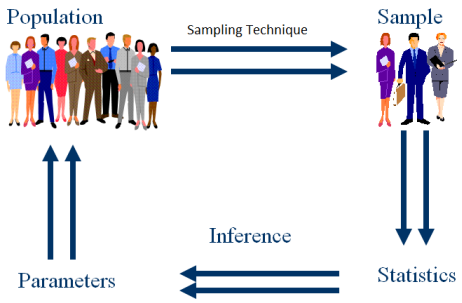
Contents

Chapter 1

ST 512 - Review

Readings: Chapters 1-8 as needed

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



Examples of parameters - (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 claims about the population 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)

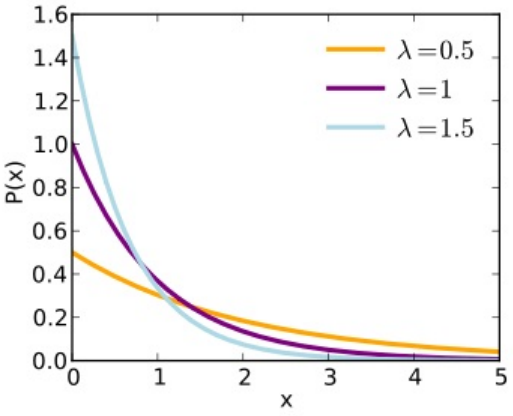
- Quantitative - A variable that is described by numerical measurements where arithmetic can be performed
  Subscales:
  Discrete - finite or countably infinite # of values (# of flowers, 0, 1, 2, ...)
  Continuous - any value in an interval is possible (Temperature, $(-459.67\,\text{deg}\,F, \infty)$)

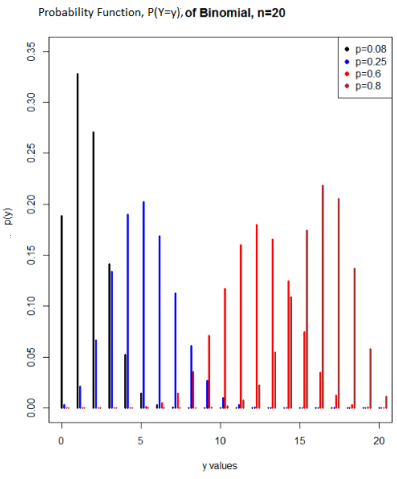
Random Variables and Things of Interest:

- Random Variable (RV) - Numeric outcome to a random process

  Things of interest

  - Distribution - pattern and frequency of observable values
    For continuous RVs, a smooth curve. For discrete a 'probability histogram.'





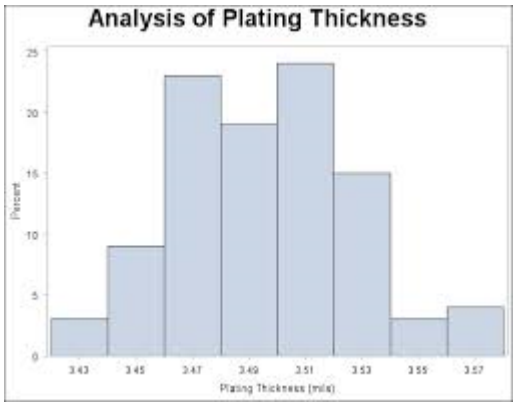
  - Mean/Median - measures of center of the distribution
    Main focus often on mean: true mean μ, RV sample mean $\bar{Y}$, observed sample mean $\bar{y}$

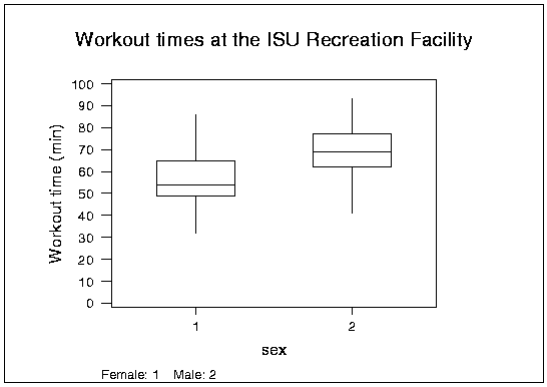
– Standard Deviation, Variance, IQR, Range - measures of spread for the distribution For our purposes, SD and Variance are most important: true variance σ^2, true SD σ, observed sample variance s^2, observed SD s

Graphical Descriptions of RV's:

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

- Boxplot - Uses the Five Number Summary (min, Q_1, M, Q_3, max) to display the realizations of a RV





Statistics are RVs!

The distribution of a statistic is called a sampling distribution

We almost always require that our sample is a **random sample** (RS), or equivalently that our random variables are **iid** (independent and identically distributed).

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

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

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

Two general methods for inference:

1. Confidence Interval (CI) - range of values we believe contain the parameter with some level of confidence

   For an observed $(1-\alpha)100\%$ confidence interval (c_L, c_U) we can say

   We are $(1-\alpha)100\%$ 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 $(1-\alpha)100\%$ probability of producing an interval that contains the parameter.

   i.e. If the experiment were done repeatedly and an interval made for each sample, $(1-\alpha)100\%$ of the intervals would contain the parameter value.

