 68.09 \pm 1.963 * 2.52/\sqrt{928} = (67.928, 68.252)$$

2. Change in average son's height (inches) per one inch increase in midparent height.
3. β_1 is an unknown parameter.
4. $\hat{\beta}_1 = 0.65$ son inches/midparent inch.
5. $\hat{\beta}_1 = 0.65$ is an observed value of a statistic.
6. β_1 is the slope of the population mean, $\hat{\beta}_1$ is the slope from the SLR of the observed data. $\hat{\beta}_1 = \beta_1$ is very unlikely.
7. $\widehat{SE}(\hat{\beta}_1) = \sqrt{MS[E]/S_{xx}} = 0.041$.
8. Add and subtract 1.963 times the SE to get $(0.566, 0.727)$
9. $y = 23.942 + 0.646x$
10. $r^2 = 21.05\%$
11. $\mu(68) = \beta_0 + 68\beta_1$
12. $\hat{\mu}(68) = \hat{\beta}_0 + 68\hat{\beta}_1 = 67.889$
13. Probably not! $\mu(68) = \beta_0 + 68\beta_1$ is unknown and $\hat{\mu}(68)$ is only an estimate.
14. This question is asking for the square root of the estimate of variation due to experimental error or $\hat{\sigma} = \sqrt{MS[E]} = 2.24$.
15. Same as the previous question as an assumption on our model is that the errors have the same variance (and hence square root) at every point on the line. (Assumption of homoskedasticity.)
16. $SE(\hat{\beta}_0 + 68\hat{\beta}_1) = 0.075$. Expressions given by

$$\begin{aligned} \widehat{SE}(\hat{\mu}(68)) &= \sqrt{MS[E]\left(\frac{1}{n} + \frac{(68-\bar{x})^2}{\sum(x_i - \bar{x})^2}\right)} \\ &= \sqrt{(1,68)' MS[E](X'X)^{-1}(1,68)} \end{aligned}$$

    X a (928×2) *design matrix*.
17. $\widehat{SE}(\hat{\mu}(72)) > \widehat{SE}(\hat{\mu}(68))$ as it is further from $\bar{x}$, where we have the most information about our data.
18. ρ, which is the population correlation coefficient (a parameter).

19. $r = 0.459$, moderate, positive. P-value ¡ 0.0001, there is a significant linear relationship.

20. These are all parameters and describe the mean and standard deviation of the sons‘ heights, the mean and standard deviation of the midparent‘s heights, and the correlation between them.

21. The confidence interval is

$$\left(\frac{\frac{1+r}{1-r} e^{-2z/\sqrt{n-3}} - 1}{\frac{1+r}{1-r} e^{-2z/\sqrt{n-3}} + 1}, \frac{\frac{1+r}{1-r} e^{2z/\sqrt{n-3}} - 1}{\frac{1+r}{1-r} e^{2z/\sqrt{n-3}} + 1} \right)$$

    or $(0.406, 0.508)$.

22. Residuals reasonably symmetric, no heavy tails. Model fit is ok.

23. Based on this study alone, no. This is an observational study as no midparent heights were assigned by the researchers. However, if science and genetics are brought in, a causal relationship might be inferred.

Recall the example about a random sample of students taking the same exam:

IQ	TIME	GRADE
105	10	75
110	12	79
120	6	68
116	13	85
122	16	91
130	8	79
114	20	98
102	15	76

Consider the additive regression model for the GRADE of subject i, Y_i, in which the mean of Y_i is a linear function of IQ and Time ($X_{i1} = \text{IQ}$ and $X_{i2} = \text{TIME}$) for subjects $i = 1, \ldots, 8$:

$$Y = \beta_0 + \beta_1 \text{IQ} + \beta_2 \text{TIME} + \text{error}$$

or

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + E_i$$

Let's write out the model in matrix form:

$$\mathbf{y}=\begin{pmatrix} 75 \\ 79 \\ 68 \\ 85 \\ 91 \\ 79 \\ 98 \\ 76 \end{pmatrix} \quad \mathbf{X}=\begin{pmatrix} 1 & 105 & 10 \\ 1 & 110 & 12 \\ 1 & 120 & 6 \\ 1 & 116 & 13 \\ 1 & 122 & 16 \\ 1 & 130 & 8 \\ 1 & 114 & 20 \\ 1 & 102 & 15 \end{pmatrix} \quad \boldsymbol{\beta}=\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 8 & 919 & 100 \\ 919 & 106165 & 11400 \\ 100 & 11400 & 1394 \end{pmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 28.90 & -0.23 & -0.22 \\ -0.23 & 0.0018 & 0.0011 \\ -0.22 & 0.0011 & 0.0076 \end{pmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{pmatrix} 0.74 \\ 0.47 \\ 2.10 \end{pmatrix}$$

$$SS(E) = \mathbf{e}'\mathbf{e} = (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) = 45.8, \quad \mathbf{e}'\mathbf{e}/df = 9.15$$

$$\widehat{\boldsymbol{\Sigma}} = MS(E)(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 264.45 & -2.07 & -2.05 \\ -2.07 & 0.017 & 0.010 \\ -2.05 & 0.010 & 0.070 \end{pmatrix}$$

Some questions to answer:

1. What is the estimate for β_1? Interpretation?
2. What is the standard error of $\hat{\beta}_1$?
3. Is $\beta_1 = 0$ plausible, while controlling for possible linear associations between Test Score and Study time? $(t(0.025, 5) = 2.57)$
4. Estimate the mean grade among the population of ALL students with $IQ = 113$ who study $TIME = 14$ hours.
5. Report a standard error for this mean.
6. Report a 95% confidence interval for this mean.
7. What is the estimate of the error variance?

Some answers:

1. $\hat{\beta}_1 = 0.47$ (second element of $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, estimated average exam points per IQ point for students studying the same amount)
2. $\sqrt{0.017} = 0.13$ (square root of middle element of $\widehat{\mathbf{\Sigma}}$)
3. $H_0 : \beta_1 = 0$, T-statistic: $t = (\hat{\beta}_1 - 0)/SE(\hat{\beta}_1)$
   Observed value is $t = .47/\sqrt{.017} = .47/.13 = 3.6 > 2.57$ ("$\hat{\beta}_1$ differs significantly from 0.")
4. Unknown population mean: $\theta = \beta_0 + \beta_1(113) + \beta_1(14)$
   Estimate : $\hat{\theta} = (1, 113, 14) * \hat{\beta} = 83.6$
5. $\text{Var}((1, 113, 14) * \hat{\boldsymbol{\beta}}) = (1, 113, 14)\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})(1, 113, 14)'$
   or $(1, 113, 14)\widehat{\mathbf{\Sigma}}(1, 113, 14)' = 1.3$ or $SE(\hat{\theta}) = \sqrt{1.3} = 1.14$
6. $\hat{\theta} \pm t(0.025, 5)SE(\hat{\theta})$ or $83.6 \pm 2.57(1.14)$ or $(80.7, 86.6)$
7. The estimate of the error variance is the $MS(E) = 9.15$

Continuing this example, consider this sequence of analyses:

1. Regress GRADE on IQ.
2. Regress GRADE on IQ and TIME.
3. Regress GRADE on TIME IQ TI where TI = TIME*IQ.

ANOVA (Grade on IQ)

SOURCE	DF	SS	MS	F	p-value
IQ	1	15.9393	15.9393	0.153	0.71
Error	6	625.935	104.32		

It appears that IQ has nothing to do with grade, but we did not look at study time.

Looking at the *multiple* regression we get

The REG Procedure

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	596.11512	298.05756	32.57	0.0014
Error	5	45.75988	9.15198		
Corrected Total	7	641.87500			

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.73655	16.26280	0.05	0.9656
IQ	1	0.47308	0.12998	3.64	0.0149
Time	1	2.10344	0.26418	7.96	0.0005

Now the test for dependence on IQ is significant $p = 0.0149$. Why? This is a slightly different test. This is testing if IQ is important after taking into account the linear relationship between Grade and Time.

Now recall when we fit the interaction model we found the following (here type I SS are included):

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS
Intercept	1	72.20608	54.07278	1.34	0.2527	52975
IQ	1	-0.13117	0.45530	-0.29	0.7876	15.93930
Time	1	-4.11107	4.52430	-0.91	0.4149	580.17582
TI	1	0.05307	0.03858	1.38	0.2410	14.69521

This model now appears to be over-fit as the type III tests (tests done after accounting for all other variables in the model) are all non-significant.

We can perform a LOF test to see if the interaction model or the MLR model is preferred. We can find $SS(R)_f$ by summing the type I SS,

$$SS(R)_f = 15.939 + 580.176 + 14.695 = 610.8103$$

The $SS(R)_r$ can be found by subtracting off the TI type I SS from $SS(R)_f$ or by adding all type I SS except TI giving

$$SS(R)_r = 610.8103 - 14.695 = 596.115$$

Now the numerator of our LOF statistic is

$$\frac{610.8103 - 596.115}{3 - 2} = \frac{14.695}{1} = 14.7$$

(Note this is the type I SS for TI!). Our LOF stat is then

$$F = 14.7/7.766 = 1.893$$

(MSE found from output in earlier notes.) Comparing this to $F(1, 4, 0.05) = 7.709$ we fail to reject H_0 in favor of H_A. That is, the additive model is adequate.

A random sample of $n = 31$ trees is drawn from a population of trees. On each tree, indexed by i, three variables are measured:

- x_{i1}: "girth", tree diameter in inches
- x_{i2}: "height" (in feet)
- Y_i: volume of timber, in cubic feet.

Given x_1 and x_2, a MLR model for these data is given by

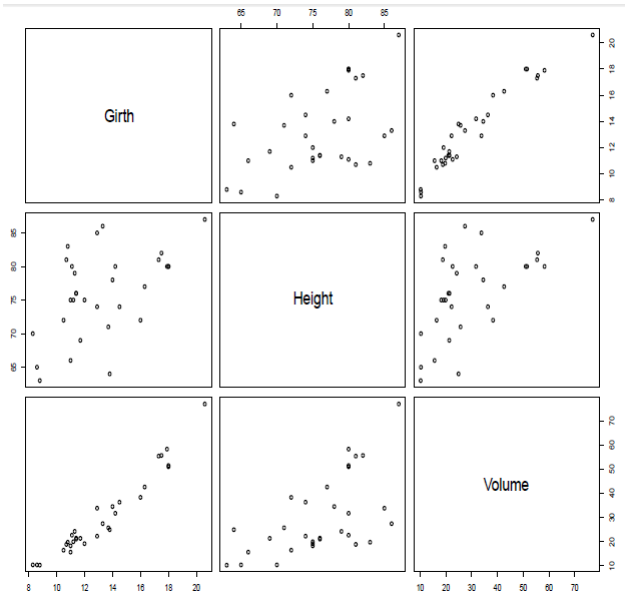
$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + E_i \text{ for } i = 1, \ldots, n$$

where errors are assumed iid normal w/ constant variance σ^2.

For trees with x_1, x_2 the model for mean volume is

$$\mu(x_1, x_2) = E(Y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

A scatterplot matrix



Consider all trees with girth $x_{01} = 15\ in$ and height $x_{02} = 80\ ft$.

- Estimate the mean volume among these trees, along with a standard error and 95% confidence interval. $Note: t(28, 0.025) = 2.048$
- Obtain a 95% prediction interval of y_0, the volume from an individual tree sampled from this population of 80 footers, with girth 15 inches.

SAS generates $\hat{\beta}$ and $\widehat{Var}(\hat{\beta}) = MSE * (X'X)^{-1}$

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	7684.16251	3842.08126	254.97	<.0001
Error	28	421.92136	15.06862		
Corrected Total	30	8106.08387			

Root MSE	3.88183	R-Square	0.9480
Dependent Mean	30.17097	Adj R-Sq	0.9442
Coeff Var	12.86612		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-57.98766	8.63823	-6.71	<.0001
Girth	1	4.70816	0.26426	17.82	<.0001
Height	1	0.33925	0.13015	2.61	0.0145

Covariance of Estimates

Variable	Intercept	Girth	Height
Intercept	74.6189461	0.4321713812	-1.050768886
Girth	0.4321713812	0.0698357838	-0.017860301
Height	-1.050768886	-0.017860301	0.0169393298

Recall: Let $\mathbf{W}$ denote a $p \times 1$ random vector with mean $\boldsymbol{\mu}_{\mathbf{W}}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{W}}$. Suppose $\mathbf{a}$ is a $p \times 1$ (fixed) vector of coefficients. Then

$$\boxed{\begin{array}{rcl} E(\mathbf{a}'\mathbf{W}) & = & \mathbf{a}'\boldsymbol{\mu}_{\mathbf{W}} \\ \text{Var}(\mathbf{a}'\mathbf{W}) & = & \mathbf{a}'\boldsymbol{\Sigma}_{\mathbf{W}}\mathbf{a}. \end{array}}$$

1. Consider all trees with Girth 15 and Height 80 To estimate mean volume among these trees, along with an estimated standard error, take $\mathbf{x}_0' = (1, 15, 80)$ and consider

$$\hat{\mu}(\mathbf{x}_0) = \mathbf{x}_0'\hat{\boldsymbol{\beta}}$$

$$E(\mathbf{x}_0'\hat{\boldsymbol{\beta}}) = \mathbf{x}_0'\boldsymbol{\beta}$$

$$Var(\mathbf{x}_0'\hat{\boldsymbol{\beta}}) = \mathbf{x}_0'\hat{\boldsymbol{\Sigma}}\mathbf{x}_0$$

Substitution of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}} = MSE(\mathbf{X}'\mathbf{X})^{-1}$ gives the estimates:

$$\begin{array}{rcl} \hat{\mu}(\mathbf{x}_0) & = & (1, 15, 80) \begin{pmatrix} -58.0 \\ 4.71 \\ 0.34 \end{pmatrix} \\ & = & 39.8 \\ \widehat{\text{Var}}(\hat{\mu}(\mathbf{x}_0)) & = & (1, 15, 80) \begin{pmatrix} 74.62 & 0.43 & -1.05 \\ 0.43 & 0.070 & -0.018 \\ -1.05 & -0.018 & 0.017 \end{pmatrix} \begin{pmatrix} 1 \\ 15 \\ 80 \end{pmatrix} \\ & = & 0.72 \\ \widehat{SE}(\hat{\mu}(\mathbf{x}_0)) & = & \sqrt{.72} = 0.849 \end{array}$$

Thus, a 95% CI is given by

$$39.8 \pm 2.048 * 0.849 = (38.061, 41.539)$$

2. 95% Prediction limits? Same idea but we add and subtract $t(.025, 28)\sqrt{.72 + MS(E)}$.

Chapter 7

ST 512 - General Linear Models

Readings: 12.1-12.6 pg 540 - 583

A general linear model:
Models which can include both qualitative and/or quantitative explanatory variables are called general linear models or GLMs. (Again, the 'linearity' pertains to the parameters, not the explanatory variables.)

How to write qualitative variables in a GLM format?

ANOVA (Analysis of Variance, i.e. comparing mean squares) revisited:

The One-Way ANOVA model is used when we wish to compare the means of t different groups. (One-Way corresponds to having only one factor of interest.)

Often a completely randomized experimental design will be analyzed using an ANOVA model.

One form of the One-Way ANOVA model is

$$Y_{ij} = \mu + \tau_i + E_{ij}$$

- E_{ij} are i.i.d. $N(0, \sigma^2)$
- $i = 1, ..., t$ describes the treatment group
- $j = 1, ..., n_i$ represents the number of observations we have in treatment group i.

We will consider 'balanced' designs for now, where $n_i = n$, same number of replicates for each treatment. Total number of observations $= N = nt$

Unknown parameters:

- μ - overall population mean (avg of treatment population means)
- τ_i - difference between (population) mean for treatment i and μ
- σ^2 - (population) variance within a given treatment group (constant across groups)

Goals of One-Way ANOVA: Determine

1. if all treatment means are equal.
2. if treatment means not equal, which means differ from eachother.

One-Way ANOVA example:
An experiment was done to determine if there was a difference between antibiotic types in terms of their binding fraction in bovines. There were N=20 bovines that were randomly assigned to one of t=5 types of antibiotics (the levels of the factor, since only one factor these levels are also the treatments), yielding n=4 replicates for each treatment. The data given here, labeled in terms of the One-Way ANOVA format:

Binding Fraction (Y)	Antibiotic	True Trt Mean	Sample Mean
$Y_{11} = 29.6$	Penicillin G	$\mu + \tau_1$	$\bar{y}_{1+} = 28.6$
$Y_{12} = 24.3$	Penicillin G		
$Y_{13} = 28.5$	Penicillin G		
$Y_{14} = 32.0$	Penicillin G		
$Y_{21} = 27.3$	Tetracyclin	$\mu + \tau_2$	$\bar{y}_{2+} = 31.4$
$Y_{22} = 32.6$	Tetracyclin		
$Y_{23} = 30.8$	Tetracyclin		
$Y_{24} = 34.8$	Tetracyclin		
$Y_{31} = 5.8$	Streptomycin	$\mu + \tau_3$	$\bar{y}_{3+} = 7.8$
$Y_{32} = 6.2$	Streptomycin		
$Y_{33} = 11.0$	Streptomycin		
$Y_{34} = 8.3$	Streptomycin		
$Y_{41} = 21.6$	Erythromycin	$\mu + \tau_4$	$\bar{y}_{4+} = 19.1$
$Y_{42} = 17.4$	Erythromycin		
$Y_{43} = 18.3$	Erythromycin		
$Y_{44} = 19.0$	Erythromycin		
$Y_{51} = 29.2$	Chloramphenicol	$\mu + \tau_5$	$\bar{y}_{5+} = 27.8$
$Y_{52} = 32.8$	Chloramphenicol		
$Y_{53} = 25.0$	Chloramphenicol		
$Y_{54} = 24.2$	Chloramphenicol		
$\bar{y}_{++} = 22.9$			

Goal: Test if the population means for these 5 treatments are plausibly equal.
If so, which treatment means differ significantly?

Modeling the binding fraction experiment:
One-Way ANOVA model is appropriate:

$$Y_{ij} = \mu + \tau_i + E_{ij}$$

for $i = 1, \ldots, 5$ and $j = 1, \ldots, 4$, where E_{ij} are i.i.d. $N(0, \sigma^2)$ errors.

To test $H_0 : \tau_1 = \tau_2 = \ldots = \tau_5 = 0$, we just carry out One-Way ANOVA table and look at global p-value.

Table for balanced one-way ANOVA:

Source	DF	SS	MS	F
Treatments	$t-1$	$SS(T)$	$MS(T) = \frac{SS(T)}{(t-1)}$	$F = \frac{MS(T)}{MS(E)}$
Error	$t(n-1)$	$SS(E)$	$MS(E) = \frac{SS(E)}{(N-t)}$	
Total	$nt-1$	$SS(TOT)$		

where

$$\begin{aligned} SS(T) &= \sum_{i=1}^{t}\sum_{j=1}^{n}(\bar{y}_{i+} - \bar{y}_{++})^2 = n\sum_{i=1}^{t}(\bar{y}_{i+} - \bar{y}_{++})^2 \\ SS(E) &= \sum_{i=1}^{t}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i+})^2 \\ SS(Tot) &= \sum_{i=1}^{t}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{++})^2 \end{aligned}$$

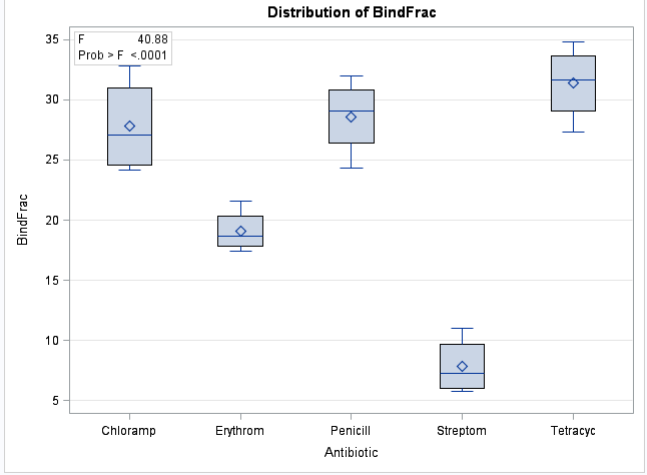
Note: SS(T) is also called SS(Between) and SS(E) is also called SS(Within).

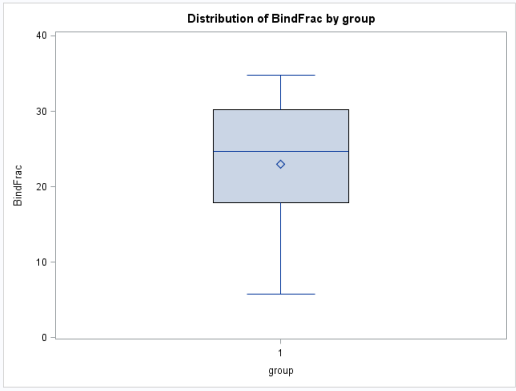
For our example,

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1480.823000	370.205750	40.88	<.0001
Error	15	135.822500	9.054833		
Corrected Total	19	1616.645500			

R-Square	Coeff Var	Root MSE	BindFrac Mean
0.915985	13.12023	3.009125	22.93500

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Antibiotic	4	1480.823000	370.205750	40.88	<.0001





Conclusion about treatment means begin equal?

Parameter estimates:

- $\hat{\mu} = \bar{y}_{++}$
- $\hat{\tau}_i = \bar{y}_{i+} - \bar{y}_{++}$ and standard errors of treatment means are $\sqrt{\frac{MS(E)}{n}}$
- To compare treatment means we look at $\hat{\tau}_i - \hat{\tau}_j = \bar{y}_{i+} - \bar{y}_{j+}$ with SE $= \sqrt{\frac{2MS(E)}{n}}$

Matrix formulation of the GLM representation of the One-way ANOVA model: (will allow us to make inference just as we've done previously!)

$$\mathbf{y}=\begin{pmatrix} 29.6 \\ 24.3 \\ 28.5 \\ 32.0 \\ 27.3 \\ 32.6 \\ 30.8 \\ 34.8 \\ 5.8 \\ 6.2 \\ 11.0 \\ 8.3 \\ 21.6 \\ 17.4 \\ 18.3 \\ 19.0 \\ 29.2 \\ 32.8 \\ 25.0 \\ 24.2 \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix} \quad \boldsymbol{\beta}=\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}$$

SAS proc glm code and output are given below:

```
proc glm data=binding;
class antibiotic;
model bindfrac=antibiotic/solution inverse;
run;
```

The GLM Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1480.823000	370.205750	40.88	<.0001
Error	15	135.822500	9.054833		
Corrected Total	19	1616.645500			

R-Square	Coeff Var	Root MSE	BindFrac Mean
0.915985	13.12023	3.009125	22.93500

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Antibiotic	4	1480.823000	370.205750	40.88	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Antibiotic	4	1480.823000	370.205750	40.88	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	31.37500000	B	1.50456251	20.85	<.0001
Antibiotic Chloramp	-3.57500000	B	2.12777270	-1.68	0.1136
Antibiotic Erythrom	-12.30000000	B	2.12777270	-5.78	<.0001
Antibiotic Penicill	-2.77500000	B	2.12777270	-1.30	0.2118
Antibiotic Streptom	-23.55000000	B	2.12777270	-11.07	<.0001
Antibiotic Tetracyc	0.00000000	B	.	.	.

X'X Generalized Inverse (g2)							
	Intercept	Dummy001	Dummy002	Dummy003	Dummy004	Dummy005	BindFrac
Intercept	0.25	-0.25	-0.25	-0.25	-0.25	0	31.375
Dummy001	-0.25	0.5	0.25	0.25	0.25	0	-3.575
Dummy002	-0.25	0.25	0.5	0.25	0.25	0	-12.3
Dummy003	-0.25	0.25	0.25	0.5	0.25	0	-2.775
Dummy004	-0.25	0.25	0.25	0.25	0.5	0	-23.55
Dummy005	0	0	0	0	0	0	0
BindFrac	31.375	-3.575	-12.3	-2.775	-23.55	0	135.8225

Estimates of the β's still found by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{pmatrix} 27.8 \\ 0.8 \\ 3.6 \\ -20.0 \\ -8.7 \end{pmatrix}$$

Estimates for the five treatment means obtained by using combinations from the $\hat{\boldsymbol{\beta}}$ vector

$$\mu(x_1, x_2, x_3, x_4) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

$$\begin{aligned}
\text{Trt 1 estimate} &= \hat{\mu}(1,0,0,0) = \hat{\beta}_0 + \hat{\beta}_1 = 28.6 \\
\text{Trt 2 estimate} &= \hat{\mu}(0,1,0,0) = \hat{\beta}_0 + \hat{\beta}_2 = 31.4 \\
\text{Trt 3 estimate} &= \hat{\mu}(0,0,1,0) = \hat{\beta}_0 + \hat{\beta}_3 = \ 7.8 \\
\text{Trt 4 estimate} &= \hat{\mu}(0,0,0,1) = \hat{\beta}_0 + \hat{\beta}_4 = 19.1 \\
\text{Trt 5 estimate} &= \hat{\mu}(0,0,0,0) = \hat{\beta}_0 \qquad\quad = 27.8
\end{aligned}$$

For standard errors of the $\hat{\beta}$'s we still have our variance-covariance matrix $\hat{\boldsymbol{\Sigma}} = MS(E)(\mathbf{X}^T\mathbf{X})^{-1}$

$$\hat{\boldsymbol{\Sigma}} = MS(E)(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 2.3 & -2.3 & -2.3 & -2.3 & -2.3 \\ & 4.5 & 2.3 & 2.3 & 2.3 \\ & & 4.5 & 2.3 & 2.3 \\ & & & 4.5 & 2.3 \\ & & & & 4.5 \end{pmatrix}$$

Note the pattern, what is the reason for it?

To get SE's of our treatment mean estimates we can use vectors: Let $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}$ be defined by

$$\mathbf{a}^T = (1,1,0,0,0), \mathbf{b}^T = (1,0,1,0,0), \mathbf{c}^{-1} = (1,0,0,1,0), \mathbf{d}^T = (1,0,0,0,1).$$

Then

$$\begin{array}{lcll}
\hat{\mu}(1,0,0,0) & = & \hat{\beta}_0 + \hat{\beta}_1 & = \mathbf{a}^T\hat{\beta} \\
\hat{\mu}(0,1,0,0) & = & \hat{\beta}_0 + \hat{\beta}_2 & = \mathbf{b}^T\hat{\beta} \\
\hat{\mu}(0,0,1,0) & = & \hat{\beta}_0 + \hat{\beta}_3 & = \mathbf{c}^T\hat{\beta} \\
\hat{\mu}(0,0,0,1) & = & \hat{\beta}_0 + \hat{\beta}_4 & = \mathbf{d}^T\hat{\beta} \\
\hat{\mu}(0,0,0,0) & = & \hat{\beta}_0 & = \hat{\beta}_0
\end{array}$$

and for a balanced design the variances are all the same and are given by

$$\mathbf{a}'\hat{\mathbf{\Sigma}}\mathbf{a} = \mathbf{b}'\hat{\mathbf{\Sigma}}\mathbf{b} = \mathbf{c}'\hat{\mathbf{\Sigma}}\mathbf{c} = \mathbf{d}'\hat{\mathbf{\Sigma}}\mathbf{d} = \widehat{\text{Var}}(\hat{\beta}_0) = \widehat{\text{Var}}(\hat{\beta}_0 + \hat{\beta}_j) = 2.3$$

so the estimated SE for any sample treatment mean is $\sqrt{2.3} = 1.5$.

Recall from one-way ANOVA that

$$\widehat{SE}(\bar{y}_{i+}) = \sqrt{\frac{MS(E)}{n}} = \sqrt{\frac{9.1}{4}} = \sqrt{2.3} = 1.5$$

and that for differences between treatment means

$$\widehat{SE}(\bar{y}_{i+} - \bar{y}_{j+}) = \sqrt{\frac{2MS(E)}{n}} = \sqrt{4.5} = 2.1$$

Now we have a framework to use both quantitative and categorical explanatory variables at the same time!

A general linear model for 5k times of men AND women:
Using X_1=Age, we fit a quadratic model $\mu(x_1) = \beta_0 + \beta_1 x_1 + \beta_1 x_1^2$ for to predict the mean pace for runners. Consider modeling gender (a categorical variable) as well.
Let x_3 be defined by

$$x_3 = \begin{cases} 1 & \text{female} \\ 0 & \text{male} \end{cases}$$

Some candidate models:

1. The 'null' model:
$$\mu(x_1, x_3) = \beta_0$$

2. The One-Way ANOVA model in GLM form:
$$\mu(x_1, x_3) = \beta_0 + \beta_3 x_3$$

3. The SLR model using Age:
$$\mu(x_1, x_3) = \beta_0 + \beta_1 x_1$$

4. The MLR model quadratic in Age:

$$\mu(x_1, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$$

5. A GLM that allows for different intercepts for our parabolas:

$$\mu(x_1, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_3$$

This model has intercept β_0 for males ($x_3 = 0$) and intercept $\beta_0 + \beta_3$ for females ($x_3 = 1$).

Equation for males: $\beta_0 + \beta_1 x_1 + \beta_2 x_1^2$

Equation for females: $(\beta_0 + \beta_3) + \beta_1 x_1 + \beta_2 x_1^2$

6. A GLM that allows for different intercepts and for different shapes of the parabolas:

$$\mu(x_1, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_3 + \beta_4 x_1 x_3 + \beta_5 x_1^2 x_3$$

Intercepts as in the previous model. 'Linear' term for males is β_1 and is $\beta_1 + \beta_4$ for females. 'Quadratic' term for males is β_2 and is $\beta_2 + \beta_5$ for females.

Equation for males: $\beta_0 + \beta_1 x_1 + \beta_2 x_1^2$

Equation for females: $(\beta_0 + \beta_3) + (\beta_1 + \beta_4) x_1 + (\beta_2 + \beta_5) x_1^2$

We can fit these models in proc glm using the following code: (Note: each model must be done in a separate proc glm statement)

```
proc glm;
class sex;
title 'Model 2';
model pace=sex;
title 'Model 3';
model pace=age;
title 'Model 4';
model pace=age age*age;
title 'Model 5';
model pace=age age*age sex;
title 'Model 6';
model pace=age age*age sex sex*age sex*age*age;
run;
```

The GLM Procedure

Dependent Variable: pace

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	170.7413698	170.7413698	43.70	<.0001
Error	158	617.3533455	3.9072997		
Corrected Total	159	788.0947153			

R-Square	Coeff Var	Root MSE	pace Mean
0.216651	21.67274	1.976689	9.120625

Source	DF	Type I SS	Mean Square	F Value	Pr > F
sex	1	170.7413698	170.7413698	43.70	<.0001

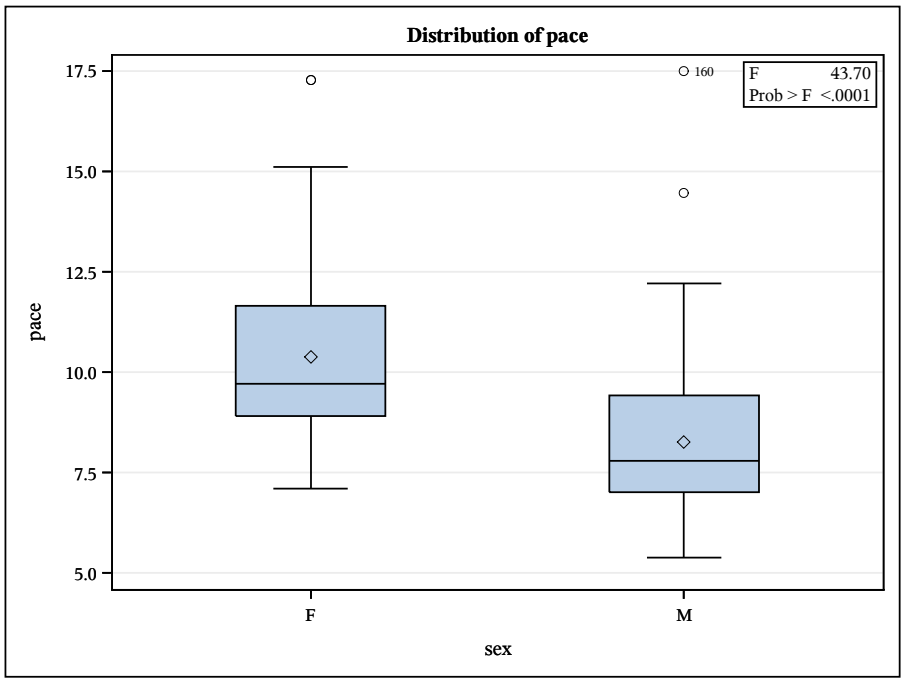
Source	DF	Type III SS	Mean Square	F Value	Pr > F
sex	1	170.7413698	170.7413698	43.70	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	8.266140351	B	0.20280402	40.76	<.0001
sex F	2.103346829	B	0.31818512	6.61	<.0001
sex M	0.000000000	B	.	.	.