2. Hypothesis Test (HT) - test to determine if a given value is reasonable for a parameter

   For a hypothesis test, the p-value means

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

   Statistical significance implies

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

Common ways make inference about μ (true mean):

One sample Z-test

When the true SD, σ, is known and $\bar{Y}$ has a normal distribution the sampling distribution of the statistic

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

100(1-α)%Ci for μ is

$$\bar{Y} \pm z_{\alpha/2}\sigma/\sqrt{n}$$

HT: for $H_0 : \mu = \mu_0 \;\; vs \;\; H_A : \mu > \mu_0 \;\; or \;\; \mu < \mu_0 \;\; or \;\; \mu \neq \mu_0$

$$\text{Test Statistic: } Z = \frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}}$$

$$RR : \{z_{obs} : z_{obs} > z_\alpha\} \;\; or \;\; \{z_{obs} : z_{obs} < -z_\alpha\} \;\; or \;\; \left\{z_{obs} : |z_{obs}| > z_{\alpha/2}\right\}$$
$$P - value : P(Z > z_{obs}) \;\; or \;\; P(Z < z_{obs}) \;\; or \;\; 2 * P(Z > |z_{obs}|)$$

One sample T-test

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

$$T = \frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1} \;\; \text{(valid if RS and Y is normal)}$$

100(1-α)% CI for μ is

$$\bar{Y} \pm t_{\alpha/2,n-1}S/\sqrt{n}$$

HT: for $H_0 : \mu = \mu_0 \;\; vs \;\; H_a : \mu > \mu_0 \;\; or \;\; \mu < \mu_0 \;\; or \;\; \mu \neq \mu_0$

$$\text{Test Statistic: } T = \frac{\bar{Y} - \mu_0}{S/\sqrt{n}}$$

$$RR : \{t_{obs} : t_{obs} > t_{\alpha,n-1}\} \;\; or \;\; \{t_{obs} : t_{obs} < -t_{\alpha,n-1}\} \;\; or \;\; \left\{t_{obs} : |t_{obs}| > t_{\alpha/2,n-1}\right\}$$
$$P - value : P(T_{n-1} > t_{obs}) \;\; or \;\; P(T_{n-1} < t_{obs}) \;\; or \;\; 2 * P(T_{n-1} > |t_{obs}|)$$

Inference about two (true) means, μ_1 and μ_2 or $\mu_d = \mu_1 - \mu_2$:

Paired Data: (Paired t-test) Assume **differences** are a RS and normally distributed
100(1-α)% CI for μ_d is

$$\overline{D} \pm t_{\alpha/2,n-1} S_D/\sqrt{n} = \bar{Y} - \bar{X} \pm t_{\alpha/2,n-1} S_{\bar{Y}-\bar{X}}/\sqrt{n}$$

HT: for $H_0 : \mu_d = \Delta_0$ *vs* $H_a : \mu_d > \Delta_0$ *or* $\mu_d < \Delta_0$ *or* $\mu_d \neq \Delta_0$

$$\text{Test Statistic: } T = \frac{\bar{Y} - \bar{X} - \Delta_0}{S_d/\sqrt{n}}$$

$$RR : \{t_{obs} : t_{obs} > t_{\alpha,n-1}\} \quad or \quad \{t_{obs} : t_{obs} < -t_{\alpha,n-1}\} \quad or \quad \left\{t_{obs} : |t_{obs}| > t_{\alpha/2,n-1}\right\}$$
$$P - value : P(T_{n-1} > t_{obs}) \quad or \quad P(T_{n-1} < t_{obs}) \quad or \quad 2 * P(T_{n-1} > |t_{obs}|)$$

Independent Samples: Assume populations are independent RS's with each population having a normal distribution
Equal Variance (Two-sample pooled t-test):
100(1-α)% CI for μ_d is

$$\bar{Y} - \bar{X} \pm t_{\alpha/2,n_1+n_2-2} \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

HT: for $H_0 : \mu_d = \Delta_0$ *vs* $H_a : \mu_d > \Delta_0$ *or* $\mu_d < \Delta_0$ *or* $\mu_d \neq \Delta_0$

$$\text{Test Statistic: } T = \frac{\bar{Y} - \bar{X} - \Delta_0}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$RR : \{t_{obs} : t_{obs} > t_{\alpha,n_1+n_2-2}\} \quad or \quad \{t_{obs} : t_{obs} < -t_{\alpha,n_1+n_2-2}\} \quad or \quad \left\{t_{obs} : |t_{obs}| > t_{\alpha/2,n_1+n_2-2}\right\}$$
$$P - value : P(T_{n_1+n_2-2} > t_{obs}) \quad or \quad P(T_{n_1+n_2-2} < t_{obs}) \quad or \quad 2 * P(T_{n_1+n_2-2} > |t_{obs}|)$$

Unequal Variance (Two-sample t-test)
100(1-α)% CI for μ_d is

$$\bar{Y} - \bar{X} \pm t_{\alpha/2,\widehat{df}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

HT: for $H_0 : \mu_d = \Delta_0$ *vs* $H_a : \mu_d > \Delta_0$ *or* $\mu_d < \Delta_0$ *or* $\mu_d \neq \Delta_0$

$$\text{Test Statistic: } T = \frac{\bar{Y} - \bar{X} - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$RR : \left\{t_{obs} : t_{obs} > t_{\alpha,\widehat{df}}\right\} \quad or \quad \left\{t_{obs} : t_{obs} < -t_{\alpha,\widehat{df}}\right\} \quad or \quad \left\{t_{obs} : |t_{obs}| > t_{\alpha/2,\widehat{df}}\right\}$$
$$P - value : P(T_{\widehat{df}} > t_{obs}) \quad or \quad P(T_{\widehat{df}} < t_{obs}) \quad or \quad 2 * P(T_{\widehat{df}} > |t_{obs}|)$$

$$\widehat{df} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2/(n_1-1) + \left(\frac{s_2^2}{n_2}\right)^2/(n_2-1)}$$

Extension of two-sample pooled t-test to inference about t (true) means, $\mu_1, \mu_2, ..., \mu_t$:

Analysis used for a completely randomized design.

One Way ANOVA model:

$$Y_{ij} = \mu_i + E_{ij}$$

where $E_{ij} \sim^{iid} N(0, \sigma^2)$, $i = 1, 2, ...t$, and $j = 1, 2, ..., n$ (total sample size $= nt = N$)

μ_i is the mean for group i and σ^2 is the common variance for each population.

An alternative model:

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

where μ is the overall (grand) mean, τ_i is the effect for the i^{th} treatment, and $E_{ij} \sim^{iid} N(0, \sigma^2)$
Data and labeling:

Corn Syrup	Replicate #	'L' measurement	Label
26	1	51.89	y_{11}
26	2	51.52	y_{12}
26	3	52.69	y_{13}
42	1	47.21	y_{21}
42	2	48.57	y_{22}
42	3	47.57	y_{23}
55	1	41.43	y_{31}
55	2	42.31	y_{32}
55	3	42.31	y_{33}

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

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

P-value from ANOVA table tests

$$H_0 : \mu_1 = \mu_2 = ... = \mu_t \text{ vs } H_A : \text{at least 1 mean differs}$$

Chapter 2

ST 512 - Contrasts and Multiple Comparisons

Readings 9.1-9.5

For the next section or two we will continue to consider experiments that have a quantitative response and categorical predictors only (i.e. experiments that can be analyzed using One-Way ANOVA). To start with let's just look at a completely randomized experimental design (i.e. all treatments are randomly assigned to the experimental units). We will look at contrasts of the parameters that will lead directly into the way to analyze Multi-Way ANOVA models where a 'factorial' treatment structure is used.

Consider the traditional balanced One-Way ANOVA model (**ANOVA (Analysis of Variance, i.e. comparing mean squares)**). That is,

- we have a *continuous* response, Y
- *qualitative or categorical* predictor(s) which we call our **factor(s)** (often we will use a continuous factor that is only observed at a few values as our categorical factor)
- if a single factor then the **levels** of the factor are our **treatments**, if multiple factors the combinations of levels from the factors is the treatment. (Either way, t = total number of treatments)

(One-Way corresponds to having only one factor of interest.)

One form of the One-Way ANOVA model is

$$Y_{ij} = \mu_i + E_{ij}$$

- E_{ij} are iid $N(0, \sigma^2)$
- $i = 1, ..., t$ describes the treatment group
- $j = 1, ..., n_i$ represents the number of replications 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_i - (true) mean response for all members of population i
  Estimate:
  $$\bar{Y}_{i\bullet} = \frac{\sum_{j=1}^{n} Y_{ij}}{n}$$
  Average of all replicates for that treatment.

- σ^2 - (true) variance within a given treatment group (assumed constant across groups)
  Estimate:
  $$MS(E) = SS(E)/(N-t) = \frac{\sum_{i=1}^{t}\sum_{j=1}^{n}\left(Y_{ij} - \bar{Y}_{i\bullet}\right)^2}{N-t}$$
  Simplifies to $\frac{S_1^2+...+S_t^2}{t}$ in balanced designs. Essentially an 'avg' of each groups variance.

Goals of One-Way ANOVA: Determine

1. if all treatment group means are equal.

2. if treatment group means are not equal, which means differ from each other.

One-Way ANOVA example:
An experiment was done to determine if there was a difference between antibiotic types in terms of their mean 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
$y_{11} = 29.2$	Chloramphenicol	μ_1	$\bar{y}_{1\bullet} = 27.8$
$y_{12} = 32.8$	Chloramphenicol		
$y_{13} = 25.0$	Chloramphenicol		
$y_{14} = 24.2$	Chloramphenicol		
$y_{21} = 21.6$	Erythromycin	μ_2	$\bar{y}_{2\bullet} = 19.1$
$y_{22} = 17.4$	Erythromycin		
$y_{23} = 18.3$	Erythromycin		
$y_{24} = 19.0$	Erythromycin		
$y_{31} = 29.6$	Penicillin G	μ_3	$\bar{y}_{3\bullet} = 28.6$
$y_{32} = 24.3$	Penicillin G		
$y_{33} = 28.5$	Penicillin G		
$y_{34} = 32.0$	Penicillin G		
$y_{41} = 5.8$	Streptomycin	μ_4	$\bar{y}_{4\bullet} = 7.8$
$y_{42} = 6.2$	Streptomycin		
$y_{43} = 11.0$	Streptomycin		
$y_{44} = 8.3$	Streptomycin		
$y_{51} = 27.3$	Tetracyclin	μ_5	$\bar{y}_{5\bullet} = 31.4$
$y_{52} = 32.6$	Tetracyclin		
$y_{53} = 30.8$	Tetracyclin		
$y_{54} = 34.8$	Tetracyclin		
$\bar{y}_{\bullet\bullet} = 22.9$			

Recall: Notation for means in One-Way ANOVA -
Overall sample mean = $\bar{y}_{\bullet\bullet}$ or $\bar{y}_{++}$ (mean over index i and index j)

Treatment i sample mean = $\bar{y}_{i\bullet}$ or $\bar{y}_{i+}$ (mean over index j)

Goal: Test if the population means for these 5 treatments are plausibly equal.
If not equal, which treatment means differ significantly?