Note: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

The GLM Procedure

Dependent Variable: pace



The GLM Procedure

Dependent Variable: pace

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.0965043	1.0965043	0.22	0.6396
Error	158	786.9982110	4.9810013		
Corrected Total	159	788.0947153			

R-Square	Coeff Var	Root MSE	pace Mean
0.001391	24.46999	2.231816	9.120625

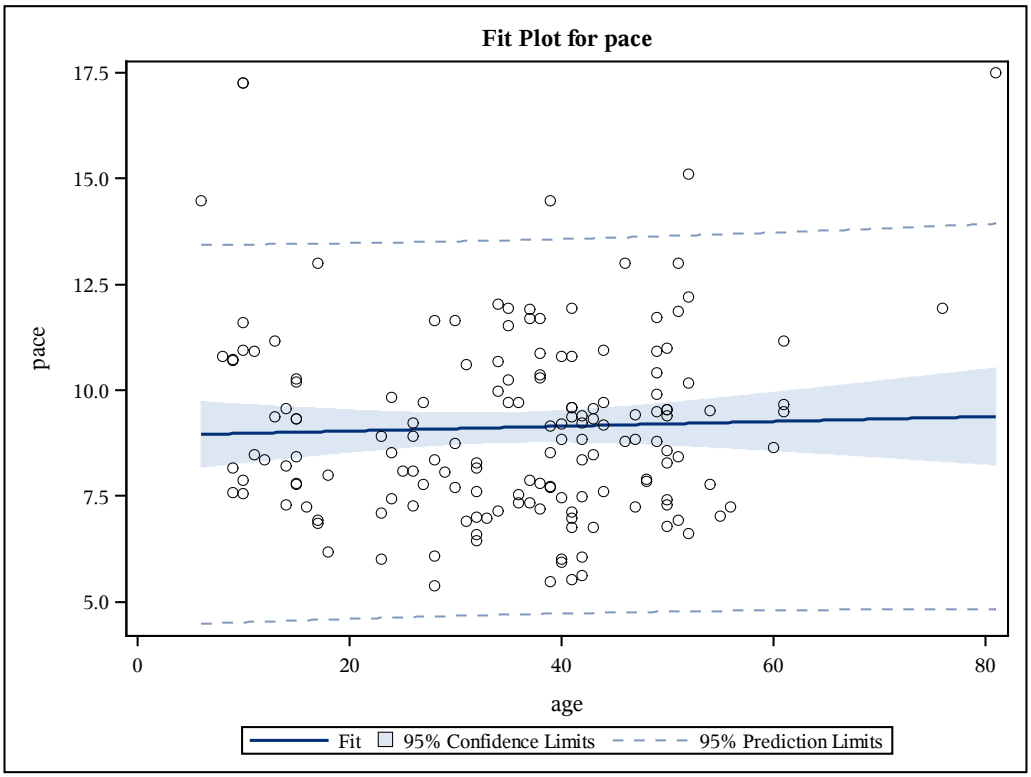
Source	DF	Type I SS	Mean Square	F Value	Pr > F
age	1	1.09650427	1.09650427	0.22	0.6396

Source	DF	Type III SS	Mean Square	F Value	Pr > F
age	1	1.09650427	1.09650427	0.22	0.6396

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	8.922709126	0.45724042	19.51	<.0001
age	0.005643654	0.01202856	0.47	0.6396

The GLM Procedure

Dependent Variable: pace



Model 4

The GLM Procedure

Dependent Variable: pace

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	113.6450003	56.8225001	13.23	<.0001
Error	157	674.4497150	4.2958581		
Corrected Total	159	788.0947153			

R-Square	Coeff Var	Root MSE	pace Mean
0.144202	22.72482	2.072645	9.120625

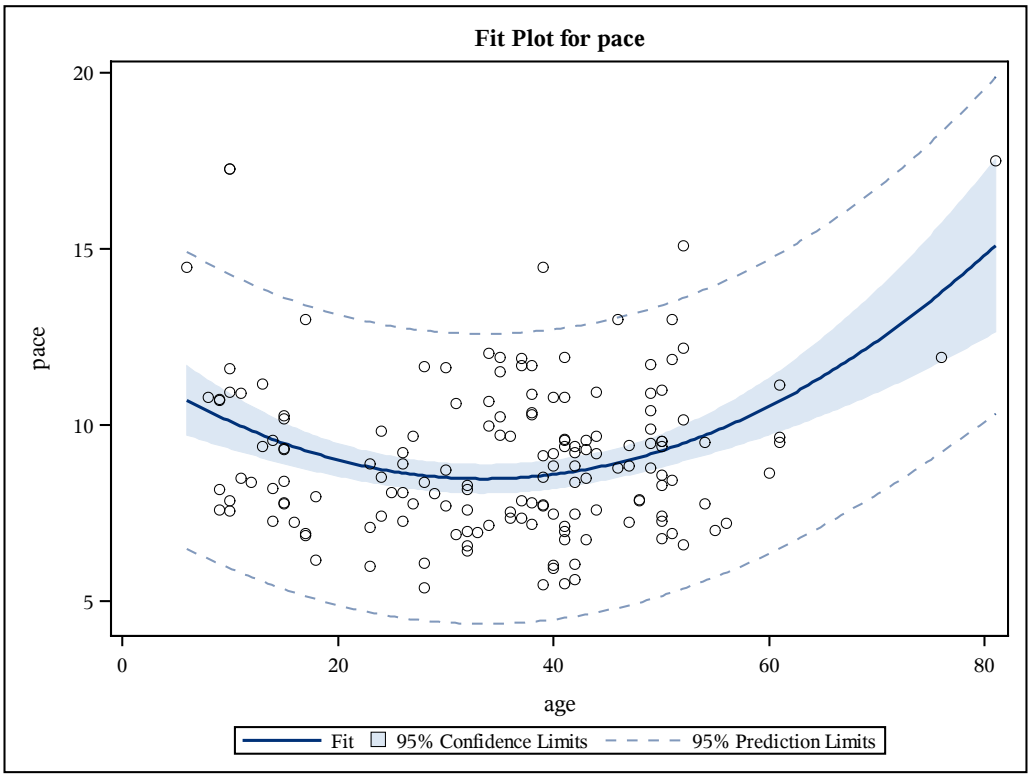
Source	DF	Type I SS	Mean Square	F Value	Pr > F
age	1	1.0965043	1.0965043	0.26	0.6141
age*age	1	112.5484960	112.5484960	26.20	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
age	1	98.5223939	98.5223939	22.93	<.0001
age*age	1	112.5484960	112.5484960	26.20	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	11.78503486	0.70215799	16.78	<.0001
age	-0.19699301	0.04113470	-4.79	<.0001
age*age	0.00293699	0.00057380	5.12	<.0001

The GLM Procedure

Dependent Variable: pace



The GLM Procedure

Dependent Variable: pace

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	290.3485074	96.7828358	30.33	<.0001
Error	156	497.7462079	3.1906808		
Corrected Total	159	788.0947153			

R-Square	Coeff Var	Root MSE	pace Mean
0.368418	19.58471	1.786248	9.120625

Source	DF	Type I SS	Mean Square	F Value	Pr > F
age	1	1.0965043	1.0965043	0.34	0.5586
age*age	1	112.5484960	112.5484960	35.27	<.0001
sex	1	176.7035071	176.7035071	55.38	<.0001

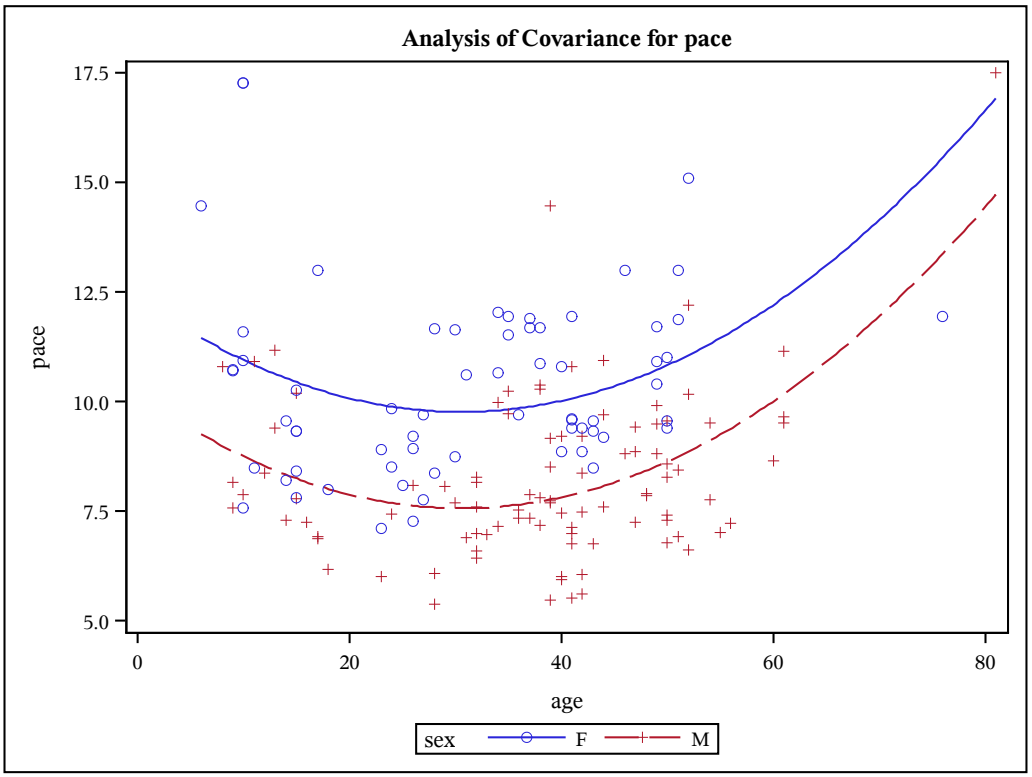
Source	DF	Type III SS	Mean Square	F Value	Pr > F
age	1	73.9438080	73.9438080	23.17	<.0001
age*age	1	102.7473738	102.7473738	32.20	<.0001
sex	1	176.7035071	176.7035071	55.38	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	10.18316690	B	0.64227743	15.85	<.0001
age	-0.17145849		0.03561638	-4.81	<.0001
age*age	0.00280792		0.00049481	5.67	<.0001
sex F	2.19792213	B	0.29534621	7.44	<.0001
sex M	0.00000000	B	.	.	.

Note: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

The GLM Procedure

Dependent Variable: pace



Model 6

The GLM Procedure

Dependent Variable: pace

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	293.5282772	58.7056554	18.28	<.0001
Error	154	494.5664380	3.2114704		
Corrected Total	159	788.0947153			

R-Square	Coeff Var	Root MSE	pace Mean
0.372453	19.64841	1.792058	9.120625

Source	DF	Type I SS	Mean Square	F Value	Pr > F
age	1	1.0965043	1.0965043	0.34	0.5599
age*age	1	112.5484960	112.5484960	35.05	<.0001
sex	1	176.7035071	176.7035071	55.02	<.0001
age*sex	1	0.0057235	0.0057235	0.00	0.9664
age*age*sex	1	3.1740464	3.1740464	0.99	0.3217

Source	DF	Type III SS	Mean Square	F Value	Pr > F
age	1	66.02141759	66.02141759	20.56	<.0001
age*age	1	87.52232536	87.52232536	27.25	<.0001
sex	1	3.34259172	3.34259172	1.04	0.3092
age*sex	1	2.85593189	2.85593189	0.89	0.3471
age*age*sex	1	3.17404636	3.17404636	0.99	0.3217

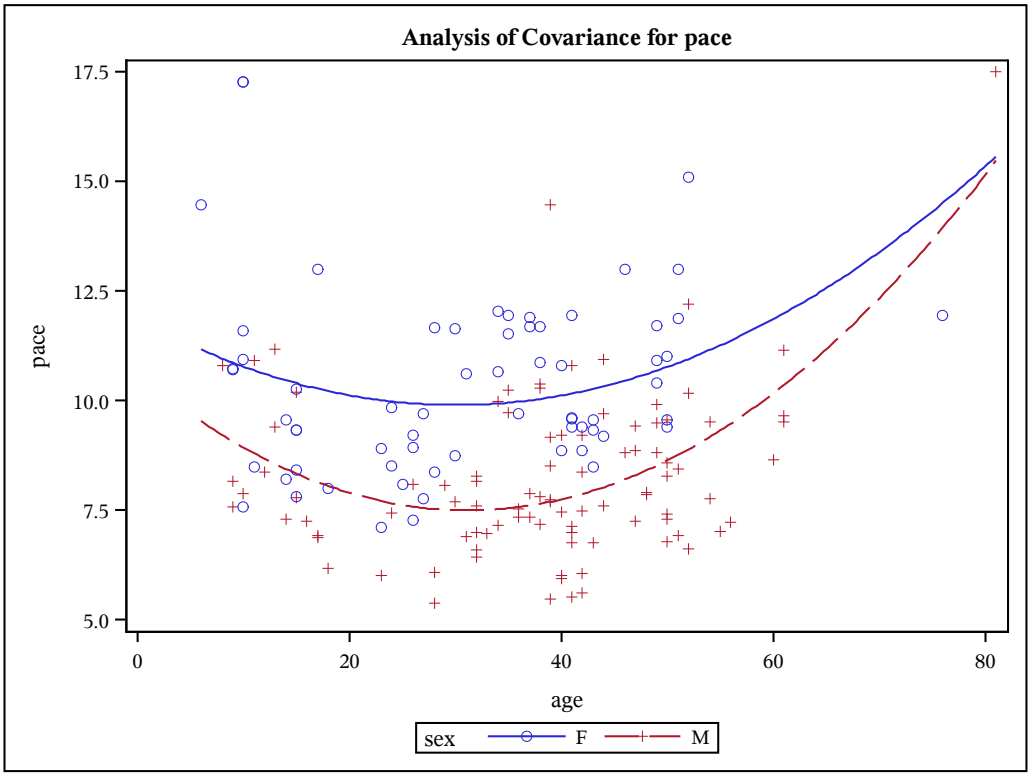
Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	10.60848468	B	0.88640608	11.97	<.0001
age	-0.19985505	B	0.04841621	-4.13	<.0001
age*age	0.00320665	B	0.00064628	4.96	<.0001
sex F	1.25727925	B	1.23237262	1.02	0.3092
sex M	0.00000000	B	.	.	.
age*sex F	0.06882008	B	0.07297821	0.94	0.3471
age*sex M	0.00000000	B	.	.	.

The GLM Procedure

Dependent Variable: pace

Parameter	Estimate		Standard Error	t Value	Pr > |t|
age*age*sex F	-0.00102594	B	0.00103197	-0.99	0.3217
age*age*sex M	0.00000000	B	.	.	.

Note: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.



Analysis of covariance, ANCOVA or ACOVA:
Recall the three principles of experimental design:

- Randomization
- Replication
- Error Reducing Methods

One method of error reduction we looked at was blocking. Here we split up the EUs and then randomize treatments to each block. In this way, every treatment occurs in each block and the block effects cancel out.

We are not always able to block, sometimes we don't have the value of the covariate until after the experiment is done. A similar method that will account for these types of covariates is called Analysis of CoVariance (ANCOVA).

Associations between covariates z and the main response variable of interest y can be used to reduce unexplained variation σ^2.

An nutrition example:
A nutrition scientist conducted an experiment to evaluate the effects of four vitamin supplements on the weight gain of laboratory animals. The experiment was conducted in a completely randomized design with $N = 20$ animals randomized to $a = 4$ supplement groups, each with sample size $n \equiv 5$. The response variable of interest is weight gain, but calorie intake z was measured concomitantly as couldn't separate EUs by this at the beginning.

Diet	$y(g)$	Diet	y	Diet	y	Diet	y
1	48	2	65	3	79	4	59
1	67	2	49	3	52	4	50
1	78	2	37	3	63	4	59
1	69	2	75	3	65	4	42
1	53	2	63	3	67	4	34
1	$\bar{y}_{1+} = 63.0$	2	$\bar{y}_{2+} = 57.4$	3	$\bar{y}_{3+} = 65.2$	4	$\bar{y}_{4+} = 48.8$
1	$s_1 = 12.3$	2	$s_2 = 14.3$	3	$s_3 = 9.7$	4	$s_4 = 10.9$

Q: Is there evidence of a vitamin supplement effect?

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	797.800000	265.933333	1.82	0.1836
Error	16	2334.400000	145.900000		
Corrected Total	19	3132.200000			

P-value < 0.05, fail to reject H_0. There is no evidence of a diet effect.

Calorie intake z was measured concomitantly:

Diet	y	z	Diet	y	z	Diet	y	z	Diet	y	z
1	48	350	2	65	400	3	79	510	4	59	530
1	67	440	2	49	450	3	52	410	4	50	520
1	78	440	2	37	370	3	63	470	4	59	520
1	69	510	2	73	530	3	65	470	4	42	510
1	53	470	2	63	420	3	67	480	4	34	430

Q: How and why could these new data be incorporated into analysis?
A: A GLM can take into account both types of variables! The method of ANCOVA can be used to reduce unexplained variation.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_z z_i + E_i \quad \text{for } i = 1, \ldots, 20$$

where x_{ij} is an indicator variable for subject i receiving vitamin supplement j:

$$x_{ij} = \begin{cases} 1 & \text{subject } i \text{ receives supplement } j \\ 0 & \text{else} \end{cases}$$

and errors $E_i \overset{iid}{\sim} N(0, \sigma^2)$.

Which Diet is being used as the baseline?
Diet 4

Proceeding with MLR analysis of this GLM:

```
proc glm data=diets;
class diet;
model gain = diet caloric;
run;
```

The GLM Procedure

Dependent Variable: gain

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1951.680373	487.920093	6.20	0.0038
Error	15	1180.519627	78.701308		
Corrected Total	19	3132.200000			

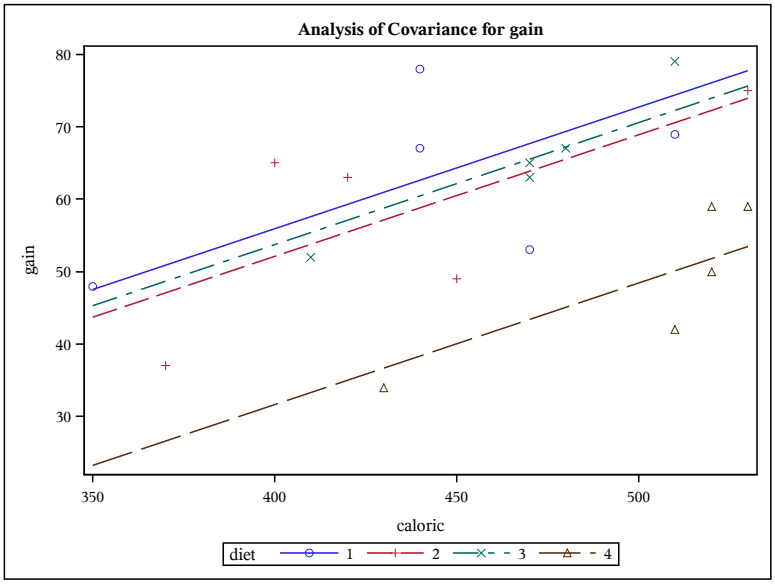
R-Square	Coeff Var	Root MSE	gain Mean
0.623102	15.11308	8.871376	58.70000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
diet	3	797.800000	265.933333	3.38	0.0463
caloric	1	1153.880373	1153.880373	14.66	0.0016

Source	DF	Type III SS	Mean Square	F Value	Pr > F
diet	3	1537.071659	512.357220	6.51	0.0049
caloric	1	1153.880373	1153.880373	14.66	0.0016

The GLM Procedure

Dependent Variable: gain



How can we estimate the mean weight gains for diet, taking into account the caloric intake?

Adjusted vs unadjusted means:

The sample mean weight gains for the four diets and for the caloric intake for each diet group were

Level of diet	N	gain		caloric	
		Mean	Std Dev	Mean	Std Dev
1	5	63.0000000	12.2678441	442.000000	58.9067059
2	5	57.8000000	14.8727940	434.000000	61.0737259
3	5	65.2000000	9.6540147	468.000000	36.3318042
4	5	48.8000000	10.8949530	502.000000	40.8656335

The means for each diet are 'unadjusted' means. According to our analysis, caloric intake has a significant effect on weight gain. However, each diet group had a different mean amount of caloric intake. What does this imply?

Unadjusted means do not make any adjustment for the facts that

1. caloric intake may vary by diet (presumably by chance, not because of diet)
2. weight gain depends on caloric intake

Adjusted means or **lsmeans** (least squares means) will estimate mean weight gains at a common value of our caloric intake (our covariate z). The value often used for comparison is $\bar{z}$, the sample mean of the covariate.

Here, $\bar{z} = (442 + 434 + 468 + 502)/4 = 461.5$. The adjusted means are then just (the sub a is to differentiate unadjusted means and adjusted means)

$$\begin{aligned}
\bar{y}_{1,a} &= \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_z(461.5) \\
\bar{y}_{2,a} &= \hat{\beta}_0 + \hat{\beta}_2 + \hat{\beta}_z(461.5) \\
\bar{y}_{3,a} &= \hat{\beta}_0 + \hat{\beta}_3 + \hat{\beta}_z(461.5) \\
\bar{y}_{4,a} &= \hat{\beta}_0 + \hat{\beta}_z(461.5)
\end{aligned}$$

To get SAS to report the estimated regression parameter vector $\hat{\boldsymbol{\beta}}$, use the `solution` option in the model statement. The default parametrization is the one we've adopted here where β_0 is the mean of the last level of the classification treatment factor:

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	-35.66310108	B	22.41252629	-1.59	0.1324
diet 1	24.29519136	B	6.19932022	3.92	0.0014
diet 2	20.44121688	B	6.35678835	3.22	0.0058
diet 3	22.12060844	B	5.80625371	3.81	0.0017
diet 4	0.00000000	B	.	.	.
caloric	0.16825319		0.04394140	3.83	0.0016

Substitution of $\hat{\boldsymbol{\beta}}$ into the expressions for adjusted means yields

$$\begin{aligned}
\bar{y}_{1,a} &= -35.7 + 24.3 + 0.17(461.5) = 66.3 \\
\bar{y}_{2,a} &= -35.7 + 20.4 + 0.17(461.5) = 62.4 \\
\bar{y}_{3,a} &= -35.7 + 22.1 + 0.17(461.5) = 64.1 \\
\bar{y}_{4,a} &= -35.7 + +0.17(461.5) = 42.0
\end{aligned}$$

These means are better for comparisons between diets as the effect of caloric intake (which affects weight gain) is constant across all the diets. Thus, we are removing its effect.

To get proc glm to produce the estimates (above), unadjusted means (above), $(\mathbf{X}^T\mathbf{X})^{-1}$ matrix (below), adjusted means with standard errors (below), and CI's (below):

```
proc glm data=diets;
class diet;
model gain = diet caloric/solution inverse;
means diet;
lsmeans diet/stderr cl;
run;
```

X'X Generalized Inverse (g2)							
	Intercept	diet 1	diet 2	diet 3	diet 4	caloric	gain
Intercept	6.3826300294	-0.938959764	-1.037487733	-0.618743867	0	-0.012315996	-35.66310108
diet 1	-0.938959764	0.4883218842	0.3000981354	0.2500490677	0	0.0014720314	24.295191364
diet 2	-1.037487733	0.3000981354	0.5134445535	0.2567222767	0	0.0016683023	20.441216879
diet 3	-0.618743867	0.2500490677	0.2567222767	0.4283611384	0	0.0008341511	22.12060844
diet 4	0	0	0	0	0	0	0
caloric	-0.012315996	0.0014720314	0.0016683023	0.0008341511	0	0.0000245339	0.1682531894
gain	-35.66310108	24.295191364	20.441216879	22.12060844	0	0.1682531894	1180.5196271

The $(\mathbf{X}^T\mathbf{X})^{-1}$ matrix is found by removing the row/column with 0's and ignoring the row/column for the response.

diet	gain LSMEAN	Standard Error	Pr > |t|
1	66.2809372	4.0588750	<.0001
2	62.4269627	4.1473443	<.0001
3	64.1063543	3.9776677	<.0001
4	41.9857458	4.3482563	<.0001

diet	gain LSMEAN	95% Confidence Limits	
1	66.280937	57.629650	74.932224
2	62.426963	53.587108	71.266818
3	64.106354	55.628156	72.584552
4	41.985746	32.717657	51.253835

We can now look at all pairwise differences of the lsmeans to see which levels differ significantly.

ANCOVA - What did we just do? We used our covariate to reduce the unexplained variation in our response, allowing a clearer picture of our treatment differences.

Huge assumptions of ANCOVA - We assume the treatment *does not* affect the covariate. In this example, we assume the diets are not causing the animals to have different caloric intake (i.e. do not cause them to eat more or less). (We also need to do our usual assumption checking.)

We can inspect this assumption. Let our covariate be our response and conduct an ANOVA using the diets as our treatments. The global p-value will test if the caloric intake means differ significantly for each diet. We hope to see no significance here!

```
proc anova data=diets;
class diet;
model caloric = diet;
run;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	14095.00000	4698.33333	1.84	0.1798
Error	16	40760.00000	2547.50000		
Corrected Total	19	54855.00000			

No evidence that treatment affects covariate.

Chapter 8

ST 512 - Contrasts and Multiple Comparisons

Readings: 9.1-9.5 326-360

Consider the traditional balanced One-Way ANOVA model. That is, we have a *continuous* response, Y, and a *qualitative or categorical* predictor which we call our **factor**. This factor has t *levels* (also the *treatments* in this case) and our interest lies in whether or not the mean response differs between the treatments.

The parametrization of the One-way ANOVA model we have looked at is

$$Y_{ij} = \mu + \tau_i + E_{ij}, \quad i = 1, 2, ...t, \quad j = 1, ..., n_i,$$

$E_{ij} \sim N(0, \sigma^2)$ (balanced implies n_i is the same for all levels). The (true) treatment mean for treatment i is given by $\mu + \tau_i = \mu_i$.

Consider the bovine antibiotic/binding percentage example from earlier. Let

$$\begin{aligned}
\mu_1 &= \mu + \tau_1 = \text{mean of Chloramphenicol treatment} \\
\mu_2 &= \mu + \tau_2 = \text{mean of Erythromycin treatment} \\
\mu_3 &= \mu + \tau_3 = \text{mean of Penicillin G treatment} \\
\mu_4 &= \mu + \tau_4 = \text{mean of Streptomycin treatment} \\
\mu_5 &= \mu + \tau_5 = \text{mean of Tetracyclin treatment}
\end{aligned}$$

There were four observations at each treatment level. Recall the p-value for testing the global hypothesis that

$$H_0 : \tau_1 = \tau_2 = \tau_3 = \tau_4 = \tau_5 = 0 \quad vs \quad H_A : \text{At least 1 differs}$$

which is equivalent to testing

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = 0 \quad vs \quad H_A : \text{At least 1 differs}$$

was < 0.0001 implying we should reject H_0 in favor of H_A. That is, at the 5% significance level there is enough evidence to conclude the mean for at least one of the antibiotics differs.

Given the answer to the previous question, the next logical question to answer is: 'Which treatment means are different?' Suppose we want first inspect the difference between the Cholramphenicol (μ_1) and Erythromycin treatment means (μ_2). In terms of μ_1 and μ_2, how can we write this question as a null and alternative hypotheses?

This is an example of a **linear combination** of treatment means. In general, a linear combination of treatment means takes the form

For the bovine experiment, which of the following are linear combinations of treatment means?

- $\theta_6 = 4\mu_1 + 3\mu_2 - 7\mu_5$
- $\theta_7 = 3\mu_1\mu_4 + 2\mu_3$
- $\theta_8 = \mu_1 + \mu_2 + \mu_3 + \mu_4 + \mu_5$
- $\theta_9 = \mu_1^2 + 3\mu_2 + 1$

If the coefficients of the linear combination sum to zero (i.e. $c_1 + c_2 + ... + c_t = 0$), the linear combination is called a **contrast**.

Is our linear combination $\mu_1 - \mu_2 = 0$ a contrast? How about any of θ_6 through θ_9?

If we want to do inference about a linear combination of treatment means we need an estimator of θ, call it $\hat{\theta}$ and we will also need a measure of variability, say $\hat{SE}(\hat{\theta})$.

An estimator is given by substitution of the sample means:

$$\hat{\theta} = c_1\bar{Y}_{1+} + c_2\bar{Y}_{2+} + ... + c_t\bar{Y}_{t+}$$

Use the output from the code that follows to estimate the following linear combinations:

$\theta_1 = \mu_1 \quad \hat{\theta}_1 = \bar{y}_{1+} =$ _____

$\theta_2 = \mu_2 \quad \hat{\theta}_2 = \bar{y}_{2+} =$ _____

$\theta_3 = \mu_3 \quad \hat{\theta}_3 = \bar{y}_{3+} =$ _____

$\theta_4 = \mu_4 \quad \hat{\theta}_4 = \bar{y}_{4+} =$ _____

$\theta_5 = \mu_5 \quad \hat{\theta}_5 = \bar{y}_{5+} =$ _____

What is the relationship between the $\hat{\beta}$ estimates (from the '/solution' option) and these means again?

```
proc glm data=binding;
class antibiotic;
model bindfrac=antibiotic/solution;
means antibiotic;
run;
```

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	31.37500000	B	1.50456251	20.85	<.0001
Antibiotic Chloramp	-3.57500000	B	2.12777270	-1.68	0.1136
Antibiotic Erythrom	-12.30000000	B	2.12777270	-5.78	<.0001
Antibiotic Penicill	-2.77500000	B	2.12777270	-1.30	0.2118
Antibiotic Streptom	-23.55000000	B	2.12777270	-11.07	<.0001
Antibiotic Tetracyc	0.00000000	B	.	.	.

Level of Antibiotic	N	BindFrac	
		Mean	Std Dev
Chloramp	4	27.8000000	3.98998747
Erythrom	4	19.0750000	1.80623919
Penicill	4	28.6000000	3.21765960
Streptom	4	7.8250000	2.38379949
Tetracyc	4	31.3750000	3.17109340

Find an estimate of our contrast $\theta = \mu_1 - \mu_2$. Find an estimate for θ_6 and θ_8.

We now have a point estimate of the quantity in our null hypothesis, in order to conduct our test we must also know about the variability of this estimate, i.e. What is $\hat{Var}(\hat{\theta})$ or $\hat{SE}(\hat{\theta})$?

The variance of a linear combination of means in One-way ANOVA has a very nice form:

Due to the normality assumption on our errors we can use a t-test. Let θ_0 be a value of interest for our contrast (often 0). To test $H_0 : \theta = \theta_0$ vs $H_A : \theta \neq \theta_0$ we can use

$$t = \frac{\hat{\theta} - \theta_0}{\hat{SE}(\hat{\theta})} \sim t_{t(n-1)} \text{ under } H_0$$

What is the value of this test statistic for our contrast $\theta = \mu_1 - \mu_2$? Compare it to $t(0.975, 15) = 2.13$, what is your conclusion? What is your interpretation?

Likewise a confidence interval can be formed using

$$\hat{\theta} \pm t(\alpha/2, t(n-1))\hat{SE}(\hat{\theta}) = \sum_{i=1}^{t} c_i \bar{y}_{i+} \pm t(\alpha/2, t(n-1))\sqrt{MS(E)\sum_{i=1}^{t}\frac{c_i^2}{n_i}}$$

What is a 95% CI for $\theta = \mu_1 - \mu_2$? Does your conclusion here match the conclusion using the test statistic?

Note: A contrast that has only two nonzero c's is called a pairwise contrast (as it looks at only two means). These can be had easily in proc glm using the code below.

A **complex** contrast is a contrast that involves more than two non-zero coefficients. For example, $\theta_{10} = \frac{\mu_1+\mu_2}{2} - \frac{\mu_3+\mu_4}{2}$ is a complex contrast.