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

$$Y_{ij} = \mu_i + E_{ij}$$

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

$$\begin{aligned} \mu_1 &= \text{mean of Cholramphenicol treatment} \\ \mu_2 &= \text{mean of Erythromycin treatment} \\ &\vdots \\ \mu_5 &= \text{mean of Tetracyclin treatment} \end{aligned}$$

To test H_0 :

$$\mu_1 = \mu_2 = \ldots = \mu_5$$

vs

H_A :

at least 1 mean differs

we use software to compute the One-Way ANOVA table and look at 'global' p-value from the table.

Table for balanced one-way ANOVA:

Source	DF	SS	MS	F	P-value
Treatments	$t-1$	$SS(T)$	$MS(T) = \frac{SS(T)}{(t-1)}$	$F = \frac{MS(T)}{MS(E)}$	$P(F_{t-1,t(n-1)} > F_{obs})$
Error	$t(n-1)$	$SS(E)$	$MS(E) = \frac{SS(E)}{(N-t)}$		
Total	$nt-1$	$SS(Tot)$			

where

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

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

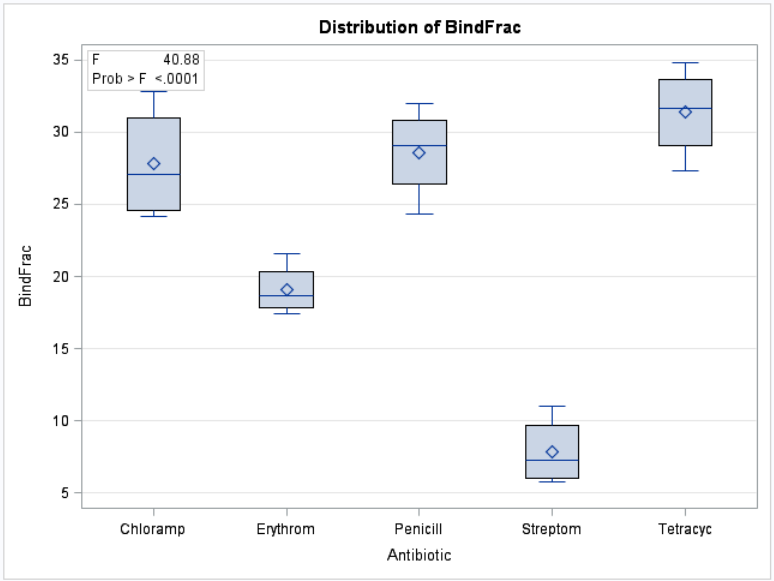
In SAS,

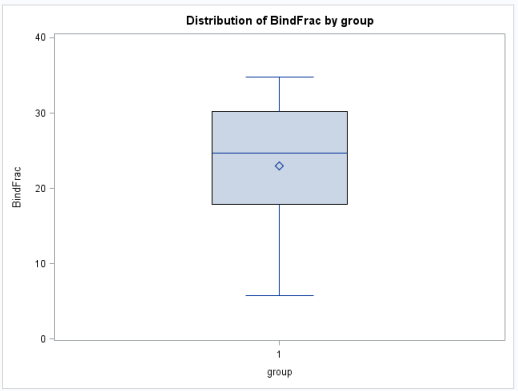
```
proc glm data=binding;
class antibiotic;
model bindfrac=antibiotic;
means antibiotic;
run;
```

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





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

Conclusion about treatment means begin equal?
p-value < 0.05 so reject H_0 in favor of H_A. That is, at the 5% S.L. there is enough evidence to say at least one antibiotic differs in terms of binding fraction.

The next step is to do pairwise comparisons of means using a multiple comparison correction.

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?

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{or} \quad \mu_1 = \mu_2 \quad vs \quad H_A : \mu_1 - \mu_2 \neq 0 \quad \text{or} \quad \mu_1 \neq \mu_2$$

We get an estimator this quantity with the corresponding sample means

$$\hat{\mu}_1 - \hat{\mu}_2 = \bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}$$

The standard error of this quantity can be found by taking the square root of the variance (recall we assume our samples are independent so covariance is 0)

$$Var(\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}) = (1^2)Var(\bar{Y}_{1\bullet}) + (-1)^2\bar{Y}_{2\bullet} + 2(1)(-1)Cov(\bar{Y}_{1\bullet}, \bar{Y}_{2\bullet})$$

$$= Var(Y_{1j})/n_1 + Var(Y_{2j})/n_2 = \sigma^2/n_1 + \sigma^2/n_2$$

For a balanced design we have

$$Var(\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}) = \sigma^2(1/n + 1/n) = 2\sigma^2/n$$

yielding a standard error of

$$SE(\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}) = \sqrt{2\sigma^2/n}$$

By the normality assumption on the data we then have a case similar to the two-sample t test with pooled variance!

$$\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet} \sim N(\mu_1 - \mu_2, \sigma^2(1/n_1 + 1/n_2)) = N(\mu_1 - \mu_2, 2\sigma^2/n)$$

We estimate σ^2 by the common pooled estimate (over all the samples, not just these two)

$$MS(E) = S_w^2 = \frac{SS(E)}{t(n-1)} = \frac{\sum_{i=1}^{t}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{i\bullet})^2}{N-t}$$

Thus, we can for a t-test for testing $H_0 : \mu_1 = \mu_2 \quad vs \quad H_A : \mu_1 \neq \mu_2$ using

$$T = \frac{\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}}{\sqrt{MS(E)(1/n_1 + 1/n_2)}} = \frac{\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}}{\sqrt{2MS(E)/n}} \sim t_{N-t} = t_{t(n-1)}$$

We can form a $(1-\alpha)100\%$ CI for $\mu_1 - \mu_2$ using

$$\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet} \pm t_{\alpha/2,N-t}\sqrt{MS(E)(1/n_1 + 1/n_2)} \quad or \quad \bar{Y}_{1\bullet} - \bar{Y}_{2\bullet} \pm t_{\alpha/2,N-t}\sqrt{2MS(E)/n}$$

and check if 0 is in the interval.

1. Create a 95% CI for $\mu_1 - \mu_2$ ($P(T_{15} > 2.13) = 0.025$) and also find the test statistic for testing the above hypotheses. What is your conclusion?

$(27.8 - 19.075) \pm 2.13\sqrt{2 * 9.0548/4} = 8.725 \pm 2.13\sqrt{4.5274} = 8.725 \pm 4.5321 = (4.19, 13.26)$

We are 95% confident the true difference in mean binding fraction between the Cholramp and Erythrom treatments is between 4.19 and 13.26.

Since 0 is not in the interval, we conclude that the treatment means differ at the 0.05 significance level.

More generally than just wanting to see which means differ, we may want to compare different functions of means.

The functions of the means will be in the form of a **linear combination** of the treatment means.
In general, a linear combination of treatment means takes the form

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

where the c_i are called the coefficients of the linear combination.

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

- $\theta_6 = 4\mu_1 + 3\mu_2 - 7\mu_5$
  yes
- $\theta_7 = 3\mu_1\mu_4 + 2\mu_3$
  no
- $\theta_8 = \mu_1 + \mu_2 + \mu_3 + \mu_4 + \mu_5$
  yes
- $\theta_9 = \mu_1^2 + 3\mu_2 + 1$
  no

If the coefficients of the linear combination sum to zero the linear combination is called a **contrast**.i.e. $c_1 + c_2 + ... + c_t = 0$ or $\sum_{i=1}^{t} c_i = 0$

Is our linear combination $\mu_1 - \mu_2 = 0$ a contrast? How about any of θ_6 through θ_9?
Yes, $\mu_1 - \mu_2$ is a contrast as 1+(-1)+0+0+0=0.
θ_6 is a contrast but θ_8 is not.
As θ_7 and θ_9 are not linear combinations, they cannot be contrasts.

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:

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

Use the output from previous to estimate the following linear combinations:

$$\theta_1 = \mu_1 \quad \hat{\theta}_1 = 27.8$$

$$\theta_2 = \mu_2 \quad \hat{\theta}_2 = 19.075$$

$$\theta_5 = \mu_5 \quad \hat{\theta}_5 = 31.375$$

$$\theta_6 = 4\mu_1 + 3\mu_2 - 7\mu_5 \quad \hat{\theta}_6 = 4 * 27.8 + 3 * 19.075 - 7 * 31.375 = -51.2$$

$$\theta_8 = \mu_1 + \mu_2 + \mu_3 + \mu_4 + \mu_5 \quad \hat{\theta}_8 = 27.8 + 19.075 + 28.6 + 7.825 + 31.375 = 114.675$$

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:

$$Var(\hat{\theta}) = \frac{c_1^2}{n_1}\sigma^2 + \frac{c_2^2}{n_2}\sigma^2 + ... + \frac{c_t^2}{n_t}\sigma^2 = \sigma^2\sum_{i=1}^{t}\frac{c_i^2}{n_i}$$

This variance involves the unknown quantity σ^2. What value can we use to estimate σ^2?

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

We can then estimate the standard error by

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

Making inference about the linear combination: (includes contrasts)

Due to the normality assumption each mean has a normal distribution and further, the linear combination will also have a normal distribution. Therefore, once we estimate the variance, we can use a t-test and a t-interval.

Relate to idea of one-sample z and one-sample t

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

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

A $(1-\alpha)100\%$ CI for θ is

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

What is the value of this test statistic and a 95% CI for the θ_8? Compare the test stat to $t_{0.025,15} = 2.13$ to make a conclusion. What is your interpretation?

$$t_{obs} = \frac{114.675}{\sqrt{9.0548\,(1^2/4 + 1^2/4 + 1^2/4 + 1^2/4 + 1^2/4)}} = \frac{114.675}{\sqrt{9.0548 * 5/4}} = 34.086$$

Rejection Region: $\{t_{obs} : |t_{obs}| > t_{0.025,15} = 2.13\} \implies$ Reject $H_0 : \theta_8 = 0$ in favor of $H_A : \theta_8 \neq 0$

At the 5% significance level, there is enough evidence to conclude that the sum of the treatment mean binding fractions is different from 0. (This is not really a useful result for this particular experiment, but this demonstrates the procedure!)

Getting the answers from SAS

Pairwise Contrasts in SAS

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.

Recall, if our global p-value is significant then our secondary goal is to find which means differ. This question is answered via looking at all pairwise comparisons (contrasts) of means.

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;

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. (To be discussed more later.)