```
proc glm data=binding;
class antibiotic;
model bindfrac=antibiotic/solution;
means antibiotic/ lsd cldiff lines;
lsmeans antibiotic/stderr pdiff;
run;
```

*Generally we'll want to use lsmeans not means, but ok here since no covariates involved and a balanced design was done.

The GLM Procedure

t Tests (LSD) for BindFrac

Note: This test controls the Type I comparisonwise error rate, not the experimentwise

Alpha	0.05
Error Degrees of Freedom	15
Error Mean Square	9.054833
Critical Value of t	2.13145
Least Significant Difference	4.5352

Comparisons significant at the 0.05 level are indicated by ***.				
Antibiotic Comparison	**Difference Between Means**	**95% Confidence Limits**		
Tetracyc - Penicill	2.775	-1.760	7.310	
Tetracyc - Chloramp	3.575	-0.960	8.110	
Tetracyc - Erythrom	12.300	7.765	16.835	***
Tetracyc - Streptom	23.550	19.015	28.085	***
Penicill - Tetracyc	-2.775	-7.310	1.760	
Penicill - Chloramp	0.800	-3.735	5.335	
Penicill - Erythrom	9.525	4.990	14.060	***
Penicill - Streptom	20.775	16.240	25.310	***
Chloramp - Tetracyc	-3.575	-8.110	0.960	
Chloramp - Penicill	-0.800	-5.335	3.735	
Chloramp - Erythrom	8.725	4.190	13.260	***
Chloramp - Streptom	19.975	15.440	24.510	***
Erythrom - Tetracyc	-12.300	-16.835	-7.765	***
Erythrom - Penicill	-9.525	-14.060	-4.990	***
Erythrom - Chloramp	-8.725	-13.260	-4.190	***
Erythrom - Streptom	11.250	6.715	15.785	***
Streptom - Tetracyc	-23.550	-28.085	-19.015	***
Streptom - Penicill	-20.775	-25.310	-16.240	***
Streptom - Chloramp	-19.975	-24.510	-15.440	***
Streptom - Erythrom	-11.250	-15.785	-6.715	***

The SAS System 12:34 Sunday, February 16,

The GLM Procedure

t Tests (LSD) for BindFrac

Note: This test controls the Type I comparisonwise error rate, not the experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	15
Error Mean Square	9.054833
Critical Value of t	2.13145
Least Significant Difference	4.5352

Means with the same letter are not significantly different.			
t Grouping	**Mean**	**N**	**Antibiotic**
A	31.375	4	Tetracyc
A			
A	28.600	4	Penicill
A			
A	27.800	4	Chloramp
B	19.075	4	Erythrom
C	7.825	4	Streptom

The GLM Procedure
Least Squares Means

Antibiotic	BindFrac LSMEAN	Standard Error	Pr > |t|	LSMEAN Number
Chloramp	27.8000000	1.5045625	<.0001	1
Erythrom	19.0750000	1.5045625	<.0001	2
Penicill	28.6000000	1.5045625	<.0001	3
Streptom	7.8250000	1.5045625	0.0001	4
Tetracyc	31.3750000	1.5045625	<.0001	5

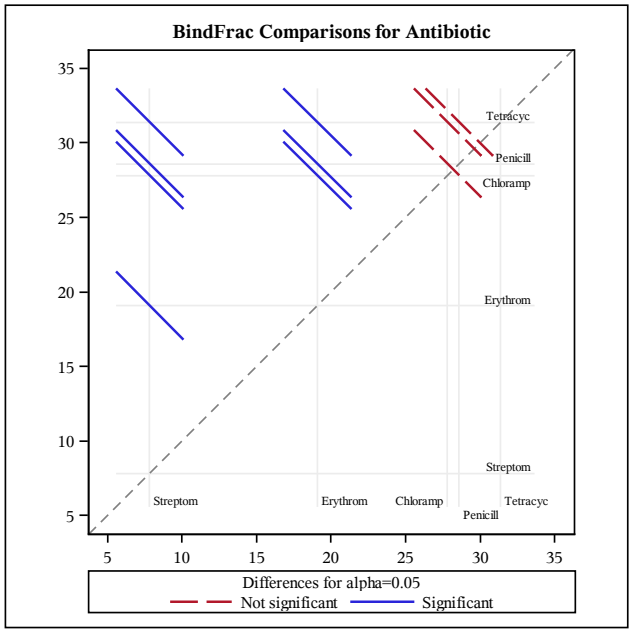
Least Squares Means for effect Antibiotic
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: BindFrac

i/j	1	2	3	4	5
1		0.0009	0.7122	<.0001	0.1136
2	0.0009		0.0004	<.0001	<.0001
3	0.7122	0.0004		<.0001	0.2118
4	<.0001	<.0001	<.0001		<.0001
5	0.1136	<.0001	0.2118	<.0001	

The SAS System 12:34 Sunday, February 16,

The GLM Procedure
Least Squares Means



Note: To ensure overall protection level, only probabilities associated with pre-planned comparisons should be used.

Use the output to construct a 95% CI for θ_{10}.

To get the estimate for θ_{10} (and a few other linear combinations of means) in SAS we can use glm but it will be easiest to use proc mixed:

```
proc glm data=binding;
class antibiotic;
model bindfrac=antibiotic/clparm;
estimate 'lsmean for trt 2' intercept 1 antibiotic 0 1 0 0 0;
estimate 'avg of trt 2 mean and trt 3 mean' intercept 2 antibiotic 0 1 1 0 0/divisor=2;
estimate 'trt 1 vs trt 5' intercept 0 antibiotic 1 0 0 0 -1;
estimate 'avg of 1 and 2 vs avg of 3 and 4' intercept 0 antibiotic 1 1 -1 -1 0/divisor=2;
run;

proc mixed data=binding;
class antibiotic;
model bindfrac=antibiotic;
lsmestimate antibiotic 'lsmean for trt 2' [1,2]/cl;
lsmestimate antibiotic 'avg of trt 2 mean and trt 3 mean' [0.5,2] [0.5,3]/cl;
lsmestimate antibiotic 'trt 1 vs trt 5' [1,1] [-1,5]/cl;
lsmestimate antibiotic 'avg of 1 and 2 vs avg of 3 and 4' [1,1] [1,2] [-1,3] [-1,4]/divisor=2 cl;
run;
```

Least Squares Means Estimate

Effect	Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Antibiotic	lsmean for trt 2	19.0750	1.5046	15	12.68	<.0001	0.05	15.8681	22.2819

Least Squares Means Estimate

Effect	Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Antibiotic	avg of trt 2 mean and trt 3 mean	23.8375	1.0639	15	22.41	<.0001	0.05	21.5699	26.1051

Least Squares Means Estimate

Effect	Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Antibiotic	trt 1 vs trt 5	-3.5750	2.1278	15	-1.68	0.1136	0.05	-8.1102	0.9602

Least Squares Means Estimate

Effect	Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Antibiotic	avg of 1 and 2 vs avg of 3 and 4	5.2250	1.5046	15	3.47	0.0034	0.05	2.0181	8.4319

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
lsmean for trt 2	19.0750000	1.50456251	12.68	<.0001	15.8681009	22.2818991
avg of trt 2 mean and trt 3 mean	23.8375000	1.06388635	22.41	<.0001	21.5698799	26.1051201
trt 1 vs trt 5	-3.5750000	2.12777270	-1.68	0.1136	-8.1102402	0.9602402
avg of 1 and 2 vs avg of 3 and 4	5.2250000	1.50456251	3.47	0.0034	2.0181009	8.4318991

Multiple Comparisons Corrections

It is not safe to go carrying out many many significance tests suggested by the data all willy-nilly. If we do, our *experiment wide* type I error rate will not be controlled.

Recall: $\alpha = P(\text{Type I Error})$

Decision	H_0 true	H_0 false
Reject H_0	Type I Error	Correct!
Fail to Reject H_0	Correct!	Type II Error

For a given test, we fix the probability of a type I error to be small (often 0.05) as it is usually considered worse than a type II error.

Consider the case with $t = 5$ (antibiotic treatments):

- the number of pairwise contrasts of the form $\theta = \mu_i - \mu_j$ is $\begin{pmatrix} 5 \\ 2 \end{pmatrix} = 10$
- each test has type I error $\alpha = 0.05$, but overall what is our type I error rate?

  i.e. P(rejecting at least one null hypothesis that is true)

- This is called the *experimentwise* or *familywise* (fwe) type I error rate. We should really control this instead of the type I error for each test!

Example: Is a certain type of coin fair (equal probability of flipping a head and a tail)?

$$H_0 : \text{Coin fair}, p = 0.5 \qquad H_A : \text{Coin biased}, p \neq 0.5$$

Experiment - flip one of these coins 10 times, if 9 or 10 heads appear or if 9 or 10 tails appear then declare coin biased.

Assuming the coin is fair,

$$\begin{aligned} \alpha &= P(\text{Concluding coin is biased}) \\ &= P(9\ heads) + P(9\ tails) + P(10\ heads) + P(10\ tails) \\ &= 2 * 10(1/2)^{10} + 2 * (1/2)^{10} = 0.021 \end{aligned}$$

This is our type I error rate for testing this particular coin (a little smaller than the usual 0.05).

Now suppose we have 100 coins of this type and we test each in the same manner. If the coins were truly fair, how many of the experiments would we expect to conclude we have a biased coin?

For a particular coin to come up heads or tails 9-10 times is very unlikely, but seeing any 1 coin of the 100 behave this way would be more likely than not.
In fact,

$$P(\text{All 100 coins identified as fair}) = 0.34$$

$$P(\text{At least 1 coin of the 100 is classsified as biased}) = 0.66$$

This is why we need to control the fwe rate when we do many data-driven comparisons!

Comparisons can be categorized as *a priori* or *post-hoc*:

- *A priori*: Significance tests which will be carried out without regard to the observed outcome of the experiment.
- *Post-hoc* or data-driven: Significance tests which are suggested by the observed outcome of the experiment.

Methods for simultaneous inference for multiple comparisons include (but there are many many of these)

- Bonferroni
- Scheffé
- Tukey

Bonferroni Correction

Suppose interest lies in exactly k contrasts. **Bonferroni correction** is to replace the usual α with

$$\alpha' = \frac{\alpha}{k}$$

By doing so our fwe rate will be less than α

$$\text{fwe rate} = \alpha^* = 1 - (1-\alpha')^k = 1 - (1 - \frac{\alpha}{k})^k \leq \alpha$$

We can now create **simultaneous** CIs. These are a group of CIs we are $(1-\alpha)\%$ confident will <u>all</u> contain their true value.

Simultaneous 95% confidence intervals for the k contrasts given by

$$a_1\bar{y}_{1+} + a_2\bar{y}_{2+} + \cdots + a_t\bar{y}_{t+} \pm t(\alpha'/2, t(n-1))\sqrt{MS(E)\sum_{i=1}^{t}\frac{a_i^2}{n_i}}$$

$$b_1\bar{y}_{1+} + b_2\bar{y}_{2+} + \cdots + b_t\bar{y}_{t+} \pm t(\alpha'/2, t(n-1))\sqrt{MS(E)\sum_{i=1}^{t}\frac{a_i^2}{n_i}}$$

$$\vdots$$

$$k_1\bar{y}_{1+} + k_2\bar{y}_{2+} + \cdots + k_t\bar{y}_{t+} \pm t(\alpha'/2, t(n-1))\sqrt{MS(E)\sum_{i=1}^{t}\frac{a_i^2}{n_i}}$$

Note: $t(\alpha'/2, t(n-1))$ might have to be obtained using software.

For the binding fraction example, consider only pairwise comparisons with Chloramphenicol (μ_1):

$$\theta_1 = \mu_1 - \mu_2, \quad \theta_2 = \mu_1 - \mu_3, \quad \theta_3 = \mu_1 - \mu_4, \quad \theta_4 = \mu_1 - \mu_5$$

We have $k = 4$, $\alpha' = 0.05/k = 0.0125$, and $t(\alpha'/2, 15) = 2.84$.

What is the Margin of Error for one of these contrasts? Find the simultaneous 95% intervals for the four contrasts. Which of these antibiotic means differ significantly from the chloramphenicol mean?

Scheffé Correction

Scheffé correction scales the t multiplier in an interesting way.

$$\text{Rather than } t(\alpha/2, t(n-1)) \text{ we use } \sqrt{(t-1)F(\alpha, t-1, t(n-1))}$$

**Doesn't depend on the number of contrasts done!

For **simultaneous** 95% confidence intervals for **any number** of contrasts, use

$$\sum_{i=1}^{t} c_i \bar{y}_{i+} \pm \sqrt{(t-1)F(\alpha, t-1, t(n-1))MS[E]\sum_{i=1}^{t}\frac{c_i^2}{n_i}}$$

For a pairwise comparisons of means, say $\mu_1 - \mu_2$, this yields

$$\bar{y}_{1+} - \bar{y}_{2+} \pm \sqrt{(t-1)F(\alpha, t-1, t(n-1))MS[E](1/n_1 + 1/n_2)}$$

For binding fraction data, what is the Margin of Error for testing one of the contrasts? (Note: $F(\alpha, t-1, t(n-1)) = 3.06$)

Tukey-Kramer Correction (or just Tukey)

Tukey's correction is the best method when making inference on **all pairwise** comparisons in balanced designs. That is, for simple contrasts of the form

$$\theta = \mu_j - \mu_k$$

it will tend to have a lower type II error rate in these cases than Scheffé and bonferroni corrections. (It has a greater chance of detecting differences i.e. is more powerful.)

- Uses multipliers from a distribution called the 'studentized range distribution'
- Denoted $q(\alpha, t, t(n-1))$

For **simultaneous** 95% confidence intervals for $\theta = \mu_j - \mu_k$ use

$$\begin{aligned}
\hat{\theta} &\pm \frac{q(\alpha, t, t(n-1))}{\sqrt{2}} \hat{SE}(\hat{\theta}) \\
\bar{y}_{j+} - \bar{y}_{k+} &\pm \frac{q(\alpha, t, t(n-1))}{\sqrt{2}} \sqrt{MS(E)(\frac{1}{n} + \frac{1}{n})} \\
\bar{y}_{j+} - \bar{y}_{k+} &\pm q(\alpha, t, t(n-1)) \sqrt{\frac{MS(E)}{n}}
\end{aligned}$$

How to do these multiple comparison corrections in SAS?

- Bonferonni can be done by manually changing the α level SAS uses. For the example above, $\alpha' = 0.05/4 = 0.0125$:

```
proc glm data=binding;
class antibiotic;
model bindfrac=antibiotic/clparm alpha=0.0125;
*Can drop intercept since it has 0 coefficient;
estimate 'theta 1' antibiotic 1 -1;
estimate 'theta 2' antibiotic 1 0 -1;
estimate 'theta 3' antibiotic 1 0 0 -1 0;
estimate 'theta 4' antibiotic 1 0 0 0 -1;
run;

proc mixed data=binding;
class antibiotic;
model bindfrac=antibiotic;
lsmestimate antibiotic 'theta 1' [1,1] [-1,2],
 'theta 2' [1,1] [-1,3],
 'theta 3' [1,1] [-1,4],
 'theta 4' [1,1] [-1,5]/cl adjust=bon;
run;
```

- Scheffe and Tukey corrections are options:

```
proc glm data=binding;
class antibiotic;
model bindfrac=antibiotic;
lsmeans antibiotic/pdiff adjust=scheffe cl;
lsmeans antibiotic/pdiff adjust=tukey cl;
run;

proc mixed data=binding;
class antibiotic;
model bindfrac=antibiotic;
lsmeans antibiotic/pdiff adjust=tukey cl;
lsmestimate antibiotic 'theta 1' [1,1] [-1,2],
  'theta 2' [1,1] [-1,3],
  'theta 3' [1,1] [-1,4],
  'theta 4' [1,1] [-1,5]/cl adjust=scheffe;
run;
```

Bonferroni GLM output:

Parameter	Estimate	Standard Error	t Value	Pr > |t|	98.75% Confidence Limits	
theta 1	8.7250000	2.12777270	4.10	0.0009	2.6893015	14.7606985
theta 2	-0.8000000	2.12777270	-0.38	0.7122	-6.8356985	5.2356985
theta 3	19.9750000	2.12777270	9.39	<.0001	13.9393015	26.0106985
theta 4	-3.5750000	2.12777270	-1.68	0.1136	-9.6106985	2.4606985

Bonferroni Mixed output:

Least Squares Means Estimates Adjustment for Multiplicity: Bonferroni												
Effect	Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P	Alpha	Lower	Upper	Adj Lower	Adj Upper
Antibiotic	theta 1	8.7250	2.1278	15	4.10	0.0009	0.0038	0.05	4.1898	13.2602	2.6893	14.7607
Antibiotic	theta 2	-0.8000	2.1278	15	-0.38	0.7122	1.0000	0.05	-5.3352	3.7352	-6.8357	5.2357
Antibiotic	theta 3	19.9750	2.1278	15	9.39	<.0001	<.0001	0.05	15.4398	24.5102	13.9393	26.0107
Antibiotic	theta 4	-3.5750	2.1278	15	-1.68	0.1136	0.4545	0.05	-8.1102	0.9602	-9.6107	2.4607

Scheffe and Tukey GLM output:

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Scheffe

Antibiotic	BindFrac LSMEAN	LSMEAN Number
Chloramp	27.8000000	1
Erythrom	19.0750000	2
Penicill	28.6000000	3
Streptom	7.8250000	4
Tetracyc	31.3750000	5

Least Squares Means for effect Antibiotic
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: BindFrac

i/j	1	2	3	4	5
1		0.0177	0.9973	<.0001	0.6003
2	0.0177		0.0091	0.0022	0.0009
3	0.9973	0.0091		<.0001	0.7881
4	<.0001	0.0022	<.0001		<.0001
5	0.6003	0.0009	0.7881	<.0001	

Antibiotic	BindFrac LSMEAN	95% Confidence Limits	
Chloramp	27.800000	24.593101	31.006899
Erythrom	19.075000	15.868101	22.281899
Penicill	28.600000	25.393101	31.806899
Streptom	7.825000	4.618101	11.031899
Tetracyc	31.375000	28.168101	34.581899

Least Squares Means for Effect Antibiotic

i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	8.725000	1.286228	16.163772
1	3	-0.800000	-8.238772	6.638772
1	4	19.975000	12.536228	27.413772
1	5	-3.575000	-11.013772	3.863772
2	3	-9.525000	-16.963772	-2.086228
2	4	11.250000	3.811228	18.688772
2	5	-12.300000	-19.738772	-4.861228
3	4	20.775000	13.336228	28.213772
3	5	-2.775000	-10.213772	4.663772
4	5	-23.550000	-30.988772	-16.111228

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey

Antibiotic	BindFrac LSMEAN	LSMEAN Number
Chloramp	27.8000000	1
Erythrom	19.0750000	2
Penicill	28.6000000	3
Streptom	7.8250000	4
Tetracyc	31.3750000	5

Least Squares Means for effect Antibiotic
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: BindFrac

i/j	1	2	3	4	5
1		0.0072	0.9953	<.0001	0.4738
2	0.0072		0.0035	0.0007	0.0003
3	0.9953	0.0035		<.0001	0.6928
4	<.0001	0.0007	<.0001		<.0001
5	0.4738	0.0003	0.6928	<.0001	

Antibiotic	BindFrac LSMEAN	95% Confidence Limits	
Chloramp	27.800000	24.593101	31.006899
Erythrom	19.075000	15.868101	22.281899
Penicill	28.600000	25.393101	31.806899
Streptom	7.825000	4.618101	11.031899
Tetracyc	31.375000	28.168101	34.581899

Least Squares Means for Effect Antibiotic

i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	8.725000	2.154598	15.295402
1	3	-0.800000	-7.370402	5.770402
1	4	19.975000	13.404598	26.545402
1	5	-3.575000	-10.145402	2.995402
2	3	-9.525000	-16.095402	-2.954598
2	4	11.250000	4.679598	17.820402
2	5	-12.300000	-18.870402	-5.729598
3	4	20.775000	14.204598	27.345402
3	5	-2.775000	-9.345402	3.795402
4	5	-23.550000	-30.120402	-16.979598

Scheffe and Tukey Mixed output:

Least Squares Means Estimates
Adjustment for Multiplicity: Scheffe

Effect	Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P	Alpha	Lower	Upper	Adj Lower	Adj Upper
Antibiotic	theta 1	8.7250	2.1278	15	4.10	0.0009	0.0177	0.05	4.1898	13.2602	1.2862	16.1638
Antibiotic	theta 2	-0.8000	2.1278	15	-0.38	0.7122	0.9973	0.05	-5.3352	3.7352	-8.2388	6.6388
Antibiotic	theta 3	19.9750	2.1278	15	9.39	<.0001	<.0001	0.05	15.4398	24.5102	12.5362	27.4138
Antibiotic	theta 4	-3.5750	2.1278	15	-1.68	0.1136	0.6003	0.05	-8.1102	0.9602	-11.0138	3.8638

Least Squares Means

Effect	Antibiotic	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Antibiotic	Chloramp	27.8000	1.5046	15	18.48	<.0001	0.05	24.5931	31.0069
Antibiotic	Erythrom	19.0750	1.5046	15	12.68	<.0001	0.05	15.8681	22.2819
Antibiotic	Penicill	28.6000	1.5046	15	19.01	<.0001	0.05	25.3931	31.8069
Antibiotic	Streptom	7.8250	1.5046	15	5.20	0.0001	0.05	4.6181	11.0319
Antibiotic	Tetracyc	31.3750	1.5046	15	20.85	<.0001	0.05	28.1681	34.5819

Differences of Least Squares Means

Effect	Antibiotic	_Antibiotic	Estimate	Standard Error	DF	t Value	Pr > |t|	Adjustment	Adj P	Alpha	Lower	Upper	Adj Lower	Adj Upper
Antibiotic	Chloramp	Erythrom	8.7250	2.1278	15	4.10	0.0009	Tukey	0.0072	0.05	4.1898	13.2602	2.1546	15.2954
Antibiotic	Chloramp	Penicill	-0.8000	2.1278	15	-0.38	0.7122	Tukey	0.9953	0.05	-5.3352	3.7352	-7.3704	5.7704
Antibiotic	Chloramp	Streptom	19.9750	2.1278	15	9.39	<.0001	Tukey	<.0001	0.05	15.4398	24.5102	13.4046	26.5454
Antibiotic	Chloramp	Tetracyc	-3.5750	2.1278	15	-1.68	0.1136	Tukey	0.4738	0.05	-8.1102	0.9602	-10.1454	2.9954
Antibiotic	Erythrom	Penicill	-9.5250	2.1278	15	-4.48	0.0004	Tukey	0.0035	0.05	-14.0602	-4.9898	-16.0954	-2.9546
Antibiotic	Erythrom	Streptom	11.2500	2.1278	15	5.29	<.0001	Tukey	0.0007	0.05	6.7148	15.7852	4.6796	17.8204
Antibiotic	Erythrom	Tetracyc	-12.3000	2.1278	15	-5.78	<.0001	Tukey	0.0003	0.05	-16.8352	-7.7648	-18.8704	-5.7296
Antibiotic	Penicill	Streptom	20.7750	2.1278	15	9.76	<.0001	Tukey	<.0001	0.05	16.2398	25.3102	14.2046	27.3454
Antibiotic	Penicill	Tetracyc	-2.7750	2.1278	15	-1.30	0.2118	Tukey	0.6928	0.05	-7.3102	1.7602	-9.3454	3.7954
Antibiotic	Streptom	Tetracyc	-23.5500	2.1278	15	-11.07	<.0001	Tukey	<.0001	0.05	-28.0852	-19.0148	-30.1204	-16.9796

Independent Contrasts

Consider a contrast θ, then

$$\theta = c_1\mu_1 + c_2\mu_2 + ... + c_t\mu_t$$

where $\sum_{i=1}^{t} c_i = 0$. The estimate of a contrast is

$$\hat{\theta} = c_1\bar{y}_{1+} + c_2\bar{y}_{2+} + ... + c_t\bar{y}_{t+}$$

and the estimated variance is given by

$$\hat{Var}(\hat{\theta}) = MS(E)\sum_{i=1}^{t}\frac{c_i^2}{n_i}$$

Recall: The idea behind ANOVA is that we partition $SS(TOT)$ into independent components $SS(Trt)$ and $SS[E]$.

Similarly, we can take $SS(Trt)$ and partition it into $t-1$ independent contrasts each with 1 df.

Orthogonal contrasts:
Let

$$\theta_1 = \sum_{i=1}^{t} c_i\mu_i \text{ and } \theta_2 = \sum_{i=1}^{t} d_i\mu_i$$

be two contrasts. θ_1 and θ_2 are **orthogonal** if

$$c_1d_1 + c_2d_2 + ... + c_td_t = \sum_{i=1}^{t} c_id_i = 0$$

A set of k contrasts is mutually orthogonal if all pairs are orthogonal.

Examples:
$(-1, 1, 0, 0, 0)$ and $(0, 0, -1, 1, 0)$ orthogonal ?

$(1, -1/2, -1/2, 0, 0)$ and $(0, 0, 0, -1, 1)$ orthogonal ?

$(-1, 1, 0, 0, 0)$ and $(0, -1, 1, 0, 0)$ orthogonal ?

Due to the joint normality, **orthogonality implies independence**!

Sums of squares for contrasts
Recall we are going to partition $SS(Trt)$ into $t-1$ independent contrasts. The sums of squares for a contrast are

$$SS(\hat{\theta}) = \frac{\hat{\theta}^2}{\left(\frac{c_1^2}{n_1} + \cdots + \frac{c_t^2}{n_t}\right)} = \frac{\hat{\theta}^2}{\left(\sum_{i=1}^{t} \frac{c_i^2}{n_i}\right)}$$

This contrast has 1 df associated with it.

We can define $MS(\hat{\theta}) = SS(\hat{\theta})/1 = SS(\hat{\theta})$ and can then test

$$H_0 : \theta = 0 \quad vs \quad H_A : \theta \neq 0$$

using the F-statistic

$$F = \frac{MS(\hat{\theta})}{MS(E)} \sim F_{1,t(n-1)}$$

Compare this to the t-test done earlier for testing a contrast

$$t = \frac{\hat{\theta} - \theta_0}{\hat{SE}(\hat{\theta})} \sim t_{t(n-1)}$$

(Remember if we square a t stat we get an F stat!)

We can also test multiple contrasts all equal to 0 at once

$$H_0 : \theta_1 = \theta_2 = ... = \theta_k = 0 \quad vs \quad H_A : \text{At least 1 } \theta \neq 0$$

using the F-statistic

$$F = \frac{\frac{SS(\hat{\theta_1}) + SS(\hat{\theta_2}) + ... + SS(\hat{\theta_k})}{k}}{MS(E)} \sim F_{k,t(n-1)}$$

How to relate this to $SS(Trt)$? Generally, if $\theta_1, \theta_2, ..., \theta_{t-1}$ are $t-1$ mutually orthogonal contrasts then

$$SS(Trt) = SS(\hat{\theta}_1) + SS(\hat{\theta}_2) + \cdots + SS(\hat{\theta}_{t-1})$$

and $df_{Trt} = df_{\hat{\theta}_1} + ... + df_{\hat{\theta}_{t-1}} = 1 + ... + 1 = t - 1$

Notice, testing all $t-1$ contrasts equal to 0 is equivalent to testing our global F-test!

Again consider the Binding Fraction data. In this case we have $5 - 1 = 4$ df for treatment. Consider the following set of 4 mutually orthogonal contrasts:

$$\begin{aligned}
\theta_1 &= (-2 \quad -1 \quad 0 \quad 1 \quad 2) \\
\theta_2 &= (2 \quad -1 \quad -2 \quad -1 \quad 2) \\
\theta_3 &= (-1 \quad 2 \quad 0 \quad -2 \quad 1) \\
\theta_4 &= (1 \quad -4 \quad 6 \quad -4 \quad 1)
\end{aligned}$$

Since these are mutually orthogonal, they are all independent.

Let's use SAS to get estimates. Test for $\theta_4 = 0$ using both the t and F tests. Then show that $SS(Trt) = SS(\theta_1) + SS(\theta_2) + SS(\theta_3) + SS(\theta_4)$ and conduct the global F test.

```
proc glm data=binding; class antibiotic; model bindfrac=antibiotic;
contrast 'theta 1' antibiotic -2 -1 0 1 2;
contrast 'theta 2' antibiotic 2 -1 -2 -1 2;
contrast 'theta 3' antibiotic -1 2 0 -2 1;
contrast 'theta 4' antibiotic 1 -4 6 -4 1; run;

proc mixed data=binding; class antibiotic; model bindfrac=antibiotic;
lsmestimate antibiotic 'theta 1' [-2,1] [-1,2] [0,3] [1,4] [2,5],
'theta 2' [2,1] [-1,2] [-2,3] [-1,4] [2,5],
'theta 3' [-1,1] [2,2] [0,3] [-2,4] [1,5],
'theta 4' [1,1] [-4,2] [6,3] [-4,4] [1,5]; run;
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Antibiotic	4	1480.823000	370.205750	40.88	<.0001

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
theta 1	1	6.7240000	6.7240000	0.74	0.4024
theta 2	1	335.1607143	335.1607143	37.01	<.0001
theta 3	1	271.9622500	271.9622500	30.04	<.0001
theta 4	1	866.9760357	866.9760357	95.75	<.0001

Least Squares Means Estimates						
Effect	Label	Estimate	Standard Error	DF	t Value	Pr > |t|
Antibiotic	theta 1	-4.1000	4.7578	15	-0.86	0.4024
Antibiotic	theta 2	34.2500	5.6296	15	6.08	<.0001
Antibiotic	theta 3	26.0750	4.7578	15	5.48	<.0001
Antibiotic	theta 4	123.18	12.5881	15	9.79	<.0001

Consider a new dataset: Data consists of the number of contaminants in IV fluids made by $t = 3$ pharmaceutical companies

	Cutter	Abbott	McGaw
	255	105	577
	264	288	515
	342	98	214
	331	275	413
	234	221	401
	217	240	260
$\bar{y}_{i+}$	273.8	204.5	396.7

Source	d.f.	Sum of squares	Mean Square	F
Treatments (or pharmacies)	2	113646	56823	5.81
Error	15	146753	9784	
Total	17	260400		

Consider the following 2 contrasts:

$$\theta_1 = \mu_M - \mu_A \quad \text{and} \quad \theta_2 = \mu_C - \frac{\mu_M + \mu_A}{2}$$

Which levels of the factor will each of these be in SAS? Rewrite these contrasts in terms of μ_1, μ_2, and μ_3.

Are these contrasts orthogonal?

Are the estimated contrasts $\hat{\theta}_1$ and $\hat{\theta}_2$ independent?

Use the output to compute $SS(\hat{\theta}_1)$ and $SS(\hat{\theta}_2)$. What should these add up to and why?

```
proc glm data=pharm; class company; model contam=company;
contrast 'McGaw vs Abbot' company -1 0 1;
estimate 'McGaw vs Abbot'  company -1 0 1;
contrast 'Cutter vs avg of McGaw and Abbot' company -1 2 -1;
estimate 'Cutter vs avg of McGaw and Abbot' company -1 2 -1/divisor=2; run;
```

Note: We should really do a multiple comparison correction for our two contrasts. Bonferonni is easiest, compare our p-values to $0.05/2 = 0.025$.