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

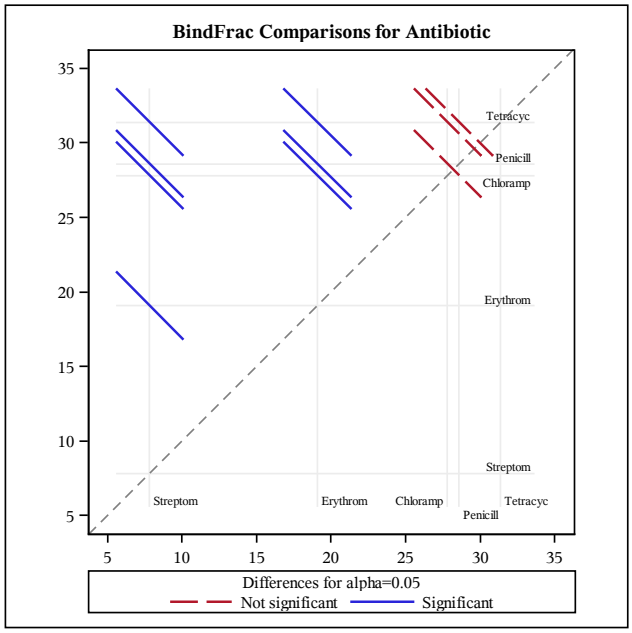
Antibiotic	BindFrac LSMEAN	Standard Error	Pr > |t|	LSMEAN Number
Chloramp	27.8000000	1.5045625	<.0001	1
Erythrom	19.0750000	1.5045625	<.0001	2
Penicill	28.6000000	1.5045625	<.0001	3
Streptom	7.8250000	1.5045625	0.0001	4
Tetracyc	31.3750000	1.5045625	<.0001	5

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 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}.

$$\theta_{10} = \frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4}{2}$$

Appropriate CI is

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

Here that is

$$\frac{27.8 + 19.075}{2} - \frac{28.6 + 7.825}{2} \pm 2.13 \sqrt{9.05483\left((1/2)^2/4 + (1/2)^2/4 + (-1/2)^2/4 + (-1/2)^2/4 + 0^2/4\right)}$$

$$23.4375 - 18.2125 \pm 2.13 \sqrt{9.05483 * 4 * (1/16)}$$

$$5.225 \pm 3.205 = (2.02, 8.43)$$

We are 95% confident the average of the Cholramp and Erythrom binding fraction means differs from the average of the Penicill and Strep binding fraction means. (Difference is between 2.02 and 8.43.)

Any linear combination in SAS.
To get the estimate for θ_{10} (and a few other linear combinations of means) in SAS we can use proc glm or use proc mixed.

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, τ_4, and τ_5 (the alternative parameterization of the model, $Y_{ij} = \mu + \tau_i + E_{ij}$).

For instance,

$$\theta_{10} = \frac{1}{2}(\mu_1 + \mu_2) - \frac{1}{2}(\mu_3 + \mu_4) = \frac{1}{2}(\mu + \tau_1 + \mu + \tau_2 - (\mu + \tau_3 + \mu + \tau_4))$$

$$= 0\mu + \frac{1}{2}\tau_1 + \frac{1}{2}\tau_2 - \frac{1}{2}\tau_3 - \frac{1}{2}\tau_4$$

In terms of syntax, we write contrast or estimate followed by a name to distinguish it. Then we do

intercept coef on μ treatment coef on τ_1 coef on τ_2 coef on τ_3 coef on τ_4 coef on τ_5

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 and a p-value.

```
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 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-wise* type I error rate will not be controlled.

For instance, we only look at all pairwise comparisons of treatment means if the global p-value is significant. Thus, these are data driven hypotheses we are testing.

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.
This is similar to the US justice system where we assume innocence until proven guilty. For most crimes it is much worse to send an innocent person to jail than to let a guilty person go free
Consider the case with $t = 5$ (antibiotic treatments):

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

  i.e. P(rejecting at least one null hypothesis that is true)
- This is called the *experiment-wise* or *family-wise* (fwe) type I error rate. We should really control this in addition to 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, what is the significance level of this test?

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

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,

$$P(\text{All 100 coins identified as fair}) = 0.12$$

$$P(\text{At least 1 coin of the 100 is classified as biased}) = 0.88$$

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

Under the assumption of independence of all tests,

$$\text{fwe rate} = P(\text{At least 1 type I error}) = \alpha^* = 1 - (1 - \alpha)^k$$

where k is the number of tests being done and α is the significance level used for each test.

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 (section 9.3)
- Tukey (section 9.5)
- Fisher's LSD (don't use this, section 9.4, you should still read this section)
- Duncan (Not required, used when there is a control treatment, section 9.7)
- Scheffé (Not required, section 9.8)

Bonferroni Correction

Suppose interest lies in exactly k linear combinations of means. **Bonferroni correction** is to replace the usual α with

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

By doing so our fwe rate will be less than α.

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

Simultaneous 95% confidence intervals for the k linear combinations of means is given by

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

$$\vdots$$

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

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

For the binding fraction example, consider only pairwise comparisons of 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?

The Margin of Errors for the CIs are all the same and are, for example,

$$t_{(\alpha'/2,15)}\sqrt{MS(E)\left(\frac{(1)^2}{4} + \frac{(-1)^2}{4} + \frac{0^2}{4} + \frac{0^2}{4} + \frac{0^2}{4}\right)} = 2.84\sqrt{(9.05)\frac{2}{4}} = 6.043$$

so that *simultaneous* 95% confidence intervals for $\theta_1, \theta_2, \theta_3, \theta_4$ are

$$\text{for } \theta_1: \quad \bar{y}_{1\bullet} - \bar{y}_{2\bullet} \pm 6.043 = 27.800 - 19.075 \pm 6.043 = (2.682, 14.768)$$

$$\text{for } \theta_2: \quad \bar{y}_{1\bullet} - \bar{y}_{3\bullet} \pm 6.043 = 27.800 - 28.600 \pm 6.043 = (-6.836, 5.236)$$

$$\text{for } \theta_3: \quad \bar{y}_{1\bullet} - \bar{y}_{4\bullet} \pm 6.043 = 27.800 - \ 7.825 \pm 6.043 = (13.939, 26.011)$$

$$\text{for } \theta_4: \quad \bar{y}_{1\bullet} - \bar{y}_{5\bullet} \pm 6.043 = 27.800 - 31.375 \pm 6.043 = (-9.611, 2.461)$$

Tukey-Kramer Correction (or just Tukey)

Tukey's correction is the best (most powerful) method when making inference on **all pairwise** contrasts 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 other multiple comparison 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 a balanced design, **simultaneous** 95% confidence intervals for $\theta = \mu_j - \mu_k$ are given by

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

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;             *Note these are different thetas than previous;
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;
```

- Tukey correction is an option you ask for:

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

proc mixed data=binding; class antibiotic;
model bindfrac=antibiotic;
lsmeans antibiotic/pdiff adjust=tukey cl;  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

Tukey Mixed output:

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

Tukey GLM output:

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

Independent Contrasts

So, what else can we do with 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 point 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 $SS(Tot)$ (with df=N-1) into independent components $SS(Trt)$ and $SS(E)$ (whose df add up to N-1).

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

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

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

(for balanced designs).

If two contrasts are orthogonal, then one contrast conveys no information about the other contrast. Hence we can break up $SS(Trt)$ into completely separate sources.

A set of k contrasts is mutually orthogonal if all pairs are orthogonal. As we have $t-1$ df, we can break $SS(Trt)$ into at most $t-1$ orthogonal contrasts.

This will allow us to attribute a certain amount of variation to given contrasts, which in turn will represent something of interest to you the researcher.

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

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

$(-1, 1, 0, 0, 0)$ and $(0, -1, 1, 0, 0)$ orthogonal ? No

Due to the joint normality of our data, **orthogonality implies independence**!

Sums of squares for contrasts
Recall we are going to partition $SS(Trt)$ into $t-1$ independent contrasts, each will have its own sum of squares. The sums of squares for a contrast are defined as

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

This contrast has 1 df associated with it.

Alternative (but equivalent) test for a contrast
Previously, we saw that we could test a linear combination (and hence a contrast) via a t-test. Now, 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 equal to 0

$$t = \frac{\hat{\theta} - 0}{\hat{SE}(\hat{\theta})} \sim t_{t(n-1)}$$

If we square a t stat we get an F stat!

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

Simultaneous test for contrasts
We can also test multiple orthogonal 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;

*I've done the estimate statements in proc mixed, but you could do them in glm instead;
*Also, we may want to do a multiple comparison correction depending on
```

Note the equivalence of the p-value for a contrast and for an estimate statement.

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

Another example

Data consists of the number of contaminants in IV fluids made by $t = 3$ pharmaceutical companies

	Cutter	Abbott	McGaw
y_{i1}	255	105	577
y_{i2}	264	288	515
y_{i3}	342	98	214
y_{i4}	331	275	413
y_{i5}	234	221	401
y_{i6}	217	240	260
$\bar{y}_{i\bullet}$	273.8	204.5	396.7

Source	d.f.	Sum of squares	Mean Square	F
Treatments (pharmacy)	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?
SAS uses alphabetical order for the factor levels so

$$\mu_1 = \mu_A \qquad \mu_2 = \mu_C \qquad \mu_3 = \mu_M$$

Rewrite these contrasts in terms of μ_1, μ_2, and μ_3.

$$\theta_1 = \mu_3 - \mu_1 = -\mu_1 + 0\mu_2 + \mu_3$$

$$\theta_2 = \mu_2 - \frac{\mu_3 + \mu_1}{2} = (-1/2)\mu_1 + 1\mu_2 + (-1/2)\mu_3$$

Are these contrasts orthogonal (and thus independent)?
Yes, -1*(-1/2)+0*1+1*(-1/2)=0

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

$$\hat{\theta}_1 = 396.7 - 204.5 = 192.2 \implies SS(\hat{\theta}_1) = \frac{192.2^2}{(-1)^2/6 + 1^2/6} = \frac{192.2^2}{1/3} = 110784.083$$

$$\hat{\theta}_2 = 273.8 - \frac{204.5 + 396.7}{2} = -26.75 \implies SS(\hat{\theta}_2) = \frac{(-26.75)^2}{(-1/2)^2/6 + 1^2/6 + (-1/2)^2/6} = 2862.25$$

These should add up to SS(Trt) since SS(Trt) has 2 df and we have 2 orthogonal contrasts.

```
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

For another good example, see page 457, example 9.3.