The GLM Procedure

Class Level Information		
Class	Levels	Values
Company	3	Abbott Cutter McGaw

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	113646.3333	56823.1667	5.81	0.0136
Error	15	146753.6667	9783.5778		
Corrected Total	17	260400.0000			

R-Square	Coeff Var	Root MSE	contam Mean
0.436430	33.91268	98.91197	291.6667

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Company	2	113646.3333	56823.1667	5.81	0.0136

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Company	2	113646.3333	56823.1667	5.81	0.0136

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
McGaw vs Abbot	1	110784.0833	110784.0833	11.32	0.0043
Cutter vs avg of McGaw and Abbot	1	2862.2500	2862.2500	0.29	0.5965

Parameter	Estimate	Standard Error	t Value	Pr > |t|
McGaw vs Abbot	192.166667	57.1068524	3.37	0.0043
Cutter vs avg of McGaw and Abbot	-26.750000	49.4559849	-0.54	0.5965

Company	contam LSMEAN
Abbott	204.500000
Cutter	273.833333
McGaw	396.666667

Polynomial contrasts

For one-way ANOVA, if the factor is actually on the interval scale but observed at only a few levels we can test for polynomial relationships.

With t levels, we can fit any polynomial of degree $t-1$ or less. (Polynomial of degree $t-1$ is equivalent to fitting ANOVA model.) That is, every polynomial model of degree $t-2$ or less is nested in the full ANOVA model!

Example: A poultry science experiment measures bodyweights of chickens from $t = 4$ diet groups. Diets are characterized by protein concentration in diet.

- Response Y = 21-day bodyweights of chickens
- A balanced CRD was done with diet and $N = 72$ total chickens. (Implying $n = 18$).

Experiment Summary:

diet group	x: level of protein	diet mean $\bar{y}_{i+}$
1	21.8	994.9
2	23.5	1000.6
3	25.2	1025.8
4	26.9	1056.0

Here we can see that our factor is actually on an interval scale but measured at only 4 levels.

Consider the One-way ANOVA model using diet and the Linear Regression model cubic in protein:

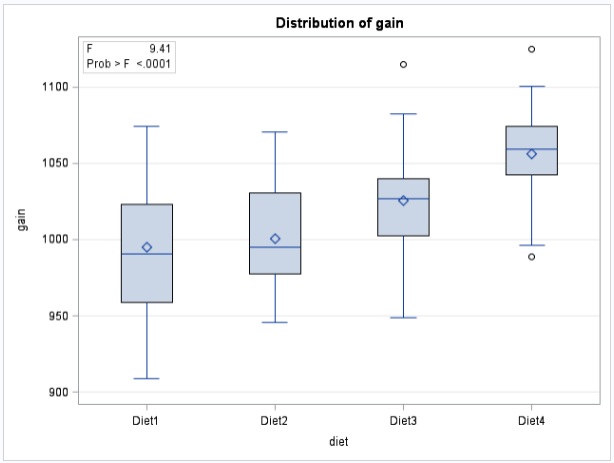
```
proc glm data=chickens; class diet;
model gain=diet; run;

proc glm data=chickens;
model gain=protein protein*protein protein*protein*protein; run;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	42020.4921	14006.8307	9.41	<.0001
Error	68	101214.9702	1488.4554		
Corrected Total	71	143235.4623			

R-Square	Coeff Var	Root MSE	gain Mean
0.293367	3.785046	38.58051	1019.288

Source	DF	Type I SS	Mean Square	F Value	Pr > F
diet	3	42020.49206	14006.83069	9.41	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
diet	3	42020.49206	14006.83069	9.41	<.0001



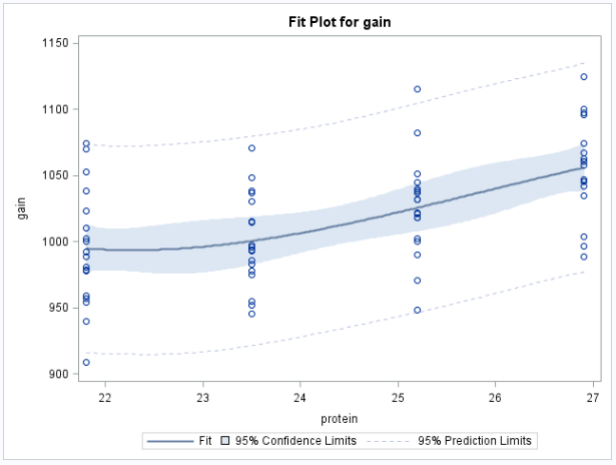
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	42020.4921	14006.8307	9.41	<.0001
Error	68	101214.9702	1488.4554		
Corrected Total	71	143235.4623			

R-Square	Coeff Var	Root MSE	gain Mean
0.293367	3.785046	38.58051	1019.288

Source	DF	Type I SS	Mean Square	F Value	Pr > F
protein	1	39129.40362	39129.40362	26.29	<.0001
protein*protein	1	2700.51253	2700.51253	1.81	0.1825
protei*protei*protei	1	190.57591	190.57591	0.13	0.7216

Source	DF	Type III SS	Mean Square	F Value	Pr > F
protein	1	232.0108148	232.0108148	0.16	0.6942
protein*protein	1	213.5806599	213.5806599	0.14	0.7060
protei*protei*protei	1	190.5759142	190.5759142	0.13	0.7216

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	9025.320090	19740.84174	0.46	0.6490
protein	-966.091671	2446.98971	-0.39	0.6942
protein*protein	38.179905	100.79101	0.38	0.7060
protei*protei*protei	-0.493645	1.37959	-0.36	0.7216



Investigate the two ANOVA tables. Also inspect the Type I Sums of Squares for the MLR model, these would be useful in model building! We can test for the adequacy of a model linear or quadratic in protein rather than the full ANOVA (equivalent to the cubic model).

If we have equally spaced levels (differences between levels all the same), we can actually write down the different polynomial parts in terms of contrasts. Below gives the contrasts you would use for equally spaced levels for 3, 4, or 5 levels.

Factor	Poly.		Coefficients for					
levels	Degree	contrast	$\bar{y}_{1+}$	$\bar{y}_{2+}$	$\bar{y}_{3+}$	$\bar{y}_{4+}$	$\bar{y}_{5+}$	$SS(\hat{\theta}_i)$
3	1	$\hat{\theta}_1$	-1	0	1			$R(\beta_1\|\beta_0)$
	2	$\hat{\theta}_2$	1	-2	1			$R(\beta_2\|\beta_0,\beta_1)$
4	1	$\hat{\theta}_1$	-3	-1	1	3		$R(\beta_1\|\beta_0)$
	2	$\hat{\theta}_2$	1	-1	-1	1		$R(\beta_2\|\beta_0,\beta_1)$
	3	$\hat{\theta}_3$	-1	3	-3	1		$R(\beta_3\|\beta_0,\beta_1,\beta_2)$
5	1	$\hat{\theta}_1$	-2	-1	0	1	2	$R(\beta_1\|\beta_0)$
	2	$\hat{\theta}_2$	2	-1	-2	-1	2	$R(\beta_2\|\beta_0,\beta_1)$
	3	$\hat{\theta}_3$	-1	2	0	-2	1	$R(\beta_3\|\beta_0,\beta_1,\beta_2)$
	4	$\hat{\theta}_4$	1	-4	6	-4	1	$R(\beta_4\|\beta_0,\beta_1,\beta_2,\beta_3)$

Rightmost column indicates extra SS in MLR of the form

$$\mu(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots .$$

The contrast corresponding to a polynomial of degree p can be used to test for a p^{th} degree association:

- large $|\hat{\theta}_1|$ indicates linear association between y and x.
- large $|\hat{\theta}_2|$ indicates quadratic association between y and x.
- large $|\hat{\theta}_3|$ indicates cubic association between y and x.

```
proc glm data=chickens; class diet; model gain=diet;
contrast 'linear' diet -3 -1 1 3;
contrast 'quadratic' diet 1 -1 -1 1;
contrast 'cubic' diet -1 3 -3 1; run;
```

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
linear	1	39129.40362	39129.40362	26.29	<.0001
quadratic	1	2700.51253	2700.51253	1.81	0.1825
cubic	1	190.57590	190.57590	0.13	0.7216

Note the equivalence between this output and the linear regression model cubic in protein. Conclusion? Go ahead and use the linear model rather than the full ANOVA. No need to look at pairwise comparisons etc.

F-ratio for lack-of-fit:
To test for lack-of-fit of a polynomial (*reduced*) model of degree p, use extra sum-of-squares F-ratio on $t-1-p$ and $N-t$ *df*:

$$F = \frac{SS(\text{lack of fit})/(t-1-p)}{MS(E)_{full}}$$

where

$$\begin{aligned} SS(\text{lack-of-fit}) &= SS(Trt) - SS(R)_{poly} \\ &= SS(E)_{poly} - SS(E)_{full} \end{aligned}$$

For illustration purposes (i.e. this isn't necessary), let's test if the quadratic model is sufficient or if the full ANOVA model is necessary.

Chapter 9

ST 512 - Analysis of Factorial Designs (Multiway ANOVA)

Readings: 13.1-13.6 584-633

We've looked at one-way ANOVA so far. This models the t treatments using $t-1$ degrees of freedom. However, the treatments may actually be combinations of more than 1 factor of interest. What we'll be able to do now is answer the question, which factor(s) (not treatment(s)) are important!

A multiway ANOVA example
An observational study was done to investigate the Cholesterol levels of different groups of people. There were two factors in this experiment

- Age - levels = Younger than 50 (Young), Older than 50 (Old)
- Gender - levels = Male, Female

Therefore, we have 2x2=4 treatments. Label the treatments as

- OF = Old Female (group 1)
- OM = Old Male (group 2)
- YF = Young Female (group 3)
- YM = Young Male (group 4)

Within each treatment group we have $n_j = n = 7$ observations (a balanced design). To investigate if any treatment means differ, we can do a One-Way ANOVA analysis by fitting the model

$$Y_{ij} = \mu + \tau_i + E_{ij}$$

where $i = 1(OF), 2(OM), 3(YF), 4(YM)$, $j = 1, 2, \ldots, 7$ and E_{ij} i.i.d. $N(0, \sigma^2)$. We constrain that $\sum_{i=1}^{t} \tau_i = 0$.

The data and One-Way ANOVA output are given below:

Treatment	Cholesterol level							avg	std. dev.
OF (i=1)	262	193	224	201	161	178	265	$\bar{y}_{1+} = 212.0$	$s_1 = 40$
OM (i=2)	192	253	248	278	232	267	289	$\bar{y}_{2+} = 251.3$	$s_2 = 32$
YF (i=3)	221	213	202	183	185	197	162	$\bar{y}_{3+} = 194.7$	$s_3 = 20$
YM (i=4)	271	192	189	209	227	236	142	$\bar{y}_{4+} = 209.4$	$s_4 = 41$

```
proc glm data=cholesterol;
class Treatment;
model Chol=Treatment;
lsmeans Treatment/cl pdiff adjust=tukey;
run;
```

Treatment	Chol LSMEAN	LSMEAN Number
OF	212.000000	1
OM	251.285714	2
YF	194.714286	3
YM	209.428571	4

Least Squares Means for effect Treatment
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: Chol

i/j	1	2	3	4
1		0.1707	0.7841	0.9990
2	0.1707		0.0250	0.1322
3	0.7841	0.0250		0.8538
4	0.9990	0.1322	0.8538	

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	12280.85714	4093.61905	3.46	0.0323
Error	24	28434.57143	1184.77381		
Corrected Total	27	40715.42857			

Treatment	Chol LSMEAN	95% Confidence Limits	
OF	212.000000	185.149211	238.850789
OM	251.285714	224.434925	278.136503
YF	194.714286	167.863497	221.565075
YM	209.428571	182.577782	236.279360

R-Square	Coeff Var	Root MSE	Chol Mean
0.301627	15.87245	34.42054	216.8571

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Treatment	3	12280.85714	4093.61905	3.46	0.0323

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Treatment	3	12280.85714	4093.61905	3.46	0.0323

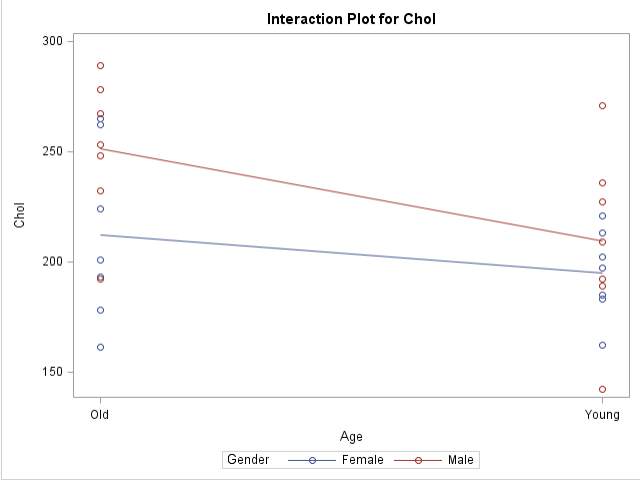
Least Squares Means for Effect Treatment

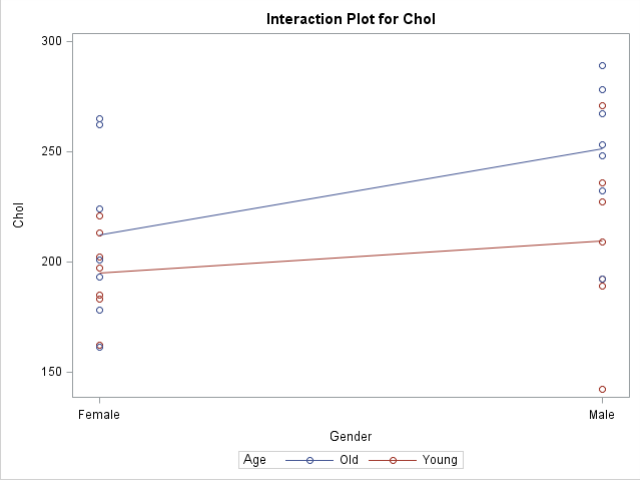
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-39.285714	-90.040133	11.468704
1	3	17.285714	-33.468704	68.040133
1	4	2.571429	-48.182990	53.325847
2	3	56.571429	5.817010	107.325847
2	4	41.857143	-8.897276	92.611561
3	4	-14.714286	-65.468704	36.040133

Conclusion from ANOVA table p-value is that the treatment means, $\mu + \tau_i$ or equivalently μ_i, are not plausibly equal (using $\alpha = 0.05$). From the lsmeans statement we can see that Young Females and Old Males are the only groups that differ significantly.

Now suppose we want to decide what effects the Age factor and the Gender factor have on the response. That is, rather than just inspect treatment mean differences, can we say that Age is significant for predicting Blood Pressure? What about Gender?

We can investigate these by looking at contrasts! Consider the plots below:





The **main effect** of factor A is the change in the response for switching levels of factor A (averaged over all other factors).
What contrast would test for the *main effect* of Age?

$\theta_{Age} =$

What contrast would test for the *main effect* of Gender?

$\theta_{Gender} =$

There is a third contrast of interest. This contrast represents the **interaction** between Age and Gender.

An **interaction** between factor A and factor B implies that the effect of factor A on the response depends on the level of factor B (and vice-versa).

In terms of the plots above, Age and Gender interact if the lines are not parallel (can look at either plot). What contrast can we write to investigate this?

$\theta_{Age*Gender} =$

1. Check that θ_{Age}, θ_{Gender}, and $\theta_{Age*Gender}$ are mutually orthogonal.

2. Use the previous output to find estimates for each contrast and provide standard errors. Recall: $\hat{\theta} = \sum c_i \bar{y}_{i+}$ and $\hat{SE}(\hat{\theta}) = \sqrt{MS(E) \sum_{i=1}^{t} \frac{c_i^2}{n_i}}$

   $\hat{\theta}_{Age} =$

   $\hat{\theta}_{Gender} =$

   $\hat{\theta}_{Age*Gender} =$

3. Find the sums of squares for each contrast. How many degrees of freedom are associated with each contrast? Recall: $SS(\hat{\theta}) = \frac{\hat{\theta}^2}{\sum \frac{c_i^2}{n_i}}$.

   $SS(\hat{\theta}_{Age}) =$

   $SS(\hat{\theta}_{Gender}) =$

   $SS(\hat{\theta}_{Age*Gender}) =$

4. Formulate a test of $H_0 : \theta_i = 0$ for each of these three contrasts. Obtain the F-ratio for each of these tests. Compare them to the F-critical value $F(1, 24, 0.05) = 4.26$) and make a decision about the importance of each effect.

   $F_{Age} =$ _____ $num\ df =$ _____ $den\ df =$ _____

   $F_{Gender} =$ _____ $num\ df =$ _____ $den\ df =$ _____

   $F_{Age*Gender} =$ _____ $num\ df =$ _____ $den\ df =$ _____

5. What do you notice about the sum of these sums of squares? Find the F-statistic for a test of $H_0 : \theta_{Age} = \theta_{Gender} = \theta_{Age*Gender} = 0$ vs H_A : at least one differs. What do you notice about this test and the overall F-test from the ANOVA table in the One-Way model?

Notice what we've done: Very similar to partitioning the $SS(Tot)$ into $SS(Model)$ and $SS(E)$, we've partitioned $SS(Trt)$, which has $t - 1 = 4 - 1 = 3$ degrees of freedom into 3 independent components that represent different effects of interest! We can test for each effect to learn more about our factors rather than just the treatment means.

This is the idea of Multi-Way ANOVA! (Although it gets a little bit more complicated when a factor has more than 2 levels.)

Let's look at how we could get these contrasts in SAS:

```
proc glm data=cholesterol;  class Treatment;  model Chol=Treatment;
contrast 'Age Main Effect Contrast' intercept 0 treatment 0.5 0.5 -0.5 -0.5;
contrast 'Gender Main Effect Contrast' intercept 0 treatment 0.5 -0.5 0.5 -0.5;
contrast 'Age*Gender Interaction Effect Contrast' intercept 0 treatment 0.5 -0.5 -0.5 0.5;
estimate 'Age Main Effect Estimate' intercept 0 treatment 0.5 0.5 -0.5 -0.5;
estimate 'Gender Main Effect Estimate' intercept 0 treatment 0.5 -0.5 0.5 -0.5;
estimate 'Age*Gender Interaction Effect Estimate' intercept 0 treatment 0.5 -0.5 -0.5 0.5;  run;
```

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
Age Main Effect Contrast	1	6121.285714	6121.285714	5.17	0.0323
Gender Main Effect Contrast	1	5103.000000	5103.000000	4.31	0.0488
Age*Gender Interaction Effect Contrast	1	1056.571429	1056.571429	0.89	0.3544

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Age Main Effect Estimate	29.5714286	13.0097426	2.27	0.0323
Gender Main Effect Estimate	-27.0000000	13.0097426	-2.08	0.0488
Age*Gender Interaction Effect Estimate	-12.2857143	13.0097426	-0.94	0.3544

Using the estimate and contrast statements in One-Way ANOVA:
We need to write our contrast in terms of the model parameters μ, τ_1, τ_2, τ_3, and τ_4.

For instance,

$$\theta_{Gender} = \frac{1}{2}(\mu_1 + \mu_3) - \frac{1}{2}(\mu_2 + \mu_4) = \frac{1}{2}(\mu + \tau_1 + \mu + \tau_3 - \mu - \tau_2 - \mu - \tau_4)$$

$$= 0\mu + \frac{1}{2}\tau_1 - \frac{1}{2}\tau_2 + \frac{1}{2}\tau_3 - \frac{1}{2}\tau_4$$

In terms of syntax, we write contrast or estimate followed by a name to distinguish it. Then we do

intercept $\underline{\text{coef on } \mu}$ treatment $\underline{\text{coef on } \tau_1}$ $\underline{\text{coef on } \tau_2}$ $\underline{\text{coef on } \tau_3}$ $\underline{\text{coef on } \tau_4}$

A contrast statements will give you the contrast sum of squares and a p-value.

An estimate statement will estimate any 'estimable' function of parameters (coefficients don't have to sum to 0).

Two-Way ANOVA example

Rather than fit a One-Way ANOVA model, we can use a different parameterization of that model called a Two-Way ANOVA model.

The Two-Way ANOVA model for the cholesterol measurements is:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + E_{ijk}$$

$i = 1, 2(old, young) \quad j = 1, 2(female, male)$ and $k = 1, 2, \ldots, 7$.
We still assume $E_{ijk} \overset{iid}{\sim} N(0, \sigma^2)$. With parameter constraints:

$$\alpha_1 + \alpha_2 = 0, \beta_1 + \beta_2 = 0, \text{ and}$$

$$(\alpha\beta)_{11} + (\alpha\beta)_{12} = 0, (\alpha\beta)_{21} + (\alpha\beta)_{22} = 0, (\alpha\beta)_{11} + (\alpha\beta)_{21} = 0, (\alpha\beta)_{12} + (\alpha\beta)_{22} = 0$$

This is called a 2×2 design since we have 2 factors with 2 levels each and the treatments are found by *crossing* the levels of the factors. (A three factor design with 2, 3, and 5 levels would be a 2×3×5 design.)

- Y_{ijk} is the response for replicate k at level i of Age and level j of Gender
- μ represents the overall mean of cholesterol,

$$\text{estimate is } \hat{\mu} = \bar{Y}_{+++}$$

- α_i represents the 'effect' for being at level i of Age,

$$\text{estimate is } \hat{\alpha}_i = \bar{Y}_{i++} - \bar{Y}_{+++}$$

- β_i represents the 'effect' for being at level j of Gender,

$$\text{estimate is } \hat{\beta}_j = \bar{Y}_{+j+} - \bar{Y}_{+++}$$

- $(\alpha\beta)_{ij}$ represents the 'joint effect' for being at level i of Age and level j of Gender,

$$\text{estimate is } \hat{(\alpha\beta)}_{ij} = \bar{Y}_{ij+} - \bar{Y}_{i++} - \bar{Y}_{+j+} + \bar{Y}_{+++}$$

- E_{ijk} is a random error

For level i of Age and level j of Gender we are model the mean cholesterol as

$$\mu_{ij} = \mu(A_i B_j) = E(Y_{ijk}) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

and, as you might expect, the estimate for level i of Age and level j of Gender is

$$\hat{\mu_{ij}} = \mu(\widehat{A_i B_j}) = \bar{Y}_{+++} + (\bar{Y}_{i++} - \bar{Y}_{+++}) + (\bar{Y}_{+j+} - \bar{Y}_{+++}) + (\bar{Y}_{ij+} - \bar{Y}_{i++} - \bar{Y}_{+j+} + \bar{Y}_{+++}) = \bar{Y}_{ij+}$$

The two-way ANOVA model can be fit easily in proc glm using the following code:

```
proc glm data=cholesterol;
class Age Gender;
model Chol=Age Gender Age*Gender;
run;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	12280.85714	4093.61905	3.46	0.0323
Error	24	28434.57143	1184.77381		
Corrected Total	27	40715.42857			

R-Square	Coeff Var	Root MSE	Chol Mean
0.301627	15.87245	34.42054	216.8571

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Age	1	6121.285714	6121.285714	5.17	0.0323
Gender	1	5103.000000	5103.000000	4.31	0.0488
Age*Gender	1	1056.571429	1056.571429	0.89	0.3544

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Age	1	6121.285714	6121.285714	5.17	0.0323
Gender	1	5103.000000	5103.000000	4.31	0.0488
Age*Gender	1	1056.571429	1056.571429	0.89	0.3544

The sums of squares for each effect are equal to the sums of squares for our contrasts in the one-way ANOVA model.

When looking at Two-Way ANOVA output,

1. Inspect the overall ANOVA table p-value. If significant,
2. Inspect the interaction p-value.
   (a) If significant, both factors are important for predicting the response (why?).
   (b) If not significant, investigate the main effect p-values for significance to determine which factors are important for predicting the response.

Types of Effects Investigated There are three basic effects that are usually of interest in Two-way ANOVA:

- Simple Effects
- Main Effects
- Interaction Effects

Use this table of means from the cholesterol example in the following questions:

	Gender	
	female(j=1)	male(j=2)
Age Older (i=1)	$\hat{\mu}_{11}$=212.0 (previously $\hat{\mu}_1$)	$\hat{\mu}_{12}$=251.3 (previously $\hat{\mu}_2$)
Young (i=2)	$\hat{\mu}_{21}$=194.7 (previously $\hat{\mu}_3$)	$\hat{\mu}_{22}$=209.4 (previously $\hat{\mu}_4$)

1. If we denote the mean for a level combination as μ_{ij} or $\mu(A_iB_j)$ (i.e. mean at factor A level i and factor B level b), how many different means are we attempting to estimate in a 2×2 Two-way ANOVA design? How many means in a general a×b Two-way ANOVA design?

2. **Simple Effects**

   (a) In the 2×2 set-up, a simple effect for factor A is the difference in the level means of factor A *at a given level of factor B*. That is,

   Simple effect of A at level 1 of B is defined as $\mu[AB_1] = \mu(A_2B_1) - \mu(A_1B_1) = \mu_{21} - \mu_{11}$,

   Simple effect of A at level 2 of B is defined as $\mu[AB_2] = \mu(A_2B_2) - \mu(A_1B_2) = \mu_{22} - \mu_{12}$,

   (b) Define the simple effects of factor B.

   (c) For the Cholesterol example, estimate these four simple effects and explain what each measures.

3. **Main Effects**

(a) Main effects in a 2x2 experiment are the averages of the simple effects. Define the main effect of factor A as

$$\mu[A] = \frac{1}{2}(\mu[AB_1] + \mu[AB_2]) = \frac{1}{2}(\mu_{22} - \mu_{12} + \mu_{21} - \mu_{11}) = \frac{1}{2}(\mu_{22} + \mu_{21}) - \frac{1}{2}(\mu_{12} + \mu_{11})$$

Our θ_{Age} contrast from before!

(b) Define the main effect for factor B.

*****Main effects should (usually) only be looked at when interaction effects are not significant.

(c) For the Cholesterol example, estimate these two main effects and explain what each measures.

4. **Interaction Effects**

(a) Interaction effects in a 2x2 experiment are the average of the *difference* of simple effects. If nonzero, this implies that when the level of factor B is changed, factor A acts differently with respect to the response. Define the interaction between factor A and B as

$$\mu[AB] = \frac{1}{2}(\mu[AB_2] - \mu[AB_1]) = \frac{1}{2}(\mu_{22} - \mu_{12} - \mu_{21} + \mu_{11}).$$

Our $\theta_{Age*Gender}$ contrast from before!

(b) Show the average of the difference of simple effects for factor B yields the same answer as above.

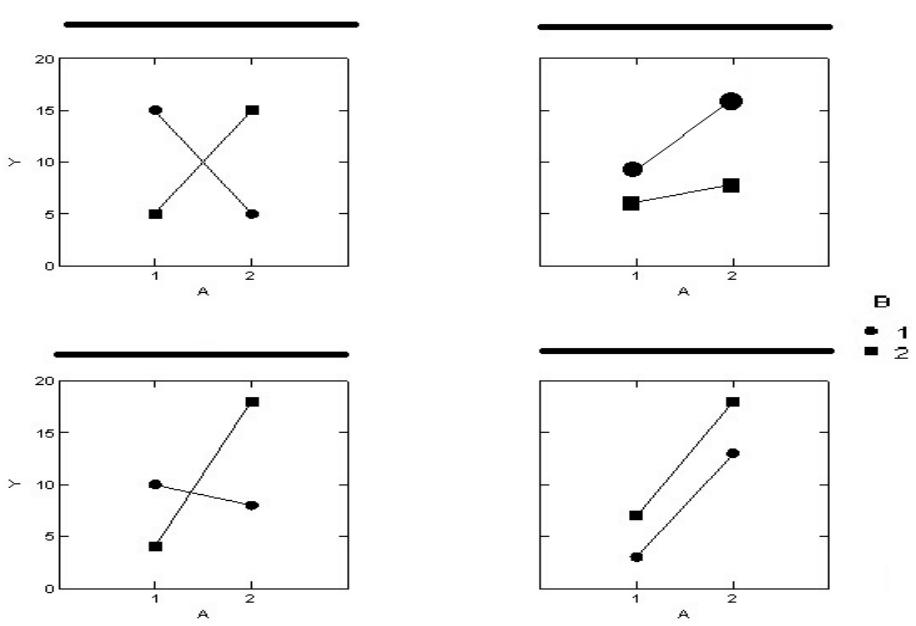
(c) For the Cholesterol example, estimate this interaction effect and explain what it measures.

(d) Interactions are easily seen in a 2x2 experiment by looking at an 'interaction plot'.

i. If an interaction effect causes the relationship between the levels of a factor to change, this is called a **qualitative** interaction.

ii. If the interaction effect just changes the magnitude of the relationship between the levels of a factor, this is called a **quantitative** interaction.

(e) Label the plots below accordingly:



(f) If interactions exist, looking at main effects is usually not necessary. For example, approximately what would the main effect for factor A be equal to for the top left plot?**An interaction implies both factors are important!**

To test for the Age main effect notice,

$$\begin{aligned}
\hat{\mu}[A] &= \frac{1}{2}(\hat{\mu}_{22} - \hat{\mu}_{12} + \hat{\mu}_{21} - \hat{\mu}_{11}) \\
&= \frac{1}{2}(\bar{Y}_{22+} - \bar{Y}_{12+} + \bar{Y}_{21+} - \bar{Y}_{11+}) \\
&= \frac{1}{2}(\bar{Y}_{22+} + \bar{Y}_{21+}) - \frac{1}{2}(\bar{Y}_{12+} + \bar{Y}_{11+}) \\
&= \bar{Y}_{2++} - \bar{Y}_{1++} \\
&= \hat{\alpha}_2 - \hat{\alpha}_1
\end{aligned}$$

Our test for the main effect of Age can be written as

$$H_0 : \alpha_1 = \alpha_2 = 0 \text{ vs } H_A : \text{ At least 1 is not 0}$$

Similarly, we can test for the Gender main effect by

$$H_0 : \beta_1 = \beta_2 = 0 \text{ vs } H_A : \text{ At least 1 is not 0}$$

And we can test the interaction using

$$H_0 : (\alpha\beta)_{11} = (\alpha\beta)_{12} = (\alpha\beta)_{21} = (\alpha\beta)_{22} = 0 \text{ vs } H_A : \text{ At least 1 is not 0}$$

Thus, this parameterization of the model gives a very nice way to test for these different effects.

The general Two-Way ANOVA model:
Suppose we have a *continuous* response, Y, and two factors, A and B. Factor A has a levels and factor B has b levels. Our main interest lies in whether or not the response differs due to the factors.

Most experiments with multiple factors we will look at will have a **factorial** treatment structure. This implies 'treatments' are combinations of the levels of different factors (also called a crossed design). For the most part we will have **complete** experiments (i.e. observations at each level combination).

This particular experiment would be an a×b experiment. The parametrization of the Two-way ANOVA model we will use is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + E_{ijk}, \quad i = 1, ..., a, \quad j = 1, ..., b, k = 1, ..., n \text{ balanced design}$$

with normality assumption on the errors and sum to zero constraints on the parameters (to get a unique solution).

To construct the ANOVA table, we will still take the total sum of squares and split it up. Now we split it into a few more parts than previous:

ANOVA table for a×b experiment with $N = abn$ = total # of obs

Source	df	Sum of Squares (SS)	MS
A	$a-1$	$SS(A) = \sum_i \sum_j \sum_k (\bar{y}_{i++} - \bar{y}_{+++})^2$	SS(A)/(a-1)
B	$b-1$	$SS(B) = \sum_i \sum_j \sum_k (\bar{y}_{+j+} - \bar{y}_{+++})^2$	SS(B)/(b-1)
AB	$(a-1)(b-1)$	$SS(AB) = \sum_i \sum_j \sum_k (\bar{y}_{ij+} - \bar{y}_{i++} - \bar{y}_{+j+} + \bar{y}_{+++})^2$	SS(AB)/((a-1)(b-1))
Error	$ab(n-1)$	$SS(E) = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{ij+})^2$	SS(E)/(ab(n-1))
Total	$N-1$	$SS(Tot) = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{+++})^2$	

The F-statistics are given by $MS(A)/MS(E)$, $MS(B)/MS(E)$, and $MS(AB)/MS(E)$ respectively.

Note: $SS(A) + SS(B) + SS(AB) = SS(Trt)$ in One-Way ANOVA

$(a-1) + (b-1) + (a-1)(b-1) = t - 1$ (degrees of freedom in One-Way ANOVA)

We still have that $SS(A) + SS(B) + SS(AB) + SS(E) = SS(Tot)$ and same for the degrees of freedom.

An a×b example

An entomologist records energy expended (y) by $N = 27$ honeybees at $a = 3$ temperature (A) levels ($20, 30, 40^oC$) consuming liquids with $b = 2$ levels of sucrose concentration (B)($20\%, 40\%$) in a balanced, completely randomized crossed 3×2 design. The data are given below:

Temp	Suc	Sample		
20	20	3.1	3.7	4.7
20	40	5.5	6.7	7.3
30	20	6	6.9	7.5
30	40	11.5	12.9	13.4
40	20	7.7	8.3	9.5
40	40	15.7	14.3	15.9

```
proc glm data=ent;
class Temp Suc;
model Energy=Temp|Suc; *Vertical Bar fits all combinations of Temp and Suc (main effects and interactions);
run;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	271.9444444	54.3888889	70.43	<.0001
Error	12	9.2666667	0.7722222		
Corrected Total	17	281.2111111			

R-Square	Coeff Var	Root MSE	energy Mean
0.967047	9.849135	0.878762	8.922222

Source	DF	Type I SS	Mean Square	F Value	Pr > F
temp	2	141.4577778	70.7288889	91.59	<.0001
Suc	1	116.5355556	116.5355556	150.91	<.0001
temp*Suc	2	13.9511111	6.9755556	9.03	0.0040

Source	DF	Type III SS	Mean Square	F Value	Pr > F
temp	2	141.4577778	70.7288889	91.59	<.0001
Suc	1	116.5355556	116.5355556	150.91	<.0001
temp*Suc	2	13.9511111	6.9755556	9.03	0.0040

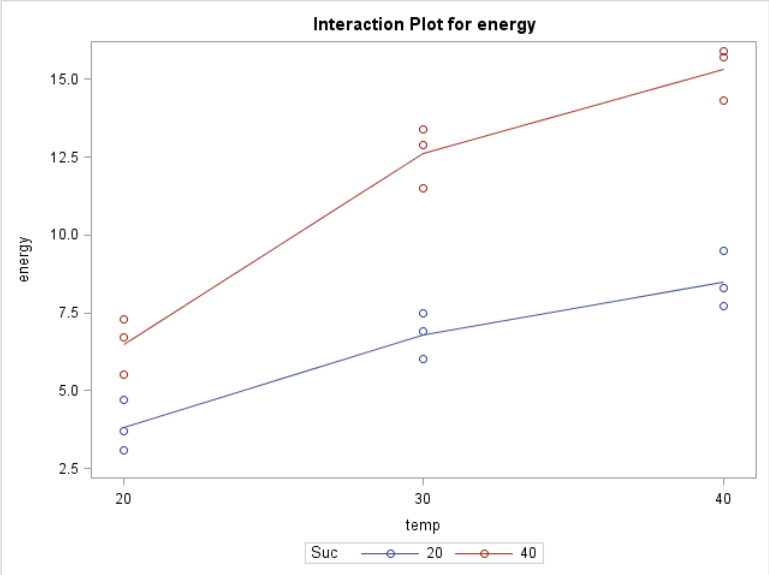
Unlike a 2×2 study, it is not possible to express interaction between Temp and Suc using 1 contrast.

Likewise, we can't estimate the main effect of Temp with only 1 contrast!

```
title 'Means for each treatment group';
proc tabulate data=ent;
class Temp Suc;
var Energy;
table Temp*mean, Energy*Suc;
run;
```

Means for each treatment group

		energy	
		Suc	
		20	40
temp			
20	Mean	3.83	6.50
30	Mean	6.80	12.60
40	Mean	8.50	15.30



Testing Interaction in an $a \times b$ experiment

Here we have (3-1)(2-1)=2 degrees of freedom for interaction. Why? Well, no interaction implies changing the level of Temp doesn't change the effect of Suc on Energy, i.e.

$$\mu_{12} - \mu_{11} = \mu_{22} - \mu_{21} \text{ or } \mu_{Temp20,Suc40} - \mu_{Temp20,Suc20} = \mu_{Temp30,Suc40} - \mu_{Temp30,Suc20}$$

as well as

$$\mu_{12} - \mu_{11} = \mu_{32} - \mu_{31} \text{ or } \mu_{Temp20,Suc40} - \mu_{Temp20,Suc20} = \mu_{Temp40,Suc40} - \mu_{Temp40,Suc20}$$

as well as

$$\mu_{32} - \mu_{31} = \mu_{22} - \mu_{21} \text{ or } \mu_{Temp40,Suc40} - \mu_{Temp40,Suc20} = \mu_{Temp30,Suc40} - \mu_{Temp30,Suc20}$$

But notice that given any two of these contrasts (move all means to one side to see these as a contrast), we can get the third contrast. So we have 3-1=2 contrasts that are needed for testing interaction.

Again, no interaction would imply **piecewise parallel lines** across all the levels of the factor on the axis.

Test for interaction effect generalizes as:

$$H_0 : (\alpha\beta)_{ij} \equiv 0 \text{ vs. } H_1 : (\alpha\beta)_{ij} \neq 0 \text{ for some } i, j$$

$$F = \frac{MS(AB)}{MS(E)}$$

on $(a-1)(b-1)$ numerator and $abn - ab$ denominator df.

For honeybee data,

$$SS(AB) = n \sum_{i=1}^{3} \sum_{j=1}^{2} (\bar{y}_{ij+} - \bar{y}_{i++} - \bar{y}_{+j+} + \bar{y}_{+++})^2 = 13.95$$

which is highly significant ($p = 0.004$) on 2 and 12 degrees of freedom.

Note: These three contrasts are not orthogonal to one another so we could not use them to find $SS(AB)$, we would have to find 2 orthogonal contrasts that represent this effect.

Since we have a significant interaction, our next step would be to analyze **simple effects** of each factor. That is, we would investigate the effects of Temperature at given levels of Sucrose and also looking at the effect of Sucrose at given levels of Temperature. This can be done in SAS:

```
proc glm data=ent; class Temp Suc;
model Energy=Temp|Suc;
lsmeans Temp*Suc/adjust=tukey pdiff cl; run;
```

temp	Suc	energy LSMEAN	95% Confidence Limits	
20	20	3.833333	2.727905	4.938761
20	40	6.500000	5.394572	7.605428
30	20	6.800000	5.694572	7.905428
30	40	12.600000	11.494572	13.705428
40	20	8.500000	7.394572	9.605428
40	40	15.300000	14.194572	16.405428

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey

temp	Suc	energy LSMEAN	LSMEAN Number
20	20	3.8333333	1
20	40	6.5000000	2
30	20	6.8000000	3
30	40	12.6000000	4
40	20	8.5000000	5
40	40	15.3000000	6

Least Squares Means for effect temp*Suc
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: energy

i/j	1	2	3	4	5	6
1		0.0274	0.0135	<.0001	0.0003	<.0001
2	0.0274		0.9979	<.0001	0.1274	<.0001
3	0.0135	0.9979		<.0001	0.2404	<.0001
4	<.0001	<.0001	<.0001		0.0011	0.0253
5	0.0003	0.1274	0.2404	0.0011		<.0001
6	<.0001	<.0001	<.0001	0.0253	<.0001	

Least Squares Means for Effect temp*Suc

i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-2.666667	-5.076699	-0.256635
1	3	-2.966667	-5.376699	-0.556635
1	4	-8.766667	-11.176699	-6.356635
1	5	-4.666667	-7.076699	-2.256635
1	6	-11.466667	-13.876699	-9.056635
2	3	-0.300000	-2.710032	2.110032
2	4	-6.100000	-8.510032	-3.689968
2	5	-2.000000	-4.410032	0.410032
2	6	-8.800000	-11.210032	-6.389968
3	4	-5.800000	-8.210032	-3.389968
3	5	-1.700000	-4.110032	0.710032
3	6	-8.500000	-10.910032	-6.089968
4	5	4.100000	1.689968	6.510032
4	6	-2.700000	-5.110032	-0.289968
5	6	-6.800000	-9.210032	-4.389968

In reality, we would probably only be interested in simple effects such as Temp 20, Suc 40 vs Temp 20, Suc 20 (i.e. the effect of Suc, holding temperature at 20). We could write estimate or contrast to get just those, but this is easier.

We could proceed to test for main effects, but we won't. Why not?

If one insists on main effects, the appropriate F-ratios are

$$F_A = \frac{SS(A)/(a-1)}{MS(E)} \text{ on } a-1, N-ab \;\; df$$

$$F_B = \frac{SS(B)/(b-1)}{MS(E)} \text{ on } b-1, N-ab \;\; df$$

Testing of the Sucrose main effect would be done just as before since we have only two levels, we would look at the marginal means of sucrose and compare them.

Testing of the Temperature main effect requires a bit more, we want to compare the three levels of temperature (20, 30, 40):

$$\frac{1}{2}(\mu_{21}+\mu_{22}) = \frac{1}{2}(\mu_{11}+\mu_{12}) \text{ or } \frac{1}{2}(\mu_{Temp30,Suc20}+\mu_{Temp30,Suc40}) = \frac{1}{2}(\mu_{Temp20,Suc20}+\mu_{Temp20,Suc40})$$

as well as

$$\frac{1}{2}(\mu_{31}+\mu_{32}) = \frac{1}{2}(\mu_{11}+\mu_{12}) \text{ or } \frac{1}{2}(\mu_{Temp40,Suc20}+\mu_{Temp40,Suc40}) = \frac{1}{2}(\mu_{Temp20,Suc20}+\mu_{Temp20,Suc40})$$

as well as

$$\frac{1}{2}(\mu_{31}+\mu_{32}) = \frac{1}{2}(\mu_{21}+\mu_{22}) \text{ or } \frac{1}{2}(\mu_{Temp40,Suc20}+\mu_{Temp40,Suc40}) = \frac{1}{2}(\mu_{Temp30,Suc20}+\mu_{Temp30,Suc40})$$

Again, given any two of these we can find the third so we really only need 2 contrasts ($df = 3 - 1 = 2$).

Another $a \times b$ Design - Interaction Not Significant

Yields on 36 tomato crops from balanced, complete, crossed design with $a = 3$ varieties (A) at $b = 4$ planting densities (B) :

Variety	Density $k/hectare$	Sample		
1	10	7.9	9.2	10.5
2	10	8.1	8.6	10.1
3	10	15.3	16.1	17.5
1	20	11.2	12.8	13.3
2	20	11.5	12.7	13.7
3	20	16.6	18.5	19.2
1	30	12.1	12.6	14.0
2	30	13.7	14.4	15.4
3	30	18.0	20.8	21.0
1	40	9.1	10.8	12.5
2	40	11.3	12.5	14.5
3	40	17.2	18.4	18.9

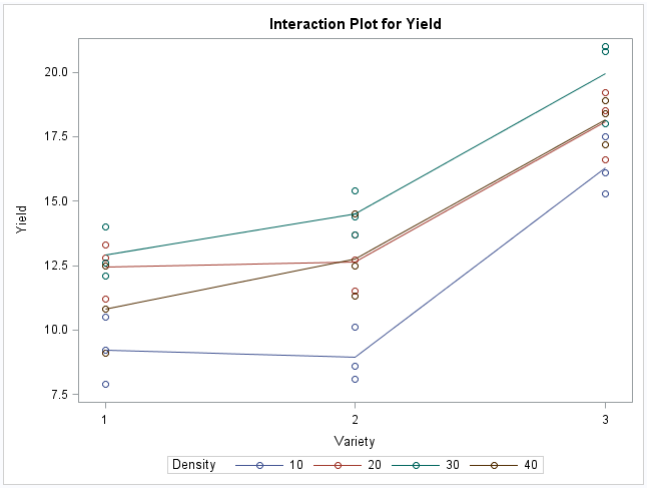
```
proc glm data=tomato; class variety density;
model Yield=Variety|Density;
lsmeans Variety Density/adjust=tukey cl; *adjust=tukey tells sas to do pdiff; run;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	422.3155556	38.3923232	24.22	<.0001
Error	24	38.0400000	1.5850000		
Corrected Total	35	460.3555556			

R-Square	Coeff Var	Root MSE	Yield Mean
0.917368	9.064568	1.258968	13.88889

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Variety	2	327.5972222	163.7986111	103.34	<.0001
Density	3	86.6866667	28.8955556	18.23	<.0001
Variety*Density	6	8.0316667	1.3386111	0.84	0.5484

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Variety	2	327.5972222	163.7986111	103.34	<.0001
Density	3	86.6866667	28.8955556	18.23	<.0001
Variety*Density	6	8.0316667	1.3386111	0.84	0.5484



As the interaction is not significant we want to look at main effects p-values. These are significant for both factors, thus both are important. We should now look at the main effect differences for both factors (since both are important, we would not look at differences for a factor deemed non-significant).

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey

Variety	Yield LSMEAN	LSMEAN Number
1	11.3333333	1
2	12.2083333	2
3	18.1250000	3

Least Squares Means for effect Variety
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: Yield

i/j	1	2	3
1		0.2249	<.0001
2	0.2249		<.0001
3	<.0001	<.0001	

Variety	Yield LSMEAN	95% Confidence Limits	
1	11.333333	10.583245	12.083422
2	12.208333	11.458245	12.958422
3	18.125000	17.374912	18.875088

Least Squares Means for Effect Variety

i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-0.875000	-2.158534	0.408534
1	3	-6.791667	-8.075201	-5.508132
2	3	-5.916667	-7.200201	-4.633132

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey

Density	Yield LSMEAN	LSMEAN Number
10	11.4777778	1
20	14.3888889	2
30	15.7777778	3
40	13.9111111	4

Least Squares Means for effect Density
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: Yield

i/j	1	2	3	4
1		0.0003	<.0001	0.0022
2	0.0003		0.1169	0.8514
3	<.0001	0.1169		0.0213
4	0.0022	0.8514	0.0213	

Density	Yield LSMEAN	95% Confidence Limits	
10	11.477778	10.611650	12.343905
20	14.388889	13.522762	15.255016
30	15.777778	14.911650	16.643905
40	13.911111	13.044984	14.777238

Least Squares Means for Effect Density

i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-2.911111	-4.548299	-1.273923
1	3	-4.300000	-5.937188	-2.662812
1	4	-2.433333	-4.070521	-0.796145
2	3	-1.388889	-3.026077	0.248299
2	4	0.477778	-1.159410	2.114966
3	4	1.866667	0.229479	3.503855

Again, analysis of replicated two (or more) factor designs often proceed according to the following directions:

1. Check for interaction
2. If no interaction, analyze main effects
3. If interaction, analyze simple effects

A three-factor example: In a balanced, complete, crossed design, $N = 36$ shrimp were randomized to $abc = 12$ treatment combinations from the factors below:

A1: Temperature at 25^o C
A2: Temperature at 35^o C
B1: Density of shrimp population at 80 shrimp/$40l$
B2: Density of shrimp population at 160 shrimp/$40l$
C1: Salinity at 10 units
C2: Salinity at 25 units
C3: Salinity at 40 units

Thus, this is a $2\times2\times3$ experiment. The response variable of interest is weight gain Y_{ijkl} after four weeks.

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + E_{ijkl}$$

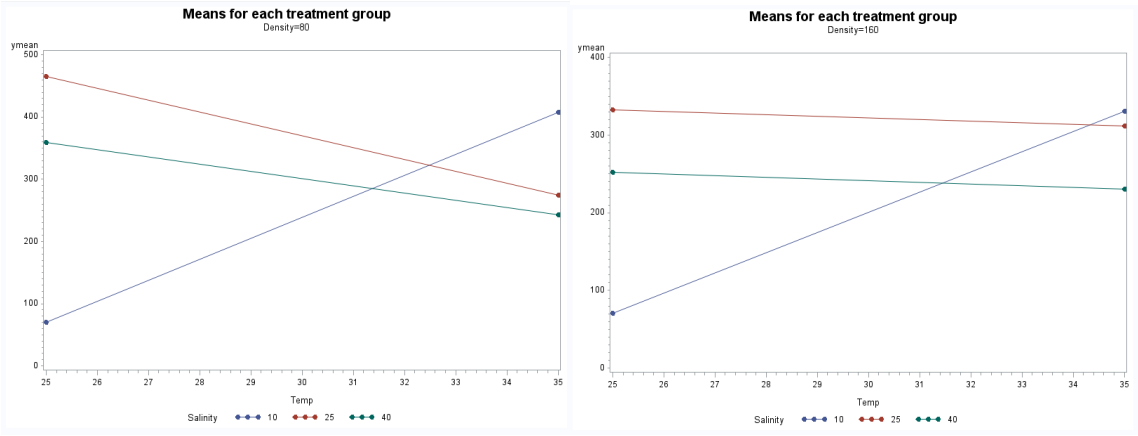
where

$$i = 1, 2, \quad j = 1, 2, \quad k = 1, 2, 3, \quad l = 1, 2, 3$$

$E_{ijkl} \overset{iid}{\sim} N(0, \sigma^2)$. Note: Many constraints are required on the parameters.

Analysis of a Multi-Way ANOVA model starts by investigating the highest order interactions and working down from there.

```
proc glm data=shrimp; class Temp Density Salinity;
model y=Temp|Density|Salinity; run;
```



Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	467636.3333	42512.3939	14.64	<.0001
Error	24	69690.6667	2903.7778		
Corrected Total	35	537327.0000			

R-Square	Coeff Var	Root MSE	y Mean
0.870301	19.30270	53.88671	279.1667

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Temp	1	15376.0000	15376.0000	5.30	0.0304
Density	1	21218.7778	21218.7778	7.31	0.0124
Temp*Density	1	8711.1111	8711.1111	3.00	0.0961
Salinity	2	96762.5000	48381.2500	16.66	<.0001
Temp*Salinity	2	300855.1667	150427.5833	51.80	<.0001
Density*Salinity	2	674.3889	337.1944	0.12	0.8909
Temp*Density*Salinit	2	24038.3889	12019.1944	4.14	0.0285

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Temp	1	15376.0000	15376.0000	5.30	0.0304
Density	1	21218.7778	21218.7778	7.31	0.0124
Temp*Density	1	8711.1111	8711.1111	3.00	0.0961
Salinity	2	96762.5000	48381.2500	16.66	<.0001
Temp*Salinity	2	300855.1667	150427.5833	51.80	<.0001
Density*Salinity	2	674.3889	337.1944	0.12	0.8909
Temp*Density*Salinit	2	24038.3889	12019.1944	4.14	0.0285

Three-way interaction is significant implying that all three factors are important. We would now look at simple effects. For example, the effect of temperature at density 80 and salinity 10.

Interpretation of second order interaction

1^{st} **order** interaction (or two-way interaction) is between two factors
2^{nd} **order** interaction (or three-way interaction) is between three factors, etc.

Consider the AB interaction at each of three levels, $C1, C2, C3$.

To do this, look at three 2×2 tables as follows:

		B	
$(C = 1)$		B1	B2
A	A1	70	71
	A2	408	331

		B	
$(C = 2)$		B1	B2
A	A1	466	333
	A2	275	312

		B	
$(C = 3)$		B1	B2
A	A1	359.0	252
	A2	243	231

Q: How is the ABC interaction manifested here?

A: We could compute $\hat{\mu}(ABC_1), \hat{\mu}(ABC_2), \hat{\mu}(ABC_3)$ and see if these first order interactions, with C fixed, are the same. (We know they are not by the F_{ABC} ratio and p-value.)

$$\hat{\mu}(ABC_1) = 408 - 70 - (331 - 71) \approx 77$$

Exercise: Obtain $\hat{\mu}(ABC_2)$ and $\hat{\mu}(ABC_3)$ as well as AB interaction plots for $C = 1$, $C = 2$ and $C = 3$.

Chapter 10

ST 512 - Random Effects Models

Readings: 14.1-14.3 643-660

Thus far we've considered the means of our factors as the main things of interest (called fixed factors). Sometimes we won't actually be interested in the mean of a factor but rather that factor's variability. Random effects models allow for this type of inference.

Example where a random factor is of interest.

- Consider a genetics study with beef animals. Response = birthweight Y (lbs).
- $t = 5$ sires (male cows), each mated to a separate group of $n = 8$ dams (female cows).
- $N = 40$, completely randomized design.
- Interest is in variability in birth rate based on sires.

		Birthweights									
Sire #	Level	Sample								$\bar{y}_{i+}$	s_i
177	1	61	100	56	113	99	103	75	62	83.6	22.6
200	2	75	102	95	103	98	115	98	94	97.5	11.2
201	3	58	60	60	57	57	59	54	100	63.1	15.0
202	4	57	56	67	59	58	121	101	101	77.5	25.9
203	5	59	46	120	115	115	93	105	75	91.0	28.0

What statistical model is appropriate for these data?

A: One-way 'fixed' effects model?

$$Y_{ij} = \underbrace{\mu}_{\text{fixed}} + \underbrace{\tau_i}_{\text{fixed}} + \underbrace{E_{ij}}_{\text{random}}$$

where τ_i denotes the difference between the mean birthweight of population of offspring from sire i and μ, mean of whole population.

We don't really care about these 5 sires, but interest is more in the entire population of possible sires. Here **sire is a random effect**.

Flow chart for identifying a factor as fixed or random.

	Random	Fixed
Levels - selected from conceptually ∞ popn of collection of levels - finite number of possible levels	X	X
Another expt - would use same levels - would involve new levels sampled from same population	X	X
Goal - estimate varcomps - estimate longrun means	X	X
Inference - for these levels used in this expt - for the population of levels	X	X

Contrast this situation with the binding fractions example. Why not model antibiotic effects random? Why fixed?

The One-Way random effects model (One-Way implies one factor)

$$Y_{ij} = \underbrace{\mu}_{\text{fixed}} + \underbrace{T_i}_{\text{random}} + \underbrace{E_{ij}}_{\text{random}}$$

where $i = 1, 2, \ldots, t$ (number of levels) and $j = 1, \ldots, n$ (number of replicates).

- Assume $T_1, T_2, \ldots, T_t \overset{iid}{\sim} N(0, \sigma_T^2)$
- Assume $E_{11}, \ldots, E_{tn} \overset{iid}{\sim} N(0, \sigma^2)$
- Assume $T_1, T_2, \ldots, T_t$ independent of $E_{11}, \ldots, E_{tn}$

Notation:

- $T_1, T_2, \ldots$ denote *random* effects, drawn from some population of interest.
  For beef animal genetic study, with $t = 5$ and $n = 8$, the random effects $T_1, T_2, \ldots, T_5$ reflect sire-to-sire variability.
- σ_T^2 and σ^2 are called the **variance components**
- Conceptually different from one-way fixed effects model, but analysis is equivalent!

For random effects model we now have:

$$E(Y_{ij}) = E(\mu + T_i + E_{ij}) = E(\mu) + E(T_i) + E(E_{ij}) = \mu + 0 + 0 = \mu$$

and

$$Var(Y_{ij}) = Var(\mu + T_i + E_{ij}) = Var(T_i) + Var(E_{ij}) + 2cov(T_i, E_{ij}) = \sigma_T^2 + \sigma^2 + 0 = \sigma_T^2 + \sigma^2$$

- Two *components* to variability in data: σ^2, σ_T^2

ANOVA table for One-Way Random effects model is the same as the One-Way Fixed effects model!

Source	SS	df	MS
Treatment	$SS(Trt)$	$t-1$	$MS(Trt)$
Error	$SS(E)$	$N-t$	$MS(E)$
Total	$SS(Tot)$	$N-1$	

Only difference is the expected values of the Mean Squares:

- Expected mean square for error = $E(MS(E)) = \sigma^2$
- Expected mean square for treatment
  - For random effects model = $E(MS(Trt)) = \sigma^2 + n\sigma_T^2$
  - For fixed effects model = $E(MS(Trt)) = \sigma^2 + \frac{n}{t-1}\sum_{i=1}^{t} \tau_i^2$

Main hypothesis of interest for One-Way random effects model:

$$H_0 : \sigma_T^2 = 0 \qquad vs. \qquad H_A : \sigma_T^2 > 0$$

If H_0 is true then

$$F = \frac{MS(Trt)}{MS(E)} \text{ will be approximately } \frac{\sigma^2 + 0}{\sigma^2} = 1$$

If H_0 is false then

$$F = \frac{MS(Trt)}{MS(E)} \text{ will be greater than 1}$$

Compare observed F to $F(t-1, N-t, \alpha)$. Again, same as One-Way fixed effects ANOVA model.

Estimating parameters of One-Way random effects model:
Estimate of μ is still

$$\hat{\mu} = \bar{y}_{++}$$

Estimate of σ^2 is still

$$\hat{\sigma}^2 = MS[E]$$

To estimate σ_T^2 we can 'equate mean squares'. We know $E(MS(Trt)) = \sigma^2 + n\sigma_T^2$. For large samples $MS(Trt)$ will be 'close' to $E(MS(Trt))$, so

$$\hat{\sigma}_T^2 = \frac{MS[T] - MS[E]}{n}$$

For sires data, $\bar{y}_{++} = 82.6$ and

Source	SS	df	MS	Expected MS
Sire	5591	4	1398	$\sigma^2 + 8\sigma_T^2$
Error	16233	35	464	σ^2
Total	21824	39		

Therefore, $\hat{\mu} = 82.6$, $\hat{\sigma}^2 = 464(lbs^2)$, and $\hat{\sigma}_T^2 = \frac{1398-464}{8} = 117$ (lbs^2).

Note: If you get an estimated σ_T^2 that is negative, it should be set to zero.

Testing a variance component - $H_0 : \sigma_T^2 = 0$
Recall that $\sigma_T^2 = \text{Var}(T_i)$, the variance among the population of treatment effects.

$$F = \frac{MS[T]}{MS[E]}$$

reject H_0 at level α if $F > F(\alpha, t-1, N-t)$ For the sires,

$$F = \frac{1398}{464} = 3.01 > 2.64 = F(0.05, 4, 35)$$

so H_0 is rejected at $\alpha = 0.05$. (The p-value is 0.0309)

Q: "Isn't this just just like the F-test for one-way ANOVA with *fixed* effects?"
A: "Yes."

Specific questions pertaining to this study:
Consider the birthweight of a randomly sampled calf.

1. What is the estimated variance of such a calf?

2. Estimate the proportion of this variation that is due to the sire effect.

3. Estimate the proportion of this variation that is not due to the sire effect.

Other quantities of interest in random effects models:

Coefficient of variation (CV):

$$CV(Y_{ij}) = \frac{\sqrt{\text{Var}(Y_{ij})}}{|E(Y_{ij}|} = \frac{\sqrt{\sigma_T^2 + \sigma^2}}{|\mu|}$$

Note: this is *not* estimated by `Coeff Var` in PROC GLM output.

Intraclass correlation coefficient: (ρ_I)

$$\rho_I = \frac{\text{Cov}(Y_{ij}, Y_{ik})}{\sqrt{\text{Var}(Y_{ij})\text{Var}(Y_{ik})}} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma^2}$$

- Interpretation: the correlation between two responses receiving the same level of the random factor.
- Bigger values of ρ_I correspond to larger random treatment effects.

For the sires data,

$$\widehat{CV} = \frac{\sqrt{117 + 464}}{82.6} = 0.29$$

$$\hat{\rho}_I = \frac{117}{117 + 464} = 0.20$$

Interpretations:

- The estimated standard deviation of a birthweight, 24.1 (lbs), is 29% of the estimated mean birthweight, 82.6.
- The estimated correlation between any two calves with the same sire for a male parent, or the estimated *intrasire* correlation coefficient, is 0.20

Using Proc GLM and Proc Mixed for random effects models:

```
proc glm data=sires;
class sire;
model BirthWeight=Sire;
random Sire;
run;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	5591.15000	1397.78750	3.01	0.0309
Error	35	16232.75000	463.79286		
Corrected Total	39	21823.90000			

R-Square	Coeff Var	Root MSE	BirthWeight Mean
0.256194	26.08825	21.53585	82.55000

The GLM Procedure

Source	Type III Expected Mean Square
Sire	Var(Error) + 8 Var(Sire)

```
proc mixed data=sires method=type3;
class sire;
model BirthWeight=;
random Sire;
run;
```

Type 3 Analysis of Variance

Source	DF	Sum of Squares	Mean Square	Expected Mean Square	Error Term	Error DF	F Value	Pr > F
Sire	4	5591.150000	1397.787500	Var(Residual) + 8 Var(Sire)	MS(Residual)	35	3.01	0.0309
Residual	35	16233	463.792857	Var(Residual)	.	.	.	.

Covariance Parameter Estimates

Cov Parm	Estimate
Sire	116.75
Residual	463.79

(Note: σ^2 =`Var(Error)` and σ_T^2 =`Var(sire)`.)

Interval Estimation of μ

A $100(1-\alpha)\%$ confidence interval for μ can be easily derived by consideration of $SE(\bar{Y}_{++})$:

$$\begin{aligned}\bar{Y}_{++} &= \frac{1}{N}\sum_{i=1}^{t}\sum_{j=1}^{n} Y_{ij} \\ &= \frac{1}{N}\sum_{i=1}^{t}\sum_{j=1}^{n}(\mu + T_i + E_{ij}) \\ &= \mu + \bar{T}_+ + \bar{E}_{++}\end{aligned}$$

where $\bar{T}_+ = (T_1 + \cdots + T_t)/t$ and $\bar{E}_{++} = (\sum\sum E_{ij})/N$, so that

$$\begin{aligned}\text{Var}(\bar{Y}_{++}) &= \text{Var}(\bar{T} + \bar{E}_{++}) \\ &= \frac{\sigma_T^2}{t} + \frac{\sigma_2}{nt} \\ &= \frac{1}{nt}(n\sigma_T^2 + \sigma^2) \\ &= \frac{1}{nt}E(MS[T]).\end{aligned}$$

If the data are normally distributed, then

$$\frac{\bar{Y}_{++} - \mu}{\sqrt{\frac{MS[T]}{nt}}} \sim t_{t-1}$$

and a $100(1-\alpha)\%$ confidence interval for μ given by

$$\bar{Y}_{++} \pm t(\alpha/2, t-1)\sqrt{\frac{MS[T]}{nt}}$$

Sires data: $\bar{y}_{++} = 82.6, MS[T] = 1398, nt = 40$. Critical value $t(0.025, 4) = 2.78$ yields the 95% CI

$$82.6 \pm 2.78(5.91) \quad \text{or} \quad (66.1, 99.0).$$

We are 95% confident the true mean birthweight is between 66.1 and 99.0 lbs.

By adding in the statement 'estimate 'mean' intercept 1/cl' to the proc mixed code we get (can use type3 or reml method)

Estimates								
Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
mean	82.5500	5.9114	4	13.96	0.0002	0.05	66.1373	98.9627

Interval estimation for variance components:
The estimated residual variance component for the sire data was

$$\hat{\sigma}^2 = MS[E] = 464 \ lbs^2$$

A $100(1-\alpha)\%$ confidence interval for this variance component is given by

$$\left(\frac{(N-t)MS[E]}{\chi^2_{\alpha/2}}, \ \frac{(N-t)MS[E]}{\chi^2_{1-\alpha/2}} \right).$$

For the sire data and $\alpha = 0.05$ this becomes,

$$\left(\frac{(40-5)464}{53.2} < \sigma^2 < \frac{(40-5)464}{20.6} \right)$$

$$\left(\frac{35}{53.2}464 < \sigma^2 < \frac{35}{20.6}464 \right)$$

or $(305.2, 789.5) lbs^2$
We are 95% confident the true error variance is between 305.2 and 789.5 lbs^2.

Interval estimation for σ^2_T
The estimated variance component for the random sire effect was $\hat{\sigma}^2_T = 117$.
A $100(1-\alpha)\%$ confidence interval for σ^2_T is given by

$$\left(\frac{\widehat{df}\hat{\sigma}^2_T}{\chi^2_{\alpha/2,\widehat{df}}}, \ \frac{\widehat{df}\hat{\sigma}^2_T}{\chi^2_{1-\alpha/2,\widehat{df}}} \right)$$

where

$$\widehat{df} = \frac{(n\hat{\sigma}^2_T)^2}{\frac{MS[T]^2}{t-1} + \frac{MS[E]^2}{N-t}}$$

is the Satterthwaite approximation to the degrees of freedom.

For the sire data,

$$\widehat{df} = \frac{(8 \times 117)^2}{\frac{1398^2}{4} + \frac{464^2}{35}} = 1.76$$

Software must be used to obtain this non-integer degrees of freedom (using a table we'd have to round to the nearest integer):

$$\chi^2_{0.975,1.76} = 0.029, \quad \chi^2_{0.025,1.76} = 6.87$$

yielding the 95% confidence interval

$$\left(\frac{1.76(117)}{6.87} \ \frac{1.76(117)}{0.29} \right) = (30, 7051)$$

We are 95% confident the variance between sires is between 30 and 7051 lbs^2.

To get these two intervals in SAS we need to use proc mixed with 'reml' estimation rather than type3.

```
proc mixed data=sires method=reml cl;
class sire;
model BirthWeight=;
random Sire;
run;
```

Covariance Parameter Estimates				
Cov Parm	Estimate	Alpha	Lower	Upper
Sire	116.75	0.05	29.9707	7051.37
Residual	463.79	0.05	305.11	789.17

Note: The estimates of the variance components using type3 and reml estimation match here, this is not always the case.

Confidence interval for ρ_I:
A $100(1-\alpha)\%$ confidence interval for ρ_I is given by

$$\left(\frac{F_{obs}-F_{\alpha/2}}{F_{obs}+(n-1)F_{\alpha/2}}, \frac{F_{obs}-F_{1-\alpha/2}}{F_{obs}+(n-1)F_{1-\alpha/2}}\right)$$

where $F_{\alpha/2} = F(\frac{\alpha}{2}, t-1, N-t)$ and F_{obs} is the observed F-ratio for treatment effect from the ANOVA table.

For the sires, $F_{obs} = 3.01$ and $F_{0.025} = 3.179, F_{0.975} = 0.119$.
The formula gives $(-0.01, 0.75)$.

We are 95% confident the intraclass correlation coefficient is between -0.01 and 0.75.

Review of one-way random effects ANOVA

Model:

$$Y_{ij} = \underbrace{\mu}_{\text{fixed}} + \underbrace{T_i}_{\text{random}} + \underbrace{E_{ij}}_{\text{random}} \text{ for } i = 1, 2, \ldots, t \text{ and } j = 1, \ldots, n$$

with

$$T_1, T_2, \ldots, T_t \overset{iid}{\sim} N(0, \sigma_T^2) \text{ independent of } E_{11}, \ldots, E_{tn} \overset{iid}{\sim} N(0, \sigma^2)$$

Remarks:

- ($T_1, T_2, \ldots$ randomly drawn from pop'n of treatment effects.)
- Only three parameters: μ, σ, σ_T^2
- Several functions of these parameters of interest
  - $CV(Y) = \frac{\sqrt{\sigma^2+\sigma_T^2}}{\mu}$
  - $\rho_I = \text{Corr}(Y_{ij}, Y_{ik}) = \frac{\sigma_T^2}{\sigma^2+\sigma_T^2}$
- Two observations from same treatment group not independent

Exercise: match up the formulas for confidence intervals below with their targets, $\rho_I, \sigma^2, \sigma_T^2, \mu$:

$$Y_{++} \pm t(0.025, t-1)\sqrt{\frac{MS[T]}{nt}}$$

$$\left(\frac{F_{obs} - F_{1-\alpha/2}}{F_{obs} + (n-1)F_{1-\alpha/2}}, \frac{F_{obs} - F_{\alpha/2}}{F_{obs} + (n-1)F_{\alpha/2}}\right)$$

$$\left(\frac{(N-t)MS[E]}{\chi^2_{\alpha/2}}, \frac{(N-t)MS[E]}{\chi^2_{1-\alpha/2}}\right)$$

$$\left(\frac{\widehat{df}\hat{\sigma}_T^2}{\chi^2_{\alpha/2,\widehat{df}}}, \frac{\widehat{df}\hat{\sigma}_T^2}{\chi^2_{1-\alpha/2,\widehat{df}}}\right)$$

Chapter 11

ST 512 - Nested Designs

Readings: 13.7-13.8 634-642

So far, we have considered complete factorial experiments. That is, experiments where the responses were measured at every possible combination of levels of the experimental factors.

Now we will consider factors that are *nested*. That is, in which responses are measured at a subset of the combination of factor levels.

Crossed vs Nested factors
Crossed - The levels of factor A are said to be *crossed* with the levels of factor B if every level of A occurs in combination with every level of B.

Nested - Factor B is *nested* in factor A if there is a new set of levels of factor B for every different level of factor A.

A Nested Design Example:
Experiment to study effect of drug and method of administration on fasting blood sugar in diabetic patients

- First factor is drug: brand I tablet, brand II tablet, insulin injection
- Second factor is type of administration (see table)

Drug (i)	Type of Administration (j)	Mean $\bar{y}_{(i)j}$	Variance $s^2_{(i)j}$	Mean (Grand mean = $\bar{y}_{+++} = 22.3$). $\bar{y}_{(i)+}$
Brand I tablet	$30mg \times 1$	15.7	6.3	17.7
	$15mg \times 2$	19.7	9.3	
Brand II tablet	$20mg \times 1$	20	1	18.7
	$10mg \times 2$	17.3	6.3	
Insulin injection	before breakfast	28	4	30.5
	before supper	33	9	

We use the notation B(A) for factor B nested in factor A, and we use $b_{(i)j}$ to say level j of factor B for the i^{th} level of factor A.
Here, factor B is administration and factor A is Drug. We would say administration is nested in drug or B(A).

The level of factor B, 'before supper', would be labeled $b_{(3)1}$ and '10 mg x 2' would be $b_{(2)2}$.

In the following examples identify which pairs of factors are crossed and which are nested.

- The amount of vitamin A in jars of baby food might vary from brand to brand and might also vary between flavors of the same brand. To study the effect of these two factors on vitamin A content, a researcher randomly selected the three major brands of baby food in the area.
  For Brand 1 they selected carrot and pear, for Brand 2 sweet potato and green bean, and for Brand 3 pea and squash. Five jars were selected for each treatment.

- Gum arabic is used to lengthen the shelf length of emulsions. It comes from acacia trees and is processed for use in emulsions. Eight raw gum arabic samples are obtained from each of two different varieties of acacia tree (for a total of sixteen samples.) Four samples from each variety of acacia tree are randomly assigned an experimental treatment (the others act as a control). The sixteen samples are dried, and an emulsion made from each. The response is the time until the emulsion begins to separate.

- A total of 30 participants each read a story and were asked to recall some facts from the story. The dependent variable was number of facts recalled. There were two variables of interest. One was the story setting (stories were either familiar or exotic). For each type of story setting there were three different stories (for a total of 6 totally different stories). Each story was read by 5 people and the number of facts recalled were recorded.

Again, we use the notation B(A) for factor B nested in factor A, and we use $b_{(i)j}$ to say level j of factor B for the i^{th} level of factor A. The model for analysis of a two-factor nested design is:

$$Y_{ijk} = \mu + \tau_i + \beta_{(i)j} + E_{ijk}, \quad E_{ijk} \sim N(0, \sigma^2)$$

$i = 1, 2, \ldots, a \quad j = 1, 2, \ldots, b_i \quad k = 1, 2, \ldots, n_{ij}$ with some sum to zero constraints

What are the interpretations of each parameter in the above model?

- μ
- τ_i
- $\beta_{(i)j}$

A partial ANOVA table (with expected mean squares) for a nested design is given below, fill it in

source	df	SS	MS	EMS	F
A	$a-1$	SS(A)	$MS(A) =$	$\sigma^2 + nb\phi_A^2$	F=
B(A)	$a\sum_{i=1}^{a}(b_i - 1)$	SS(B(A))	$MS(B(A)) =$	$\sigma^2 + n\phi_{B(A)}^2$	F=
Error	$N - a\sum_{i=1}^{a} b_i$	SS(E)	$MS(E) =$	σ^2	

The null hypotheses of interest for a nested experiment are

$$H_0 : \tau_i = 0 \; for \; all \; i = 1, ..., a$$

$$H_0 : \beta_{(i)j} = 0 \; for \; all \; i = 1, ..., a, \; j = 1, ..., b$$

and these hypotheses jointly (the global test). In words, what are each of these hypotheses testing?

Assuming the above hypotheses are true, what value is expected for each F in the above ANOVA table? What values give evidence against these H_0's?

There is no interaction being tested here, why do you think that is not something we look at for a nested design?

Another Example: The amount of readily soluble phosphorus in a large number of soil samples was to be determined in a lab that employed six technicians. **Three worked in the morning and three at night.** To determine whether the measured values of phosphorus (lb/acre) were affected by the time of day (am or pm) and the technician making the measurement, 24 identical specimen samples were assigned to the six technicians at random (each got 4 samples). They analyzed the samples and the data was recorded.

Time	Technician	Response	Mean $\bar{y}_{(i)j}$
AM	1	42,44,43,44	$\bar{y}_{(1)1} = 43.25$
AM	2	43,44,45,42	$\bar{y}_{(1)2} = 43.50$
AM	3	47,46,47,43	$\bar{y}_{(1)3} = 45.75$
PM	1	50,49,52,50	$\bar{y}_{(2)1} = 50.25$
PM	2	49,48,49,47	$\bar{y}_{(2)2} = 48.25$
PM	3	47,51,46,48	$\bar{y}_{(2)3} = 48.00$

Note: Technician 1 in the AM is different than technician 1 in the PM.

What are the factors in this study? Which factor is nested in the other?

(Let's consider technician as fixed for now.) Write out the statistical model for these data and interpret τ_i and $\beta_{(i)j}$ in terms of the experiment.

The model is fit in SAS using the code here:

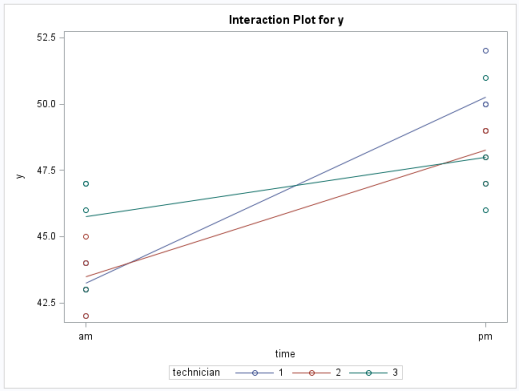
```
proc glm data=phosph;
class time technician;
model y=time technician(time);
run;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	158.0000000	31.6000000	14.22	<.0001
Error	18	40.0000000	2.2222222		
Corrected Total	23	198.0000000			

R-Square	Coeff Var	Root MSE	y Mean
0.797980	3.205832	1.490712	46.50000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
time	1	130.6666667	130.6666667	58.80	<.0001
technician(time)	4	27.3333333	6.8333333	3.08	0.0429

Source	DF	Type III SS	Mean Square	F Value	Pr > F
time	1	130.6666667	130.6666667	58.80	<.0001
technician(time)	4	27.3333333	6.8333333	3.08	0.0429



1. The first step is to check that anything in our model is useful. Which p-value tests this? What is the null hypothesis?

2. The next step in the analysis is to test the significance of the difference between technicians. Which p-value tests this? What is the null hypothesis? Interpret your conclusions.

3. We also want to inspect the difference between am and pm. Which p-value tests this? What is the null hypothesis? Interpret your conclusions.

4. Is the interaction plot given by SAS meaningful? Why?

5. Since both are significant, we want to compare the difference between technicians at a given time, then also compare the two time levels.

   These can be inspected by adding the following SAS command:

```
proc glm data=phosph plots=NONE; class time technician;
model y=time technician(time)/clparm;
lsmeans time technician(time)/adjust=tukey cl;
estimate 'effect of Time' intercept 0 time 3 -3 technician(time) 1 1 1 -1 -1 -1/divisor=3;
estimate 'effect of Tech 1/2 within Time=AM' intercept 0 time 0 0 technician(time) 1 -1 0 0 0 0;
estimate 'effect of Tech 1/3 within Time=AM' intercept 0 time 0 0 technician(time) 1 0 -1 0 0 0;
estimate 'effect of Tech 2/3 within Time=AM' intercept 0 time 0 0 technician(time) 0 1 -1  0 0 0;
estimate 'effect of Tech 1/2 within Time=PM' intercept 0 time 0 0 technician(time) 0 0 0 1 -1 0;
estimate 'effect of Tech 1/3 within Time=PM' intercept 0 time 0 0 technician(time) 0 0 0 1 0 -1;
estimate 'effect of Tech 2/3 within Time=PM' intercept 0 time 0 0 technician(time) 0 0 0 0 1 -1; run;
```

   Output from estimate statements (note, these CI's are not corrected for Multiple comparisons!)

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
effect of Time	-4.66666667	0.60858062	-7.67	<.0001	-5.94524710	-3.38808623
effect of Tech 1/2 within Time=AM	-0.25000000	1.05409255	-0.24	0.8152	-2.46456628	1.96456628
effect of Tech 1/3 within Time=AM	-2.50000000	1.05409255	-2.37	0.0291	-4.71456628	-0.28543372
effect of Tech 2/3 within Time=AM	-2.25000000	1.05409255	-2.13	0.0468	-4.46456628	-0.03543372
effect of Tech 1/2 within Time=PM	2.00000000	1.05409255	1.90	0.0739	-0.21456628	4.21456628
effect of Tech 1/3 within Time=PM	2.25000000	1.05409255	2.13	0.0468	0.03543372	4.46456628
effect of Tech 2/3 within Time=PM	0.25000000	1.05409255	0.24	0.8152	-1.96456628	2.46456628

Output from lsmeans statements (note, we probably don't care about all of the differences between technicians!)

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey

time	y LSMEAN	H0:LSMean1=LSMean2 Pr > |t|
am	44.1666667	<.0001
pm	48.8333333	

time	y LSMEAN	95% Confidence Limits	
am	44.166667	43.262574	45.070760
pm	48.833333	47.929240	49.737426

Least Squares Means for Effect time

i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-4.666667	-5.945202	-3.388131

technician	time	y LSMEAN	95% Confidence Limits	
1	am	43.250000	41.684065	44.815935
2	am	43.500000	41.934065	45.065935
3	am	45.750000	44.184065	47.315935
1	pm	50.250000	48.684065	51.815935
2	pm	48.250000	46.684065	49.815935
3	pm	48.000000	46.434065	49.565935

Least Squares Means for Effect technician(time)

i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-0.250000	-3.599943	3.099943
1	3	-2.500000	-5.849943	0.849943
1	4	-7.000000	-10.349943	-3.650057
1	5	-5.000000	-8.349943	-1.650057
1	6	-4.750000	-8.099943	-1.400057
2	3	-2.250000	-5.599943	1.099943
2	4	-6.750000	-10.099943	-3.400057
2	5	-4.750000	-8.099943	-1.400057
2	6	-4.500000	-7.849943	-1.150057
3	4	-4.500000	-7.849943	-1.150057
3	5	-2.500000	-5.849943	0.849943
3	6	-2.250000	-5.599943	1.099943
4	5	2.000000	-1.349943	5.349943
4	6	2.250000	-1.099943	5.599943
5	6	0.250000	-3.099943	3.599943

Chapter 12

ST 512 - Mixed Models

Readings: 14.1-14.8 643-699

Models with both fixed and random factors are called **Mixed Models**. Let's do some practice picking out fixed and random factors along with nested and crossed design.

Two-factor designs examples (some repeated from previous notes)

1. Entomologist records energy expended (y) by $N = 27$ honeybees
   - at three TEMPERATURES ($20, 30, 40^oC$)
   - consuming three levels of SUCROSE ($20\%, 40\%, 60\%$)

Temp	Suc	Sample		
20	20	3.1	3.7	4.7
20	40	5.5	6.7	7.3
20	60	7.9	9.2	9.3
30	20	6	6.9	7.5
30	40	11.5	12.9	13.4
30	60	17.5	15.8	14.7
40	20	7.7	8.3	9.5
40	40	15.7	14.3	15.9
40	60	19.1	18.0	19.9

   - First factor:
   - Second factor:
   - Fixed or random?
   - Crossed or nested?
   - Model:

$$Y_{ijk} = \mu + \qquad\qquad\qquad\qquad\qquad\qquad + E_{ijk}$$

2. Experiment to study effect of drug and method of administration on fasting blood sugar in a random sample of $N = 18$ diabetic patients.

Drug (i)	Type of Administration (j)	Mean $\bar{y}_{j(i)}$	Variance $s^2_{j(i)}$
($i = 1$) Brand I tablet	($j = 1$)$30mg \times 1$	15.7	6.3
	($j = 2$)$15mg \times 2$	19.7	9.3
($i = 2$) Brand II tablet	($j = 1$)$20mg \times 1$	20	1
	($j = 2$)$10mg \times 2$	17.3	6.3
($i = 3$) Insulin injection	($j = 1$) before breakfast	28	4
	($j = 2$) before supper	33	9

- First factor:
- Second factor:
- Fixed or random?
- Crossed or nested?
- Model: $Y_{ijk} = \mu +$ $+E_{ijk}$

3. An experiment is conducted to determine variability among laboratories (interlaboratory differences) in their assessment of bacterial concentration in milk after pasteurization. Milk w/ various degrees of contamination was tested by randomly drawing four samples of milk from a collection of cartons at various stages of spoilage. Each of the four samples was split into 10 parts and two were sent to each of the 5 laboratories. Y is colony-forming units/μl. Labs think they're receiving 8 independent samples.

	Sample			
Lab	1	2	3	4
1	2200	3000	210	270
	2200	2900	200	260
2	2600	3600	290	360
	2500	3500	240	380
3	1900	2500	160	230
	2100	2200	200	230
4	2600	2800	330	350
	4300	1800	340	290
5	4000	4800	370	500
	3900	4800	340	480

- First factor:
- Second factor:
- Fixed or random?
- Crossed or nested?
- Model: $Y_{ijk} = \mu +$ $+E_{ijk}$

4. An expt measures *Campylobacter* counts in $N = 120$ chickens in a processing plant, at four locations, over three days. Means (std) for $n = 10$ chickens sampled at each location tabulated below:

   - Student visits plant on three random sampled winter days.
   - On each day he samples $n = 10$ chickens at each of four locations, or sites, along the washing line: (before first washer, after 3rd washer, after microbial rinse, after chill tank)

	Location			
	Before	After	After	After
Day	Washer	Washer	mic. rinse	chill tank
1	70070.00	48310.00	12020.00	11790.00
	(79034.49)	(34166.80)	(3807.24)	(7832.05)
2	75890.00	52020.00	8090.00	8690.00
	(74551.32)	(17686.27)	(4848.01)	(5526.19)
3	95260.00	33170.00	6200.00	8370.00
	(03176.00)	(22259.08)	(5028.81)	(5720.15)

Data courtesy of Michael Bashor, General Mills

   - First factor:
   - Second factor:
   - Fixed or random?
   - Crossed or nested?
   - Model:

$$Y_{ijk} = \mu + \qquad\qquad\qquad\qquad\qquad\qquad + E_{ijk}$$

5. An experiment to assess the variability of a particular acid among plants was done. Many plants were planted and 4 were randomly selected. Then three leaves were selected from each plant to be measured.

Plant i	1			2			3			4		
Leaf j	1	2	3	1	2	3	1	2	3	1	2	3
$k = 1$	11.2	16.5	18.3	14.1	19.0	11.9	15.3	19.5	16.5	7.3	8.9	11.3
$k = 2$	11.6	16.8	18.7	13.8	18.5	12.4	15.9	20.1	17.2	7.8	9.4	10.9
$k = 3$	12.0	16.1	19.0	14.2	18.2	12.0	16.0	19.3	16.9	7.0	9.3	10.5

Data from Neter, et al (1996)

   - First factor:
   - Second factor:
   - Fixed or random?
   - Crossed or nested?
   - Model:

$$Y_{ijk} = \mu + \qquad\qquad\qquad\qquad\qquad\qquad + E_{ijk}$$

6. 5 treatments of light intensity were assigned randomly to 10 pots of plants. Each pot had two seedlings per pot. For each seedling the plant height was measured for a total of 20 measurements. (See Table 14.2 from Rao.)

Treatment	Pot	Seedling 1	Seedling 2
1	1	32.94	35.98
1	2	34.76	32.40
2	1	30.55	32.64
2	2	32.37	32.04
3	1	31.23	31.09
3	2	30.62	30.42
4	1	34.41	34.88
4	2	34.07	33.87
5	1	35.61	35.00
5	2	33.65	32.91

- First factor:
- Second factor:
- Fixed or random?
- Crossed or nested?
- Model: $Y_{ijk} = \mu +$ $+E_{ijk}$

Recap: Six types of two-factor models possible with fixed and/or random effects that are either crossed or nested.

Experiment Number			Model	Identifier
___	Y_{ijk}	=	$\mu + A_i + B_j + (AB)_{ij} + E_{ijk}$	crossed/random
___	Y_{ijk}	=	$\mu + \alpha_i + \beta_{(i)j} + E_{ijk}$	nested/fixed
___	Y_{ijk}	=	$\mu + A_i + B_{(i)j} + E_{ijk}$	nested/random
___	Y_{ijk}	=	$\mu + \alpha_i + B_j + (\alpha B)_{ij} + E_{ijk}$	crossed/mixed
___	Y_{ijk}	=	$\mu + \alpha_i + B_{(i)j} + E_{ijk}$	nested/mixed
___	Y_{ijk}	=	$\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + E_{ijk}$	crossed/fixed

- GREEK symbols parameterize FIXED, unknown treatment means
- CAPITAL letters represent RANDOM effects

In the models above there are many contraints

- for first model above, $A_i, B_i, (AB)_{ij}$ are all independent
- for second model above, $\sum \alpha_i = \sum_j \beta_{(i)j} \equiv 0$
- for third model above, $A_i, B_{(i)j}$ are all independent
- for fourth model above, $\sum \alpha_i = 0$ and $B_j, (\alpha B)_{ij}$ are all independent
- for fifth model above, $\sum \alpha_i = 0$
- for sixth model above, $\sum \alpha_i = \sum \beta_j = \sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} \equiv 0$

One method for making inference in Mixed Models is to equate mean squares to find appropriate tests. Below are tables of mean squares for these different models:

Tables of expected means squares (EMS):
When factors A and B are CROSSED, and no sum-to-zero assumptions are made on random effects, expected means associated with sums of squares are given in the table below:

Source	*df*	A, B fixed	A, B random	A fixed B random
A	$a-1$	$\sigma^2 + nb\psi_A^2$	$\sigma^2 + nb\sigma_A^2 + n\sigma_{AB}^2$	$\sigma^2 + nb\psi_A^2 + n\sigma_{\alpha B}^2$
B	$b-1$	$\sigma^2 + na\psi_B^2$	$\sigma^2 + na\sigma_B^2 + n\sigma_{AB}^2$	$\sigma^2 + na\sigma_B^2 + n\sigma_{\alpha B}^2$
AB	$(a-1)(b-1)$	$\sigma^2 + n\psi_{AB}^2$	$\sigma^2 + n\sigma_{AB}^2$	$\sigma^2 + n\sigma_{\alpha B}^2$
Error	$ab(n-1)$	σ^2	σ^2	σ^2

When factor B is NESTED in factor A, expected means associated with sums of squares are given in the table below:

Source	*df*	A, B fixed	A, B random	A fixed B random
A	$a-1$	$\sigma^2 + nb\psi_A^2$	$\sigma^2 + nb\sigma_A^2 + n\sigma_{B(A)}^2$	$\sigma^2 + nb\psi_A^2 + n\sigma_{B(A)}^2$
$B(A)$	$a(b-1)$	$\sigma^2 + n\psi_{B(A)}^2$	$\sigma^2 + n\sigma_{B(A)}^2$	$\sigma^2 + n\sigma_{B(A)}^2$
Error	$ab(n-1)$	σ^2	σ^2	σ^2

where ψ^2 and σ^2 values are defined below.

$$\psi_A^2 = \frac{1}{a-1}\sum_1^a \alpha_i^2 \quad \text{effect size of factor } A$$

$$\psi_B^2 = \frac{1}{b-1}\sum_1^b \beta_i^2 \quad \text{effect size of factor } B$$

$$\psi_{AB}^2 = \frac{1}{(a-1)(b-1)}\sum_{i=1}^a\sum_{j=1}^b (\alpha\beta)_{ij}^2 \quad \text{effect size of interaction}$$

$$\psi_{B(A)}^2 = \frac{1}{a(b-1)}\sum_{i=1}^a\sum_{j=1}^b \beta_{(i)j}^2 \quad \text{effect size of factor } B$$

$$\sigma_A^2 = \text{Var}(A_i) \quad \text{variance component for factor } A$$

$$\sigma_B^2 = \text{Var}(B_i) \quad \text{variance component for factor } B$$

$$\sigma_{AB}^2 = \text{Var}((AB)_{ij}) \quad \text{variance component for interaction}$$

$$\sigma_{B(A)}^2 = \text{Var}(B_{(i)j}) \quad \text{variance component for factor } B$$

$$\sigma^2 = \text{Var}(E_{ijk)}) \quad \text{error variance}$$

We can use these expected mean squares to determine tests for effects.

Example: For A fixed, B random in a crossed experiment

$$E(MS(A)) = \sigma^2 + nb\psi_A^2 + n\sigma_{\alpha B}^2$$

$$E(MS(AB)) = \sigma^2 + n\sigma_{\alpha B}^2$$

Thus, a test for the effect of A can be derived as

$$F = MS(A)/MS(AB)$$

If $H_0 : \alpha_i = 0$ for all i is true, then this F should be approximately 1.
If H_A : at least 1 $\alpha_i \neq 0$ is true, then this F should be larger than 1.

To determine if we are far enough from 1 we should compare to the appropriate F critical value.

Help with computing expected mean squares for balanced designs (without sum-to-zero assumptions on random effects)

1. If a factor X with index i is random then $EMS(X)$ is a linear combo of σ^2 and varcomps for all random effects containing index i. Coefficients for varcomps are limits of indexes NOT listed (summed over) in random effects.
2. If a factor X is fixed. Treat it like it is random and then just replace the varcomp for X with the effect size, ψ_X^2.

Analysis of milk example - F-tests and estimating variance components.
Recall: We have crossed random factors.

1. To test for interaction effect, use $F_{AB} = \frac{MS[AB]}{MS[E]}$ vs $F(\alpha, (a-1)(b-1), ab(n-1))$
2. To test for main effect of A, use $F_A = \frac{MS[A]}{MS[AB]}$ vs $F(\alpha, a-1, (a-1)(b-1))$
3. To test for main effect of B, use $F_B = \frac{MS[B]}{MS[AB]}$ vs $F(\alpha, b-1, (a-1)(b-1))$

Note the departure from fixed effects analysis, where $MS[E]$ is always used in the denominator!

If we use proc glm for analysis, we get the wrong analysis! (glm not intended for mixed models) Note: we are analyzing $ln(y)$ rather than y.

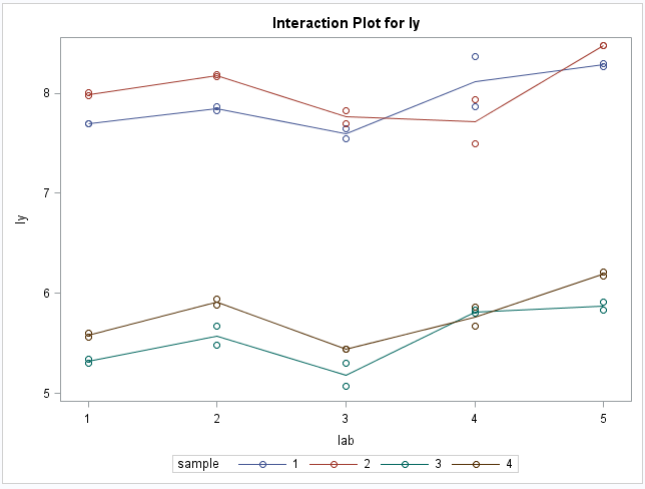
```
proc glm; class lab sample;
model ly=sample|lab;
random sample lab sample*lab; run;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	19	56.03510844	2.94921623	191.44	<.0001
Error	20	0.30810726	0.01540536		
Corrected Total	39	56.34321569			

R-Square	Coeff Var	Root MSE	ly Mean
0.994532	1.821098	0.124118	6.815577

Source	DF	Type I SS	Mean Square	F Value	Pr > F
sample	3	53.18978788	17.72992929	1150.89	<.0001
lab	4	2.30248803	0.57562201	37.37	<.0001
lab*sample	12	0.54283253	0.04523604	2.94	0.0161

Source	DF	Type III SS	Mean Square	F Value	Pr > F
sample	3	53.18978788	17.72992929	1150.89	<.0001
lab	4	2.30248803	0.57562201	37.37	<.0001
lab*sample	12	0.54283253	0.04523604	2.94	0.0161



F-statistics divide by $MS(E)$ not appropriate error term for testing A and B main effects!

To test for the main effect of the random factor A, $H_0 : \sigma_A^2 = 0$, use the F-ratio $F = MS[A]/MS[AB]$. Under H_0, $F \sim F(a-1, ab(n-1)) = F(4, 20)$, which has $F(0.05, 4, 20) = 2.87$, yielding the $\alpha = 0.05$ critical region reject H_0 if $F_{obs} > 2.87$ The correct F-ratio and p-value for testing for random LAB (A) effect:

$$F = \frac{MS[A]}{MS[AB]} = \frac{0.5756}{0.0452} = 12.72 \quad (p = 0.0003)$$

Likewise, find the correct test for the sample (B) effect. (Hint: $F(0.05, 3, 20) = 3.10$)

Can get correct analysis in proc glm by adding in the following line:

```
proc glm; class lab sample;
model ly=sample|lab;
random sample lab sample*lab;
test h=lab sample e=sample*lab; run;
```

Source	Type III Expected Mean Square
sample	Var(Error) + 2 Var(lab*sample) + 10 Var(sample)
lab	Var(Error) + 2 Var(lab*sample) + 8 Var(lab)
lab*sample	Var(Error) + 2 Var(lab*sample)

Tests of Hypotheses Using the Type III MS for lab*sample as an Error Term

Source	DF	Type III SS	Mean Square	F Value	Pr > F
lab	4	2.30248803	0.57562201	12.72	0.0003
sample	3	53.18978788	17.72992929	391.94	<.0001

Now the appropriate tests are reported. Rather than go through this, we can just use proc mixed! (see below)

Estimating variance components: The estimated variance components satisfy the following system of equations:

$$\begin{aligned} MS[E] &= \hat{\sigma}^2 \\ MS[AB] &= \hat{\sigma}^2 + n\hat{\sigma}^2_{AB} \\ &= \hat{\sigma}^2 + 2\hat{\sigma}^2_{AB} \\ MS[A] &= \hat{\sigma}^2 + nb\hat{\sigma}^2_A + n\hat{\sigma}^2_{AB} \\ &= \hat{\sigma}^2 + 8\hat{\sigma}^2_A + 2\hat{\sigma}^2_{AB} \\ MS[B] &= \hat{\sigma}^2 + na\hat{\sigma}^2_B + n\hat{\sigma}^2_{AB} \\ &= \hat{\sigma}^2 + 10\hat{\sigma}^2_B + 2\hat{\sigma}^2_{AB} \end{aligned}$$

Substitution of

$$\begin{aligned} MS[E] &= 0.0154 \\ MS[AB] &= 0.0452 \\ MS[A] &= 0.5756 \\ MS[B] &= 17.7299 \end{aligned}$$

into the system of equations yields the estimated variance components:

$$\begin{array}{rcccccc} \hat{\sigma}^2 & = & MS[E] & = & & = & 0.0154 \\ \hat{\sigma}^2_{AB} & = & \frac{MS[AB]-MS[E]}{n} & = & \frac{0.0452-0.0154}{2} & = & 0.01492 \\ \hat{\sigma}^2_{A} & = & \frac{MS[A]-MS[AB]}{nb} & = & \frac{0.5756-0.0452}{8} & = & 0.0663 \\ \hat{\sigma}^2_{B} & = & \frac{MS[B]-MS[AB]}{na} & = & \frac{17.7299-0.0452}{10} & = & 1.768 \end{array}$$

Proc mixed is our best bet for analyzing this experiment:

```
proc mixed method=type3;
class lab sample;
model ly=;
random sample|lab;
run;
```

Type 3 Analysis of Variance

Source	DF	Sum of Squares	Mean Square	Expected Mean Square	Error Term	Error DF	F Value	Pr > F
sample	3	53.189788	17.729929	Var(Residual) + 2 Var(lab*sample) + 10 Var(sample)	MS(lab*sample)	12	391.94	<.0001
lab	4	2.302488	0.575622	Var(Residual) + 2 Var(lab*sample) + 8 Var(lab)	MS(lab*sample)	12	12.72	0.0003
lab*sample	12	0.542833	0.045236	Var(Residual) + 2 Var(lab*sample)	MS(Residual)	20	2.94	0.0161
Residual	20	0.308107	0.015405	Var(Residual)	.	.	.	.

Covariance Parameter Estimates

Cov Parm	Estimate
sample	1.7685
lab	0.06630
lab*sample	0.01492
Residual	0.01541

So, what is the conclusion from the analysis of this crossed, random effects experiment?

- There is evidence of variability due to laboratory×sample interaction; interlaboratory effects vary by sample.
- The estimated parameters (μ and all variance components) of the model

$$Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + E_{ijk}$$

are

$$\begin{aligned} \hat{\sigma}^2 &= 0.0154 \\ \hat{\sigma}^2_{AB} &= 0.0149 \\ \hat{\sigma}^2_{A} &= 0.0663 \\ \hat{\sigma}^2_{B} &= 1.7685 \\ \hat{\mu} &= 6.82(\text{log scale}) \end{aligned}$$

- The standard error of $\bar{Y}_{+++}$ can be derived by

$$\begin{aligned} \bar{Y}_{+++} &= \mu + \bar{A}_{+} + \bar{B}_{+} + \overline{(AB)}_{++} + \bar{E}_{+++} \\ \text{Var}(\bar{Y}_{+++}) &= \text{Var}(\bar{A}_{+}) + \text{Var}(\bar{B}_{+}) + \text{Var}(\overline{(AB)}_{++}) + \text{Var}(\bar{E}_{+++}) \\ &= \frac{\sigma^2_A}{a} + \frac{\sigma^2_B}{b} + \frac{\sigma^2_{AB}}{ab} + \frac{\sigma^2}{abn} \end{aligned}$$

Estimation of standard error and approximation of *df*:

The standard error

$$SE(\bar{Y}_{+++}) = \sqrt{\frac{\sigma^2_A}{a} + \frac{\sigma^2_B}{b} + \frac{\sigma^2_{AB}}{ab} + \frac{\sigma^2}{abn}}$$

can be estimated by substitution of estimated variance components ($\hat{\sigma}^2$), which leads to

$$\begin{aligned} \widehat{SE}(\bar{Y}_{+++}) &= \sqrt{\frac{\hat{\sigma}^2_A}{a} + \frac{\hat{\sigma}^2_B}{b} + \frac{\hat{\sigma}^2_{AB}}{ab} + \frac{\hat{\sigma}^2}{abn}} \\ &= \text{lots of algebra and cancellations} \\ &= \sqrt{\frac{1}{nab}\left(MS[A] + MS[B] - MS[AB]\right)} \end{aligned}$$

For the milk data, we have

$$\widehat{SE}(\bar{Y}_{+++}) = \sqrt{\frac{1}{40}(0.58 + 17.73 - 0.05)} = 0.6757$$

For a 95% confidence interval, we have a problem: we don't know how many *df* are associated with a t statistic based on this estimated SE.

Satterthwaite's approximation to degrees of freedom
To approximate the df associated with a t statistic based on a standard error of the form

$$\sqrt{c_1 MS_1 + c_2 MS_2 + \cdots + c_k MS_k}$$

(a linear combination of mean square terms), use the **Satterthwaite approximation:**

$$\widehat{df} = \frac{(c_1 MS_1 + c_2 MS_2 + \cdots + c_k MS_k)^2}{(c_1 MS_1)^2/df_1 + (c_2 MS_2)^2/df_2 + \cdots + (c_k MS_k)^2/df_k}$$

The degrees of freedom associated with $\widehat{SE}(\bar{Y}_{+++})$ is approximated by

$$\widehat{df} = \frac{(0.6757)^4}{(\frac{1}{40}17.73)^2/3 + (\frac{1}{40}0.58)^2/4 + (\frac{1}{40}0.045)^2/12} = 3.18$$

Using $t(0.025, 3.18) = 3.08$, a 95% confidence interval for the (log) mean μ among the population of all labs and samples is given by

$$6.82 \pm 3.08(0.6757) = 6.82 \pm 2.08$$

In proc mixed we can get this using:

```
proc mixed cl method=type3;
   class sample lab;
   model ly=/ddfm=satterth cl;
   random sample|lab;
run;
```

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Intercept	6.8156	0.6757	3.18	10.09	0.0016	0.05	4.7325	8.8987

More Two-factor mixed models analysis examples

- Recall the Campylobacter count chicken experiment:
  - Crossed design with two factors
    - Location (4 levels)
    - Day (3 levels)
  - $n = 10$ chickens per combo for a total of $N = 120$ observations
  - Location of measurement is fixed, Day is random and the factors are crossed (4×3 layout)

Recall: Model being fit is

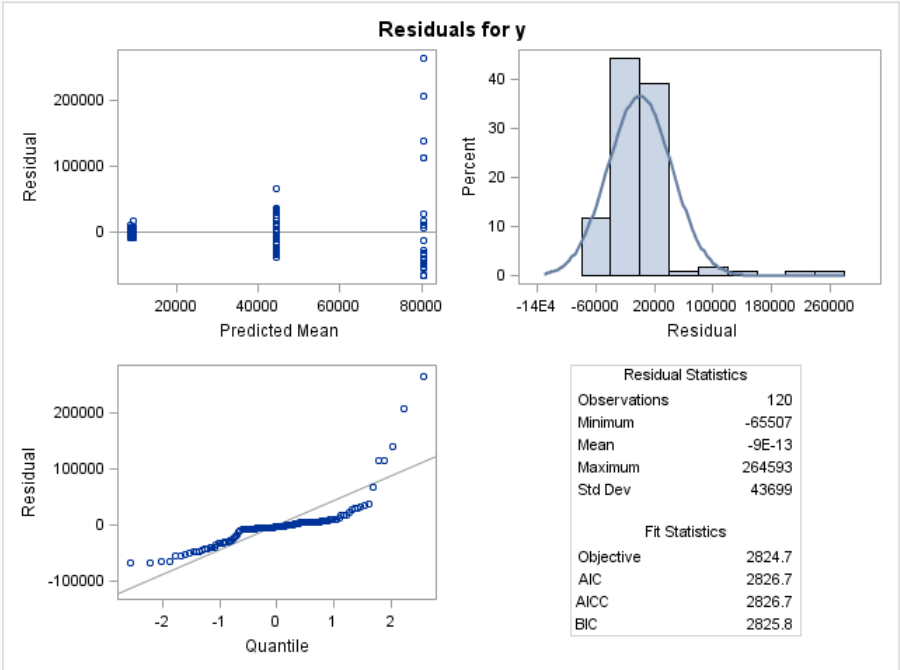
$$Y_{ijk} = \mu + \alpha_i + B_j + (\alpha B)_{ij} + E_{ijk}$$

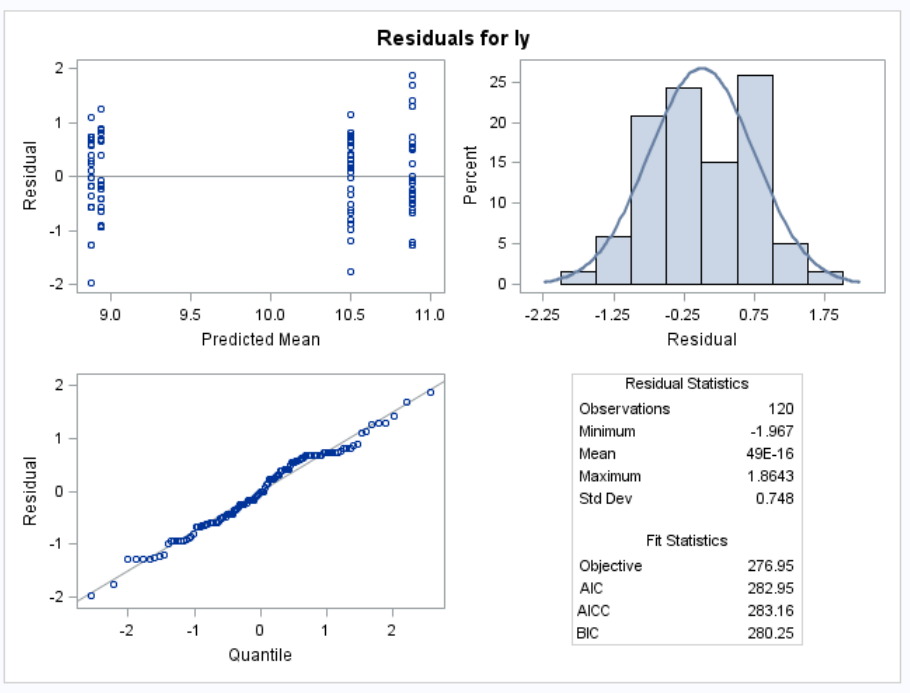
with variance components $\sigma^2_B, \sigma^2_{\alpha B}, \sigma^2$.

Fixed Factor A: location
Random Factor B: day

The Mixed model from previous was fit to this data using the code:

```
proc mixed data=bashor method=type3 plots=all;
class day location;
model ly=location; *ly=log(y) was used as non-constant variance was evident when using y as response
random day day*location;
lsmeans location/adjust=tukey cl;
run;
```





First plot = residual plot using y as response, second plot = residual plot using $log(y)$ as response. (Note: would also check conditional residuals.)

Type 3 Analysis of Variance

Source	DF	Sum of Squares	Mean Square	Expected Mean Square	Error Term	Error DF	F Value	Pr > F
location	3	97.865388	32.621796	Var(Residual) + 10 Var(day*location) + Q(location)	MS(day*location)	6	43.17	0.0002
day	2	2.787355	1.393677	Var(Residual) + 10 Var(day*location) + 40 Var(day)	MS(day*location)	6	1.84	0.2375
day*location	6	4.533565	0.755594	Var(Residual) + 10 Var(day*location)	MS(Residual)	108	1.38	0.2303
Residual	108	59.254946	0.548657	Var(Residual)	.	.	.	.

Covariance Parameter Estimates

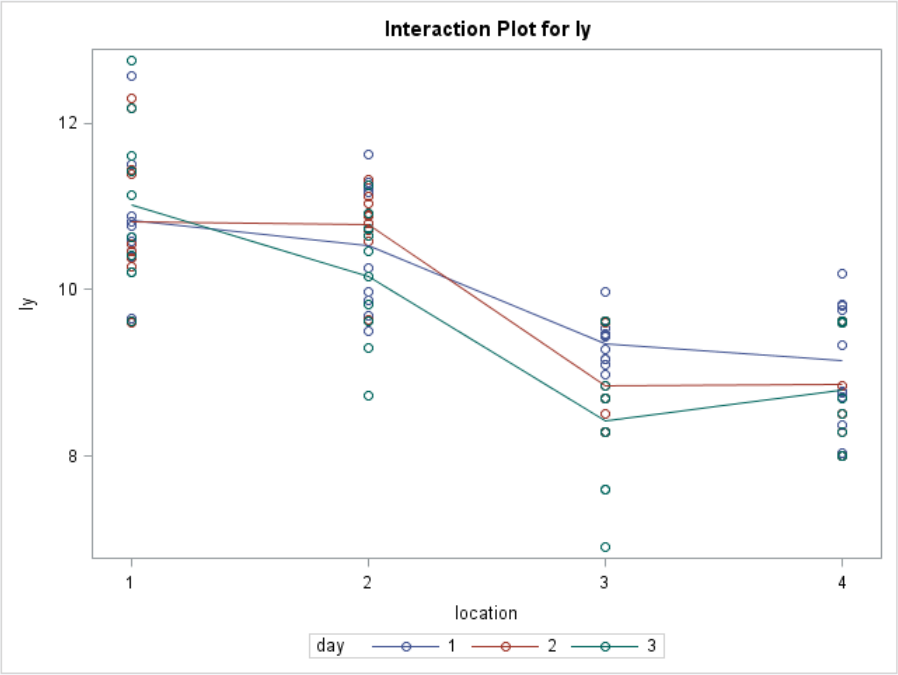
Cov Parm	Estimate
day	0.01595
day*location	0.02069
Residual	0.5487

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
location	3	6	43.17	0.0002

Least Squares Means									
Effect	location	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
location	1	10.8870	0.1747	6	62.33	<.0001	0.05	10.4596	11.3144
location	2	10.4953	0.1747	6	60.09	<.0001	0.05	10.0680	10.9227
location	3	8.8745	0.1747	6	50.81	<.0001	0.05	8.4472	9.3019
location	4	8.9394	0.1747	6	51.18	<.0001	0.05	8.5120	9.3668

Differences of Least Squares Means														
Effect	location	_location	Estimate	Standard Error	DF	t Value	Pr > |t|	Adjustment	Adj P	Alpha	Lower	Upper	Adj Lower	Adj Upper
location	1	2	0.3917	0.2244	6	1.75	0.1316	Tukey-Kramer	0.3801	0.05	-0.1575	0.9409	-0.3853	1.1686
location	1	3	2.0125	0.2244	6	8.97	0.0001	Tukey-Kramer	0.0004	0.05	1.4633	2.5617	1.2355	2.7894
location	1	4	1.9476	0.2244	6	8.68	0.0001	Tukey-Kramer	0.0005	0.05	1.3984	2.4968	1.1707	2.7245
location	2	3	1.6208	0.2244	6	7.22	0.0004	Tukey-Kramer	0.0015	0.05	1.0716	2.1700	0.8439	2.3977
location	2	4	1.5559	0.2244	6	6.93	0.0004	Tukey-Kramer	0.0018	0.05	1.0067	2.1051	0.7790	2.3329
location	3	4	-0.06488	0.2244	6	-0.29	0.7823	Tukey-Kramer	0.9907	0.05	-0.6141	0.4843	-0.8418	0.7121

Output from model above. Below is the interaction plot (found using proc glm).



In the model we estimate the variance components by:

$$\begin{aligned}
\hat{\sigma}^2 &= MS[E] = 0.55 \\
\hat{\sigma}^2_{\alpha B} &= \frac{MS[AB] - MS[E]}{n} \\
&= \frac{0.76 - 0.55}{10} = 0.021 \\
\hat{\sigma}^2_{B} &= \frac{MS[B] - MS[AB]}{na} \\
&= \frac{1.39 - 0.76}{40} = 0.016
\end{aligned}$$

To test $H_0 : \sigma^2_{\alpha B} = 0$, use

$$F_{AB} = \frac{MS[AB]}{MS[E]} = \frac{0.76}{0.55} = 1.38$$

on $(a-1)(b-1) = 6$ and $ab(n-1) = 108$ df. The p-value is 0.2303, providing no evidence of a random day $\times$ location interaction effect.

The variance component for this random effect is estimated by 0.021. Intepretation: Since we failed to reject, there is no evidence that day-to-day variability varies by location. The estimated variance component is itself very small.

Implied correlation structure
What is the correlation of two observations taken on the same day

- at the same location?
- at different locations?

Recall that $Y_{ijk} = \mu + \alpha_i + B_j + (\alpha B)_{ij} + E_{ijk}$.

$$\begin{aligned}
Corr(Y_{ijk_1}, Y_{ijk_2}) &= \frac{\text{Cov}(Y_{ijk_1}, Y_{ijk_2})}{\sigma^2 + \sigma^2_B + \sigma^2_{\alpha B}} \\
&= \frac{\text{Cov}(B_i, B_i) + \text{Cov}((\alpha B)_{ij}, (\alpha B)_{ij})}{\sigma^2 + \sigma^2_B + \sigma^2_{\alpha B}} \\
&= \frac{\sigma^2_B + \sigma^2_{\alpha B}}{\sigma^2 + \sigma^2_B + \sigma^2_{\alpha B}} \\
Corr(Y_{1jk_1}, Y_{2jk_2}) &= \frac{\text{Cov}(Y_{1jk_1}, Y_{2jk_2})}{\sigma^2 + \sigma^2_B + \sigma^2_{\alpha B}} \\
&= \frac{\text{Cov}(B_i, B_i)}{\sigma^2 + \sigma^2_B + \sigma^2_{\alpha B}} \\
&= \frac{\sigma^2_B}{\sigma^2 + \sigma^2_B + \sigma^2_{\alpha B}}
\end{aligned}$$

Estimates of these correlations are

- $\frac{0.016+0.021}{0.016+0.021+0.55} = \frac{0.037}{.587} = 0.063$
- $\frac{0.016}{0.016+0.021+0.55} = \frac{0.016}{.587} = 0.027$

Which is which?

Some analysis of fixed effects

Consider testing for a fixed effect of location. That is, test the hypothesis that average bacteria counts are constant across the locations,

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$$

$$F_A = \frac{MS[A]}{MS[AB]} = \frac{32.6}{0.76} = 43.2$$

on $a - 1 = 3$ and $(a - 1)(b - 1) = 6$ df, which is significant ($p = 0.0002$).

Since significant fixed effect, we want to estimate the pairwise comparisons among location means, such as, $\alpha_4 - \alpha_3$, consider

$$\hat{\theta} = \bar{y}_{4++} - \bar{y}_{3++} = 8.940 - 8.875 = -0.065$$

Note that

$$\text{Var}(\bar{Y}_{4++} - \bar{Y}_{3++}) \neq \sigma^2(\frac{1}{nb} + \frac{1}{nb})$$

Since our Y's are not independent anymore!

What is $SE(\hat{\theta})$ and how can it be estimated?

$$\begin{aligned} \hat{\theta} &= \bar{Y}_{2++} - \bar{Y}_{1++} \\ &= \alpha_2 + \bar{B} + \overline{\alpha B}_{2+} + \overline{E}_{2++} - (\alpha_1 + \bar{B} + \overline{\alpha B}_{1+} + \overline{E}_{1++}) \\ &= \alpha_2 - \alpha_1 + \overline{\alpha B}_{2+} - \overline{\alpha B}_{1+} + \overline{E}_{2++} - \overline{E}_{1++} \end{aligned}$$

which has variance

$$\begin{aligned} \text{Var}(\hat{\theta}) &= \text{Var}(\overline{\alpha B}_{2+}) + \text{Var}(\overline{\alpha B}_{1+}) + \text{Var}(\overline{E}_{2++}) + \text{Var}(\overline{E}_{1++}) \\ &= 2\frac{\sigma^2_{\alpha B}}{b} + 2\frac{\sigma^2}{nb} \\ &= \frac{2}{nb}(\sigma^2 + n\sigma^2_{\alpha B}) \end{aligned}$$

which can be estimated nicely on $(a - 1)(b - 1) = 6$ df by

$$\hat{\text{Var}}(\hat{\theta}) = \frac{2}{nb} MS[AB]$$

for the chickens, where $\overline{y}_{4++} - \overline{y}_{3++} = -0.06$ the SE is

$$\sqrt{\widehat{\text{Var}}(\hat{\theta})} = \sqrt{\frac{2}{3*10}0.76} = 0.22$$

Since $t(0.025, 6) = 2.45$, a 95% c.i. for θ given by $-0.06 \pm 2.45(0.22)$.

Reporting standard errors for sample means of levels of fixed factor, like LOCATION means, is a little messier:

$$\begin{aligned}
\overline{Y}_{i++} &= \mu + \alpha_i + \overline{B} + \overline{\alpha B}_{i+} + \overline{E}_{i++} \\
\text{Var}(\overline{Y}_{i++}) &= \text{Var}(\overline{B}) + \text{Var}(\overline{\alpha B}_{i+}) + \text{Var}(\overline{E}_{i++}) \\
&= \frac{\sigma_B^2}{b} + \frac{\sigma_{\alpha B}^2}{b} + \frac{\sigma^2}{nb} \\
&= \frac{1}{nb}(n\sigma_B^2 + n\sigma_{\alpha B}^2 + \sigma^2) \\
&\quad \text{estimated by} \\
\widehat{\text{Var}}(\overline{Y}_{i++}) &= \frac{1}{nb}(n\hat{\sigma}_B^2 + n\hat{\sigma}_{\alpha B}^2 + \hat{\sigma}^2) \\
&= \text{algebra yields a linear combo of multiple EMS terms} \\
&= \frac{1}{nab}\{(a-1)EMS[AB] + EMS[B]\}
\end{aligned}$$

The standard error is estimated easily enough:

$$\begin{aligned}
\widehat{SE}(\overline{Y_{i++}}) &= \sqrt{\frac{1}{nab}\{(a-1)MS[AB] + MS[B]\}} \\
&= \sqrt{\frac{1}{120}\{(4-1)0.76 + 1.39\}} \\
&= \sqrt{0.03} = 0.175
\end{aligned}$$

but the *df* must be approximated using the Satterthwaite approach

$$\hat{df} = \frac{0.175^4}{\frac{1}{120^2}\left(\frac{((4-1)0.76)^2}{6} + \frac{1.39^2}{2}\right)} = 7.33$$

with $df_{AB} = 6, df_B = 2$. Since $t(0.025, 7.33) = 2.34$, a 95% c.i. for the population mean of location 1, for example, is $10.9 \pm 2.34(0.175)$.

SAS code to fit two-factor random effects model for plant acid data:
Recall: Both effects are random and leaf is nested in plant (since leaf 1 from plant 2 doesn't really mean the same as leaf 1 from plant 2).

$$Y_{ijk} = \mu + A_i + B_{j(i)} + E_{ijk}$$

w/ variance components $\sigma^2, \sigma^2_A, \sigma^2_{B(A)}$.

```
proc mixed method=type3 cl;
class plant leaf;
model y=/cl;
random plant leaf(plant);
run;
```

Type 3 Analysis of Variance

Source	DF	Sum of Squares	Mean Square	Expected Mean Square	Error Term	Error DF	F Value	Pr > F
plant	3	343.178889	114.392963	Var(Residual) + 3 Var(leaf(plant)) + 9 Var(plant)	MS(leaf(plant))	8	4.88	0.0324
leaf(plant)	8	187.453333	23.431667	Var(Residual) + 3 Var(leaf(plant))	MS(Residual)	24	185.39	<.0001
Residual	24	3.033333	0.126389	Var(Residual)	.	.	.	.

Covariance Parameter Estimates

Cov Parm	Estimate	Alpha	Lower	Upper
plant	10.1068	0.05	-10.3930	30.6066
leaf(plant)	7.7684	0.05	0.1142	15.4227
Residual	0.1264	0.05	0.07706	0.2446

Fit Statistics

-2 Res Log Likelihood	92.7
AIC (smaller is better)	98.7
AICC (smaller is better)	99.5
BIC (smaller is better)	96.9

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Intercept	14.2611	1.7826	3	8.00	0.0041	0.05	8.5882	19.9341

Variance components estimated as

$$\begin{aligned}
\hat{\sigma}^2 &= MS[E] = 0.13 \\
\hat{\sigma}^2_{B(A)} &= \frac{MS[B(A)] - MS[E]}{n} \\
&= \frac{23.4 - 0.13}{3} = 7.8 \\
\hat{\sigma}^2_A &= \frac{MS[A] - MS[B(A)]}{nb} \\
&= \frac{114.4 - 23.4}{9} = 10.1
\end{aligned}$$

To test for random effect of nested factor B (leaf), $H_0 : \sigma^2_{B(A)} = 0$,

$$F = \frac{MS[B(A)]}{MS[E]} = \frac{23.4}{0.13} = 185.4$$

on $(b-1)a = 8$ and $(n-1)ab = 24$ df (p-value < 0.0001).

To test for random effect of factor A (plant), $H_0 : \sigma^2_A = 0$,

$$F = \frac{MS[A]}{MS[B(A)]} = \frac{114.4}{23.4} = 4.88$$

on $a - 1 = 3$ and $(b-1)a = 8df$ with $p = 0.0324$.

So there is evidence of both a random plant effect and a random leaf effect, nested in plant. The magnitudes of these effects are quantified by the estimated variance components. The statistical significance addressed by the p-values.

Implied correlation structure for plant acids:
What is the correlation of two observations taken from the same plant

- and the same leaf?
- and different leaves?

Recall that $Y_{ijk} = \mu + A_i + B_{j(i)} + E_{ijk}$.

$$\begin{aligned}
Corr(Y_{ijk_1}, Y_{ijk_2}) &= \frac{\text{Cov}(Y_{ijk_1}, Y_{ijk_2})}{\sigma^2 + \sigma^2_A + \sigma^2_{B(A)}} \\
&= \frac{\text{Cov}(A_i, A_i) + \text{Cov}(B_{j(i)}, B_{j(i)})}{\sigma^2 + \sigma^2_A + \sigma^2_{B(A)}} \\
&= \frac{\sigma^2_A + \sigma^2_{B(A)}}{\sigma^2 + \sigma^2_A + \sigma^2_{B(A)}} \\
Corr(Y_{ij_1k_1}, Y_{ij_2k_2}) &= \frac{\text{Cov}(Y_{ij_1k_1}, Y_{2j_2k_2})}{\sigma^2 + \sigma^2_A + \sigma^2_{B(A)}} \\
&= \frac{\text{Cov}(A_i, A_i)}{\sigma^2 + \sigma^2_A + \sigma^2_{B(A)}} \\
&= \frac{\sigma^2_A}{\sigma^2 + \sigma^2_A + \sigma^2_{B(A)}}
\end{aligned}$$

Estimates of these correlations are

- $\frac{10.1+7.8}{10.1+7.8+0.13} = \frac{17.9}{18.0} = 0.99$
- $\frac{10.1}{10.1+7.8+0.13} = \frac{10.1}{18.0} = 0.56$

This means that two measurements taken on the same leaf are almost perfectly correlated. Almost all the variation in any measurement can be explained by the leaf and plant effects.

Experiment with light treatments on seedlings:
Recall we have a fixed treatment effect with a nested random effect.

- Response (y) is seedling height,
- treatments are light sources, intensities,
- experimental units are 10 pots (points on graph).

Model to fit

$$Y_{ijk} = \mu + \alpha_i + P_{(i)j} + E_{ijk}$$

α_i - treatment effects for $i = 1, 2, 3, 4, 5$
$P_{(i)j}$ - pot effects, nested in treatments, $j = 1, 2$ for each i.
E_{ijk} - seedling/experimental errors, $k = 1, 2$

$$P_{(i)j} \overset{iid}{\sim} N(0, \sigma_P^2), \quad E_{ijk} \overset{iid}{\sim} N(0, \sigma^2) \ \ (P_{(i)j} \perp E_{ijk})$$

Treatment	Pot	Seedling 1	Seedling 2
1	1	32.94	35.98
1	2	34.76	32.40
2	1	30.55	32.64
2	2	32.37	32.04
3	1	31.23	31.09
3	2	30.62	30.42
4	1	34.41	34.88
4	2	34.07	33.87
5	1	35.61	35.00
5	2	33.65	32.91

SAS code for fitting this mixed model is given below:

```
proc mixed data=plantheight method=type3;
class treatment pot;
model height=treatment;
random pot(treatment);
lsmeans treatment/adjust=tukey cl;
run;
```

Type 3 Analysis of Variance								
Source	DF	Sum of Squares	Mean Square	Expected Mean Square	Error Term	Error DF	F Value	Pr > F
Treatment	4	41.080770	10.270192	Var(Residual) + 2 Var(Pot(Treatment)) + Q(Treatment)	MS(Pot(Treatment))	5	8.40	0.0192
Pot(Treatment)	5	6.112350	1.222470	Var(Residual) + 2 Var(Pot(Treatment))	MS(Residual)	10	1.19	0.3793
Residual	10	10.264200	1.026420	Var(Residual)	.	.	.	.

Covariance Parameter Estimates	
Cov Parm	Estimate
Pot(Treatment)	0.09802
Residual	1.0264

Fit Statistics	
-2 Res Log Likelihood	50.8
AIC (smaller is better)	54.8
AICC (smaller is better)	55.8
BIC (smaller is better)	55.4

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
Treatment	4	5	8.40	0.0192

Least Squares Means									
Effect	Treatment	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Treatment	1	34.0200	0.5528	5	61.54	<.0001	0.05	32.5989	35.4411
Treatment	2	31.9000	0.5528	5	57.70	<.0001	0.05	30.4789	33.3211
Treatment	3	30.8400	0.5528	5	55.79	<.0001	0.05	29.4189	32.2611
Treatment	4	34.3075	0.5528	5	62.06	<.0001	0.05	32.8864	35.7286
Treatment	5	34.2925	0.5528	5	62.03	<.0001	0.05	32.8714	35.7136

Differences of Least Squares Means														
Effect	Treatment	_Treatment	Estimate	Standard Error	DF	t Value	Pr > |t|	Adjustment	Adj P	Alpha	Lower	Upper	Adj Lower	Adj Upper
Treatment	1	2	2.1200	0.7818	5	2.71	0.0422	Tukey	0.1826	0.05	0.1103	4.1297	-1.0162	5.2562
Treatment	1	3	3.1800	0.7818	5	4.07	0.0097	Tukey	0.0475	0.05	1.1703	5.1897	0.04380	6.3162
Treatment	1	4	-0.2875	0.7818	5	-0.37	0.7281	Tukey	0.9948	0.05	-2.2972	1.7222	-3.4237	2.8487
Treatment	1	5	-0.2725	0.7818	5	-0.35	0.7416	Tukey	0.9958	0.05	-2.2822	1.7372	-3.4087	2.8637
Treatment	2	3	1.0600	0.7818	5	1.36	0.2332	Tukey	0.6753	0.05	-0.9497	3.0697	-2.0762	4.1962
Treatment	2	4	-2.4075	0.7818	5	-3.08	0.0275	Tukey	0.1248	0.05	-4.4172	-0.3978	-5.5437	0.7287
Treatment	2	5	-2.3925	0.7818	5	-3.06	0.0281	Tukey	0.1273	0.05	-4.4022	-0.3828	-5.5287	0.7437
Treatment	3	4	-3.4675	0.7818	5	-4.44	0.0068	Tukey	0.0340	0.05	-5.4772	-1.4578	-6.6037	-0.3313
Treatment	3	5	-3.4525	0.7818	5	-4.42	0.0069	Tukey	0.0346	0.05	-5.4622	-1.4428	-6.5887	-0.3163
Treatment	4	5	0.01500	0.7818	5	0.02	0.9854	Tukey	1.0000	0.05	-1.9947	2.0247	-3.1212	3.1512

For treatment effects, use $MS(Pot(treatments))$ as error term.
For example, for $H_0 : \alpha_1 = \alpha_2 = \cdots = 0$, use

$$F = \frac{MS(\text{treatment})}{MS(\text{Pot(treatment)})} \sim F_{5-1,5(2-1)} \text{ or } F_{4,5}$$

Be careful not to use

$$F = \frac{MS(\text{treatment})}{MS(E)}$$

For these data, we get

$$F = \frac{10.27}{1.22} = 8.4(df = 4, 5, p = .0192)$$

providing evidence of a treatment effect on plant heights.

Chapter 13

ST 512 - Block Designs

Readings: 15.1-15.5 700-732

Motivation - sometimes the variability of responses among experimental units is large, making detection of differences among treatment means $\mu_1, \mu_2, \ldots, \mu_t$ difficult.

Using a Block design can help in this situation!

Example: Suppose we are testing 2 drugs (A, B) on mice

- Response is mouse activity
- 12 mice - We use a Completely Randomized Design (CRD)
- Find that treatment A increases activity significantly more than treatment B
- Fellow scientist notes that activity level may change based on the weight of the mice
  - By random chance all of the treatment A mice are 'small' and all the treatment B mice are 'big'

Want to know if the drug was truly effective. How can we account for the weight of the mice? (ANCOVA is one option, but let's consider a different option).

Should we use only small mice? only large mice? Want to make inference to population as a whole. Block design can remedy this!

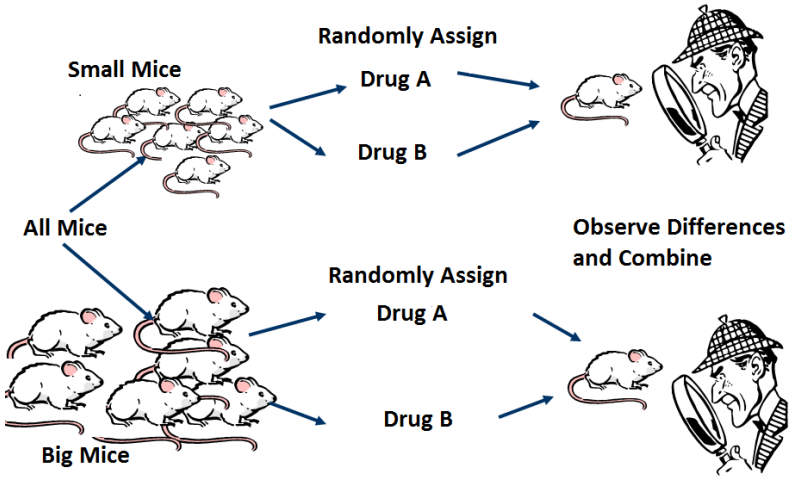
Recall:

- Blocking breaks the population up into subgroups (or blocks) **based on confounding variables that may have an effect on the response but are not of interest**
  - No need to block on variables that don't effect the response
- Experiment is run within each block then combined to make inference
- Permits inference to larger population while removing unwanted variability between units - allowing treatment effects to show up more clearly
- Technique reconciles two opposing aims of experimental design
  1. Want our subjects to be homogeneous so subtle treatment differences can be seen
  2. Usually aim to make inferences relevant to a population of interest
     - Our experimental units (subjects) must be representative of the population

The Randomized Complete Block Design (RCBD)

- Experimenter creates Blocks of subjects (Block) - Goal:
  - Try to make differences among blocks as large as possible
  - Try to make differences within blocks as small as possible
- Assume blocks are large enough to contain a complete replicate of the full set of treatments *at least once* (Complete)
- Within each block, randomization is used to assign treatments to subjects (Randomized)
- Hence, Randomized Complete Block Design (RCBD)

Example:



Randomized Complete Block Model (with fixed blocks, we assume our factor of interest is fixed)

$$Y_{ijk} = \mu + \beta_i + \tau_j + (\beta\tau)_{ij} + E_{ijk}$$

Randomized Complete Block Model (with random blocks, we assume our factor of interest is fixed)

$$Y_{ijk} = \mu + B_i + \tau_j + (B\tau)_{ij} + E_{ijk}$$

- μ overall mean
- $i = 1, ..., b$ (# blocks)
- β_i or B_i represent block effects (assumed iid $\sim N(0, \sigma_B^2)$, independent of other random effects if random blocks)
- $j = 1, ..., t$ (# treatments)
- τ_j represent treatment effects
- $(\beta\tau)_{ij}$ or $(B\tau)_{ij}$ represents the interaction between block and treatment (assumed iid $\sim N(0, \sigma_{B*Trt}^2)$, independent of other random effects if random blocks)
- $k = 1$ (# replications per block, assuming only 1)
- E_{ijk} represent random errors (assumed iid $\sim N(0, \sigma^2)$, independent of other random effects)

What might be an example of a fixed block? a random block?

- If blocks fixed - statistical inferences apply only to the blocks used
- If blocks random - statistical inferences about treatment effects apply to the entire population
- Either way, analysis for the RCBD is the same! As with CRD, can test for treatment effects by equating mean squares

To test H_0 : all $\tau_j = 0$, can look at expected mean squares to find the appropriate test

Analysis of Variance Table for RCBD

Source	df	Expected Mean Square Fixed Block	Expected Mean Square Random Block
Block	b-1	$nt\phi_B^2 + \sigma^2$	$nt\sigma_B^2 + n\sigma_{B*Trt}^2 + \sigma^2$
Treatment	t-1	$nb\phi_{Trt}^2 + \sigma^2$	$nb\phi_{Trt}^2 + n\sigma_{B*Trt}^2 + \sigma^2$
Block*Treatment	(b-1)(t-1)	$n\phi_{B*Trt}^2 + \sigma^2$	$n\sigma_{B*Trt}^2 + \sigma^2$
Error	0	σ^2	σ^2

- ϕ^2 terms are quadratic forms corresponding to the given fixed effect
- σ^2 terms represent random effects

Notice: Not possible to calculate a direct estimate of σ^2

Recall: σ^2 is the error variance, or the variance of the population of measurements made under identical experimental conditions.

In a RCB design, responses measured under identical experimental conditions are responses corresponding to a common block-treatment combination. Consequently, σ^2 is the variance of the conceptual population of responses that can be observed for a given treatment within a given block.

What to do with fixed block?

Analysis of Variance Table for RCBD

Source	df	Expected Mean Square Fixed Block	Expected Mean Square Random Block
Block	b-1	$nt\phi_B^2 + \sigma^2$	$nt\sigma_B^2 + n\sigma_{B*Trt}^2 + \sigma^2$
Treatment	t-1	$nb\phi_{Trt}^2 + \sigma^2$	$nb\phi_{Trt}^2 + n\sigma_{B*Trt}^2 + \sigma^2$
Block*Treatment	(b-1)(t-1)	$n\phi_{B*Trt}^2 + \sigma^2$	$n\sigma_{B*Trt}^2 + \sigma^2$
Error	0	σ^2	σ^2

- Fixed block, there is no ratio of mean squares to test the treatment
- If no Block*Treatment is reasonable ($\phi_{B*Trt}^2 = 0$) can use

$$F_{Trt} = \frac{MS[Trt]}{MS[B*Trt]}$$

What to do with random block?

Analysis of Variance Table for RCBD

Source	df	Expected Mean Square Fixed Block	Expected Mean Square Random Block
Block	b-1	$nt\phi_B^2 + \sigma^2$	$nt\sigma_B^2 + n\sigma_{B*Trt}^2 + \sigma^2$
Treatment	t-1	$nb\phi_{Trt}^2 + \sigma^2$	$nb\phi_{Trt}^2 + n\sigma_{B*Trt}^2 + \sigma^2$
Block*Treatment	(b-1)(t-1)	$n\phi_{B*Trt}^2 + \sigma^2$	$n\sigma_{B*Trt}^2 + \sigma^2$
Error	0	σ^2	σ^2

- For a random block, can use $F_{Trt} = \frac{MS[Trt]}{MS[B*Trt]}$ to test H_0 : all $\tau_j = 0$
- In both cases $MS[B*Trt]$ is used as error term

Recall: Interaction would imply treatment acts differently depending on block.

- If interaction truly exists:
  - Only one complete replicate in each block, power of tests will be diluted
  - If we have more than one complete replicate in each block, we can model interaction
- If interaction does not exist:
  - More advantageous to include extra blocks rather than extra replications within blocks

Advantage of block design - Differences between treatment means don't involve blocks
Assume we have a balanced design. Denote the j^{th} treatment mean by

$$\hat{\mu}_{+j} = \mu + \bar{\beta} + \tau_j$$

We may be interested in a quantity such as

$$\theta = \mu_{+2} - \mu_{+1} = \tau_2 - \tau_1$$

Estimated by

$$\hat{\theta} = \bar{y}_{+2} - \bar{y}_{+1}$$

Doesn't include the blocks!

Making inference for means or differences of means:
If blocks fixed:

- $\bar{y}_{.j}$ is interpreted as estimating $\mu + \bar{\beta} + \tau_j$
- Variance is σ^2/b
- Variance of $\hat{\theta} = \bar{y}_{+2} - \bar{y}_{+1}$ is $Var(\theta) = 2\sigma^2/b$
- σ^2 is estimated by $MS(Block * Trt)$

Thus, when blocks are fixed, inference is equivalent to inference in the two-way ANOVA model using the interaction term as the error.

Blocks random:

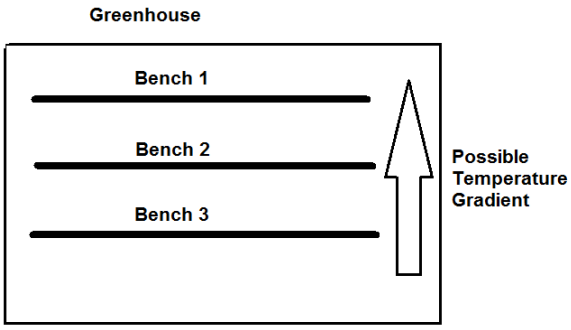
- $\bar{y}_{.j}$ is an unbiased estimate of $\mu + \tau_j$
- Variance is $(\sigma_B^2 + \sigma^2)/b$
- Degrees of freedom will need to be found using Satterthwaite's approximation!
- Variance of $\hat{\theta} = \bar{y}_{+2} - \bar{y}_{+1}$ is $Var(\theta) = 2\sigma^2/b$
- σ^2 is estimated by $MS(Block * Trt)$

Thus, when blocks are random, inference for differences of treatment means don't involve blocks!

We would of course want to make multiple comparison corrections for testing all pairwise differences of means.

Example of RCBD:

- Scientist investigating Yield of Tomato plants - Has 5 different types of fertilizers
- Plants placed in pots on 3 benches in the greenhouse
- 5 pots can fit on each bench and there is a possible temperature gradient



How can we set up our design to account for temperature gradient and reduce unexplained variation in response?

Blocks fixed or random here?

Assumptions about interaction?

Analysis can be done in proc mixed (if replication, simply add random 'Bench*Fertilizer' effect and use usual two-way mixed effects analysis):

```
proc mixed method=type3 plots=all;
class Bench Fertilizer;
model Yield=Fertilizer/residual;
random Bench;
lsmeans Fertilizer/adjust=Tukey;
run;
```

Source	DF	MS	Error Term	Error DF	P-value
Fertilizer	4	0.4164	MS(Residual)	8	0.0017
Bench	2	0.0042	MS(Residual)	8	0.8828
Residual	8	0.0338	.	.	.

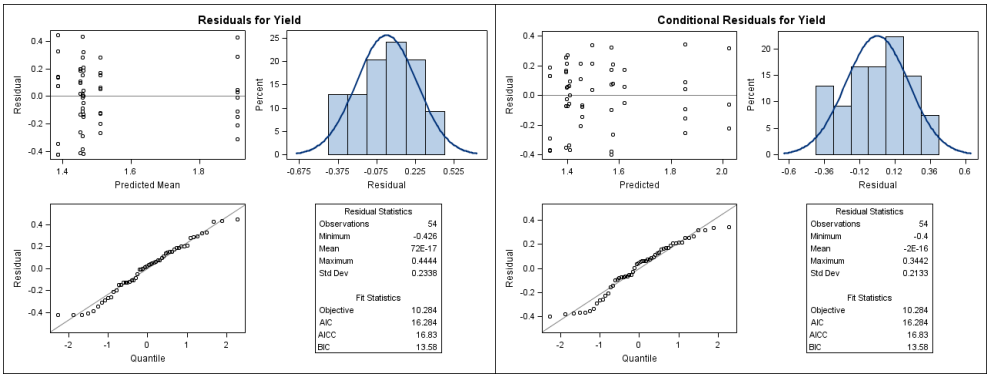
We can see we have a significant Fertilizer effect. We don't care about the block effect significance. Even if not significant we would leave this effect in!

Remember, Residual here is really Fertilizer*Bench interaction.

Differences of LSMeans

Fert	Fert	Estimate	Adj P
1	2	0.4133	0.1306
1	3	0.4267	0.1158
1	4	0.5700	0.0316
1	5	1.0367	0.0008
2	3	0.01333	1.0000
2	4	0.1567	0.8293
2	5	0.6233	0.0198
3	4	0.1433	0.8679
3	5	0.6100	0.0222
4	5	0.4667	0.0805

Should always check model assumptions:



Relationship with ANCOVA - When to use the covariate with ANCOVA and when to use it to create blocks?

- If value not known until experiment underway or over (i.e. need to sacrifice animal to know) use ANCOVA
- If covariate nearly constant for groups use Blocking
- If not clear, guide based on correlation of covariate and response:
  - if correlation is less than 0.3 ignore
  - if correlation is between 0.3 and 0.6 use blocking
  - if correlation is greater than 0.6 use ANCOVA
- Note: ANCOVA not very robust against failures of the ANOVA normality assumption - model must be correct

More Types of Block Designs

- Two blocking factors - Latin squares
- For ordinal responses - Friedman Rank sum test
- For binary response - Cochran's test
- Restrictions on blocks - Incomplete block designs

Summary

- Randomized Complete Block Design is relatively simple, but powerful technique used to eliminate the effects of selected confounding variables when comparing treatments
- Allows for easier detection of treatment effects while having results apply to larger population
- Standard ANOVA assumptions plus Treatment and Block **do not** interact
- Very simple analysis and interpretation if balanced design

Chapter 14

ST 512 - Repeated Measure and Split Plot Designs

Readings: 16.1-16742-779

Motivating Example:
Experiment investigating pesticide (3 levels) on yield of corn.

- 6 plots of land used
- CRD done (each of 3 pesticides randomly assigned to 2 plots)

plot	pest	y
1	1	53.4
2	1	46.5
1	2	54.3
2	2	57.2
1	3	55.9
2	3	57.4

Analysis? One-way ANOVA model:

$$Y_{ij} = \mu + \alpha_i + E_{ij}$$

- μ is overall mean
- α_i is effect of pesticide i
- E_{ij} random effect of plot (iid $N(0, \sigma^2)$)

Test for pesticide is

$$F = MS(Trt)/MS(E) \text{ i.e. } F = MS(Pesticide)/MS(Plot)$$

Split-plot experiment
Consider the same experiment except, after you've assigned the pesticide, you break each plot into 4 'subplots.' You randomly assign the 4 levels of a second factor, irrigation, to the subplots.

pest	plot	Yield at Irr 1	Yield at Irr 2	Yield at Irr 3	Yield at Irr 4
1	1	53.4	53.8	58.2	59.5
1	2	46.5	51.1	49.2	51.3
2	1	54.3	56.3	60.4	64.5
2	2	57.2	56.9	61.6	66.8
3	1	55.9	58.6	62.4	64.5
3	2	57.4	60.2	57.2	62.7

Total number of data points?

Is usual two-way ANOVA model appropriate here? Why or why not?

No! Errors in two-way model are assumed independent. Here, observations on the same plot are probably more alike than observations on separate plots.

Completely Randomized Split Plot model

$$Y_{ijk} = \mu + \alpha_i + S_{(i)j} + \beta_j + (\alpha\beta)_{ij} + E_{ijk}$$

- $i = 1, 2, 3$ pesticides (generally, $i = 1, ..., a$)
- $j = 1, 2, 3, 4$ irrigation (generally, $j = 1, ..., b$)
- $k = 1, 2$ plots (generally $k = 1, ..., r_i$)
- $S_{(i)j} \sim^{iid} N(0, \sigma_S^2)$ 'whole plot error' = random effect for plot j with pesticide i
  Nested as plots are different for each level of pesticide.
- $E_{ijk} \sim^{iid} N(0, \sigma^2)$ 'subplot error' = random effect between subplots (independent of S)

Definitions

- Experimental Unit - Unit on which a factor has its levels assigned
- A - **Whole plot factor (WPF)** (also called a *between plots* or *between subjects* factor)
- **Whole Plots (WP)** - E.U.'s for WPF
- B - **Subplot factor (SPF)** (also called a *within plots* or *within subjects* factor)
- **Subplots (SP)** - E.U.'s for SPF

In terms of a repeated measures study (over time) on 'subjects'. Whole plots = 'subjects', time = Subplot factor.

Suggestion: draw a picture of the layout when possible!

This is a mixed effects model! How to make inference? Look at expected mean squares!

$$Y_{ijk} = \underbrace{\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}}_{\mu_{ij}:(\text{fixed component})} + \overbrace{S_{k(i)} + E_{ijk}}^{\text{random error component}}.$$

Here, $i = 1, \ldots, a$ and $j = 1, \ldots, b$ and $k = 1, \ldots, r_i$ where r_i denotes the number of plots treated with level i of factor a. If r_i is constant, call it r.

Source	df	EMS
A : Pesticide	$a - 1 = 2$	$\sigma^2 + b\sigma_s^2 + br\psi_A^2$
Plot(A)	$(r-1)a = (2-1)3 = 3$	$\sigma^2 + b\sigma_s^2$
B: treatments	$b - 1 = 3$	$\sigma^2 + ra\psi_B^2$
$A \times B$	$(a-1)(b-1) = 6$	$\sigma^2 + r\psi_{AB}^2$
$B\times$plot(A)	$(b-1)(r-1)a = 9$	σ^2
Subplot error		
Total	$abr - 1 = 23$	

Test for the Whole plot factor is

$$F = MS(A)/MS(Plot(A)) \text{ equivalent to one way ANOVA test!}$$

Test for the sub plot factor is

$$F = MS(B)/MS(E)$$

and test for interaction is

$$F = MS(AB)/MS(E)$$

The variance of an observation is

$$Var(Y_{ijk}) = \sigma_S^2 + \sigma^2$$

Covariance between two observations on same plot is

$$Cov(Y_{ij_1k}, Y_{ij_2k}) = \sigma_S^2$$

Model allows for the correlation of observations on same plot!

$$Corr(Y_{ij_1k}, Y_{ij_2k}) = \frac{\sigma_S^2}{\sigma^2 + \sigma_S^2}$$

For the corn yields data,

Source	MS	df	EMS	F	$p-$value
A : Pesticide	128.1	2	$\sigma^2 + b\sigma_s^2 + br\psi_A^2$	3.9	0.1452
Whole plot error	$MS[S(A)] = 32.6$	3	$\sigma^2 + b\sigma_s^2$	10.1	0.0031
B: treatments	60.2	3	$\sigma^2 + ra\psi_B^2$	18.7	0.0003
$A \times B$	4.1	6	$\sigma^2 + r\psi_{AB}^2$	1.3	0.3607
$B\times$plot(A) (Subplot error)	$MS[E] = 3.2$	9	σ^2		
Total		23			

- $MS[S(A)]$ denotes mean square for WHOLE plots (nested in A)
- $MS[E]$ denotes error or subplot mean square

Analysis precedes just as multiway ANOVA -

- Check for interaction significance - if significant, look at simple effects
- If no interaction, check for main effect significance - if significant, investigate main effects

For pesticide by irrigation interaction, on $6, 9$ df:

$$F = MS[AB]/MS[E] = 4.1/3.2$$

For pesticide effect, on $2, 3$ df:

$$F = MS[A]/MS[S(A)] = 128.1/32.6$$

For irrigation effect, on $3, 9$ df:

$$F = MS[B]/MS[E] = 60.2/3.2$$

For random effect of whole plots could do a test as well, on $3, 9$ df:

$$F = MS[S(A)]/MS[E] = 32.6/3.2$$

Estimated varcomps:

$$\hat{\sigma}^2 = MS[E] = 3.2 \text{ and } \hat{\sigma_s}^2 = (MS[S(A)] - MS[E])/4 = 7.3$$

Pairwise comparisons

Several kinds of pairwise comparisons of treatment means:

1. Main effects of A: $\bar{y}_{i_1++} - \bar{y}_{i_2++}$
2. Main effects of B: $\bar{y}_{+j_1+} - \bar{y}_{+j_2+}$
3. Simple effects of A: $\bar{y}_{i_1j+} - \bar{y}_{i_2j+}$
4. Simple effects of B: $\bar{y}_{ij_1+} - \bar{y}_{ij_2+}$
5. Interaction effects: $\bar{y}_{i_1j_1+} - \bar{y}_{i_2j_2+}$

Skipping the algebra, the standard errors for all of these comparisons, save # 3 and #5, can be estimated 'cleanly.' That is, with single MS terms and integer df. (See table 16.6, careful of errata)

Comparison	Variance	Estimate	df
$\bar{Y}_{i_1++} - \bar{Y}_{i_2++}$	$\frac{2}{rb}(\sigma^2 + b\sigma_s^2)$	$\frac{2}{rb}MS[S(A)]$	$(r-1)a$
$\bar{Y}_{+j_1+} - \bar{Y}_{+j_2+}$	$\frac{2}{ra}\sigma^2$	$\frac{2}{ra}MS[E]$	$(r-1)(b-1)a$
$\bar{Y}_{i_1j+} - \bar{Y}_{i_2j+}$	$\frac{2}{r}(\sigma^2 + \sigma_s^2)$	$\frac{2}{r}(\hat{\sigma}^2 + \hat{\sigma}_s^2)$	messy
$\bar{Y}_{ij_1+} - \bar{Y}_{ij_2+}$	$\frac{2}{r}\sigma^2$	$\frac{2}{r}MS[E]$	$(r-1)(b-1)a$
$\bar{Y}_{i_1j_1+} - \bar{Y}_{i_2j_2+}$	$\frac{2}{r}(\sigma^2 + \sigma_s^2)$	$\frac{2}{r}(\hat{\sigma}^2 + \hat{\sigma}_s^2)$	messy

To analyze data from a CRSPD in SAS, PROC MIXED can be used:

```
proc mixed data=cornsp method=type3;
class pest plot irr;
model yield = pest|irr/ddfm=satterthwaite;
random plot(pest);
lsmeans trt pest/adjust=tukey cl;
*lsmeans trt|pest/adjust=tukey cl;  /* if there were interaction */
run;
```

Type 3 Analysis of Variance								
Source	DF	Sum of Squares	Mean Square	Expected Mean Square	Error Term	Error DF	F Value	Pr > F
pest	2	256.275833	128.137917	Var(Residual) + 4 Var(plot(pest)) + Q(pest,pest*irr)	MS(plot(pest))	3	3.93	0.1452
irr	3	180.697917	60.232639	Var(Residual) + Q(irr,pest*irr)	MS(Residual)	9	18.66	0.0003
pest*irr	6	24.490833	4.081806	Var(Residual) + Q(pest*irr)	MS(Residual)	9	1.26	0.3607
plot(pest)	3	97.806250	32.602083	Var(Residual) + 4 Var(plot(pest))	MS(Residual)	9	10.10	0.0031
Residual	9	29.058750	3.228750	Var(Residual)	.	.	.	.

Covariance Parameter Estimates	
Cov Parm	Estimate
plot(pest)	7.3433
Residual	3.2287

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
pest	2	3	3.93	0.1452
irr	3	9	18.66	0.0003
pest*irr	6	9	1.26	0.3607

Least Squares Means									
Effect	irr	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
irr	1	54.1167	1.3274	4.9	40.77	<.0001	0.05	50.6841	57.5492
irr	2	56.1500	1.3274	4.9	42.30	<.0001	0.05	52.7174	59.5826
irr	3	58.1667	1.3274	4.9	43.82	<.0001	0.05	54.7341	61.5992
irr	4	61.5500	1.3274	4.9	46.37	<.0001	0.05	58.1174	64.9826

Differences of Least Squares Means														
Effect	irr	_irr	Estimate	Standard Error	DF	t Value	Pr > |t|	Adjustment	Adj P	Alpha	Lower	Upper	Adj Lower	Adj Upper
irr	1	2	-2.0333	1.0374	9	-1.96	0.0816	Tukey-Kramer	0.2708	0.05	-4.3802	0.3135	-5.2720	1.2053
irr	1	3	-4.0500	1.0374	9	-3.90	0.0036	Tukey-Kramer	0.0156	0.05	-6.3968	-1.7032	-7.2886	-0.8114
irr	1	4	-7.4333	1.0374	9	-7.17	<.0001	Tukey-Kramer	0.0003	0.05	-9.7802	-5.0865	-10.6720	-4.1947
irr	2	3	-2.0167	1.0374	9	-1.94	0.0838	Tukey-Kramer	0.2766	0.05	-4.3635	0.3302	-5.2553	1.2220
irr	2	4	-5.4000	1.0374	9	-5.21	0.0006	Tukey-Kramer	0.0026	0.05	-7.7468	-3.0532	-8.6386	-2.1614
irr	3	4	-3.3833	1.0374	9	-3.26	0.0098	Tukey-Kramer	0.0405	0.05	-5.7302	-1.0365	-6.6220	-0.1447

Model is very flexible. Suppose that the irrigation factor was actually a combination of two factors:

The factor B is really a 2×2 combination of irrigation and cultivar:

B	Irr	Cult
1	no	1
2	no	2
3	yes	1
4	yes	2

The 3 *df* for the within plot factor B can be broken up into three 1 df components due to main effect of irr, main effect of Cult and interaction. Same with the AB interaction.

Model is

$$Y_{ijkl} = \mu + \alpha_i + S_{(i)k} + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + E_{ijkl}$$

(Could also have multiple whole plot factors as well.)

Easily coded up in SAS:

```
proc mixed data=cornsp method=type3;
class pest plot irr2 cult;
model yield = pest|irr2|cult/ddfm=satterthwaite;
random plot(pest);
lsmeans irr2 cult/adjust=tukey cl;
run;
```

Type 3 Analysis of Variance								
Source	DF	Sum of Squares	Mean Square	Expected Mean Square	Error Term	Error DF	F Value	Pr > F
pest	2	256.275833	128.137917	Var(Residual) + 4 Var(plot(pest)) + Q(pest,pest*irr2,pest*cult,pest*irr2*cult)	MS(plot(pest))	3	3.93	0.1452
irr2	1	133.953750	133.953750	Var(Residual) + Q(irr2,pest*irr2,irr2*cult,pest*irr2*cult)	MS(Residual)	9	41.49	0.0001
pest*irr2	2	17.747500	8.873750	Var(Residual) + Q(pest*irr2,pest*irr2*cult)	MS(Residual)	9	2.75	0.1171
cult	1	44.010417	44.010417	Var(Residual) + Q(cult,pest*cult,irr2*cult,pest*irr2*cult)	MS(Residual)	9	13.63	0.0050
pest*cult	2	1.385833	0.692917	Var(Residual) + Q(pest*cult,pest*irr2*cult)	MS(Residual)	9	0.21	0.8109
irr2*cult	1	2.733750	2.733750	Var(Residual) + Q(irr2*cult,pest*irr2*cult)	MS(Residual)	9	0.85	0.3815
pest*irr2*cult	2	5.357500	2.678750	Var(Residual) + Q(pest*irr2*cult)	MS(Residual)	9	0.83	0.4670
plot(pest)	3	97.806250	32.602083	Var(Residual) + 4 Var(plot(pest))	MS(Residual)	9	10.10	0.0031
Residual	9	29.058750	3.228750	Var(Residual)	.	.	.	.

Covariance Parameter Estimates	
Cov Parm	Estimate
plot(pest)	7.3433
Residual	3.2287

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
pest	2	3	3.93	0.1452
irr2	1	9	41.49	0.0001
pest*irr2	2	9	2.75	0.1171
cult	1	9	13.63	0.0050
pest*cult	2	9	0.21	0.8109
irr2*cult	1	9	0.85	0.3815
pest*irr2*cult	2	9	0.83	0.4670

Least Squares Means										
Effect	irr2	cult	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
irr2	no		55.1333	1.2219	3.61	45.12	<.0001	0.05	51.5922	58.6745
irr2	yes		59.8583	1.2219	3.61	48.99	<.0001	0.05	56.3172	63.3995
cult		1	56.1417	1.2219	3.61	45.95	<.0001	0.05	52.6005	59.6828
cult		2	58.8500	1.2219	3.61	48.16	<.0001	0.05	55.3088	62.3912

Differences of Least Squares Means															
Effect	irr2	cult	_irr2	_cult	Estimate	Standard Error	DF	t Value	Pr > |t|	Adjustment	Adj P	Alpha	Lower	Upper	Adj Lower	Adj Upper
irr2	no		yes		-4.7250	0.7336	9	-6.44	0.0001	Tukey-Kramer	0.0001	0.05	-6.3845	-3.0655	-6.3844	-3.0656
cult		1		2	-2.7083	0.7336	9	-3.69	0.0050	Tukey-Kramer	0.0050	0.05	-4.3678	-1.0489	-4.3678	-1.0489

Split-plot in blocks (RCBSPD):
A RCBSPD is a design where the whole plot part of the experiment is in a RCBD (block usually random). That is, the whole plot factor randomization is done in a block manner.

Again consider the previous experiment. Now suppose the six plots come from two farms, with three plots in each farm. Suppose that the three pesticide treatments are randomized to plots within farms.

Renumbering plots (1,2,1,2,1,2) as (1,2,3,4,5,6) and supposing plots (2,3,6) come from farm 1 and plots (1,4,5) from farm 2, the data are given as

farm	pest	plot	Yield-IrrN,Cult1	Yield-IrrN,Cult2	Yield-IrrY,Cult1	Yield-IrrY,Cult2
2	1	1	53.4	53.8	58.2	59.5
1	1	2	46.5	51.1	49.2	51.3
1	2	3	54.3	56.3	60.4	64.5
2	2	4	57.2	56.9	61.6	66.8
2	3	5	55.9	58.6	62.4	64.5
1	3	6	57.4	60.2	57.2	62.7

At the whole plot level (ignoring the split-plot factor), the *df* for the Block Design are given by

Source	*df*
A : Pesticide	(3-1)=2
Farms i.e. Block	(2-1)=1
Error i.e. Block*Pesticide	(3-1)(2-1)=2
Total	5

so that an F-ratio for the pesticide effect is based on $df = 2, 2$. (As with the RCBD, the Farm*Pesticide interaction is used as error here.)

In general, for a RCBSPD with a levels of a whole-plot level (A) randomized to r blocks (for a total of ra plots) and b levels of a split-plot factor (B) within each plot, the model and ANOVA table are given by

$$\begin{aligned} Y_{ijk} &= \mu + \alpha_i + R_k + \beta_j + (\alpha\beta)_{ij} + (SR)_{ik} + +E_{ijk} \\ &= \mu_{ij} + R_k + (SR)_{ik} + +E_{ijk} \end{aligned}$$

where

- i denotes level of A,
- j denotes level of B,
- k denotes block.

$R_k \overset{iid}{\sim} N(0, \sigma_r^2)$ and $SR_{ik} \overset{iid}{\sim} N(0, \sigma_{sr}^2)$. All random errors are mutually independent.

Source	df	EMS
A	$a-1$	$\sigma^2 + b\sigma_{sr}^2 + br\psi_A^2$
Blocks	$r-1$	$\sigma^2 + b\sigma_{sr}^2 + ab\sigma_r^2$
Whole plot error (Block$\times A$)	$(r-1)(a-1)$	$\sigma^2 + b\sigma_{sr}^2$
B	$b-1$	$\sigma^2 + ar\psi_B^2$
AB	$(a-1)(b-1)$	$\sigma^2 + r\psi_{AB}^2$
Error ($B\times$Blocks(A))	$a(b-1)(r-1)$	σ^2
Total	$abr-1$	

Using this table of expected mean squares, what is the test used for A? test used for AB?

RCBSPD Analysis in SAS:

```
proc mixed data=cornsp method=type3;
class pest farm irr2 cult;
model yield = pest|irr2|cult;
random farm farm*pest;
run;
```

Type 3 Analysis of Variance

Source	DF	Sum of Squares	Mean Square	Expected Mean Square	Error Term	Error DF	F Value	Pr > F
pest	2	256.275833	128.137917	Var(Residual) + 4 Var(pest*farm) + Q(pest,pest*irr2,pest*cult,pest*irr2*cult)	MS(pest*farm)	2	6.64	0.1309
irr2	1	133.953750	133.953750	Var(Residual) + Q(irr2,pest*irr2,irr2*cult,pest*irr2*cult)	MS(Residual)	9	41.49	0.0001
pest*irr2	2	17.747500	8.873750	Var(Residual) + Q(pest*irr2,pest*irr2*cult)	MS(Residual)	9	2.75	0.1171
cult	1	44.010417	44.010417	Var(Residual) + Q(cult,pest*cult,irr2*cult,pest*irr2*cult)	MS(Residual)	9	13.63	0.0050
pest*cult	2	1.385833	0.692917	Var(Residual) + Q(pest*cult,pest*irr2*cult)	MS(Residual)	9	0.21	0.8109
irr2*cult	1	2.733750	2.733750	Var(Residual) + Q(irr2*cult,pest*irr2*cult)	MS(Residual)	9	0.85	0.3815
pest*irr2*cult	2	5.357500	2.678750	Var(Residual) + Q(pest*irr2*cult)	MS(Residual)	9	0.83	0.4670
farm	1	59.220417	59.220417	Var(Residual) + 4 Var(pest*farm) + 12 Var(farm)	MS(pest*farm)	2	3.07	0.2219
pest*farm	2	38.585833	19.292917	Var(Residual) + 4 Var(pest*farm)	MS(Residual)	9	5.98	0.0223
Residual	9	29.058750	3.228750	Var(Residual)	.	.	.	.

Covariance Parameter Estimates

Cov Parm	Estimate
farm	3.3273
pest*farm	4.0160
Residual	3.2287

Note: In practice, you should probably use the default method of estimation in proc mixed called REML for most mixed models.