Chapter 3

ST 512 - Analysis of Factorial Designs (Multiway ANOVA)

Readings: 14.1-14.6, especially pg 886-887,891,904

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 treatments) 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	$\bar{y}_{1\bullet} = 212.0$	$s_1 = 40$
OM (i=2)	192	253	248	278	232	267	289	$\bar{y}_{2\bullet} = 251.3$	$s_2 = 32$
YF (i=3)	221	213	202	183	185	197	162	$\bar{y}_{3\bullet} = 194.7$	$s_3 = 20$
YM (i=4)	271	192	189	209	227	236	142	$\bar{y}_{4\bullet} = 209.4$	$s_4 = 41$

```
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

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

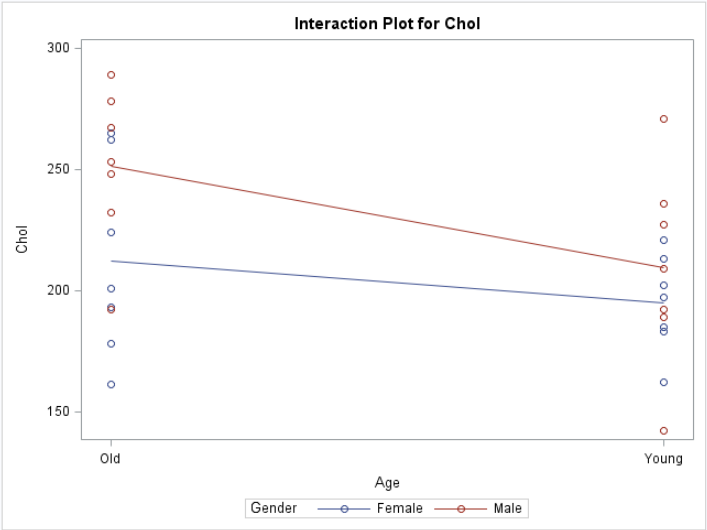
Source	DF	Type I SS	Mean Square	F Value	Pr > F
Treatment	3	12280.85714	4093.61905	3.46	0.0323

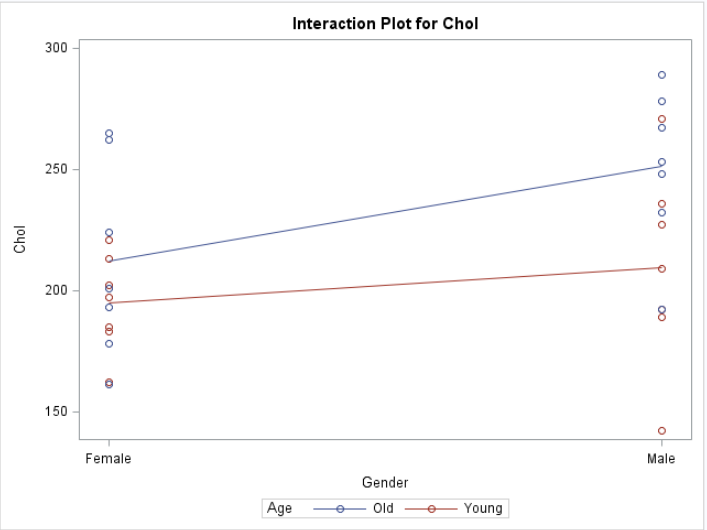
Source	DF	Type III SS	Mean Square	F Value	Pr > F
Treatment	3	12280.85714	4093.61905	3.46	0.0323

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 (profile or interaction) plots below:





Define 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?

$\theta_{Age} = \frac{1}{2}(\mu_3 + \mu_4) - \frac{1}{2}(\mu_1 + \mu_2) =$ avg young - avg old

What contrast would test for the *main effect* of Gender?

$\theta_{Gender} = \frac{1}{2}(\mu_1 + \mu_3) - \frac{1}{2}(\mu_2 + \mu_4) =$ avg female - avg male

If the main effect contrast for factor A is significantly different from 0, then that **factor** is important for predicting the response.

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?

$\theta_{Age*Gender} = \frac{1}{2}(\mu_1 - \mu_3) - \frac{1}{2}(\mu_2 - \mu_4) = \frac{1}{2}(\mu_1 - \mu_2 - \mu_3 + \mu_4)$

1. Check that θ_{Age}, θ_{Gender}, and $\theta_{Age*Gender}$ are mutually orthogonal.
   Age and Gender main effects: (-1/2)*(1/2)+(-1/2)*(-1/2)+(1/2)*(1/2)+(1/2)*(-1/2)=0

   Age main effect and interaction effect: (-1/2)*(1/2)+(-1/2)*(-1/2)+(1/2)*(-1/2)+(1/2)*(1/2)=0

   Gender main effect and interaction effect: (1/2)*(1/2)+(-1/2)*(-1/2)+(1/2)*(-1/2)+(-1/2)*(1/2)=0
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 $\hat{SE}(\hat{\theta}) = \sqrt{MS(E)\sum_{i=1}^{t} \frac{c_i^2}{n_i}}$

   $\hat{\theta}_{Age} = (-1/2) * 212 + (-1/2) * 251.3 + (1/2) * 194.7 + (1/2) * 209.4 = -29.6$

   $\hat{\theta}_{Gender} = (1/2) * 212 + (-1/2) * 251.3 + (1/2) * 194.7 + (-1/2) * 209.4 = -27$

   $\hat{\theta}_{Age*Gender} = (1/2) * 212 + (-1/2) * 251.3 + (-1/2) * 194.7 + (1/2) * 209.4 = -12.3$

   $\hat{SE}(\hat{\theta}) = \sqrt{1184.77(1/7)(1/2)^2 * 4} = 13.01$ for all three contrasts.
3. Find the sums of squares for each contrast. How many degrees of freedom are associated with each contrast? Recall: $SS(\hat{\theta}) = \frac{\hat{\theta}^2}{\sum \frac{c_i^2}{n_i}}$.

   $SS(\hat{\theta}_{Age}) = \frac{(-29.6)^2}{(1/7)*4*(1/2)^2} = 6133.12$

   $SS(\hat{\theta}_{Gender}) = \frac{(-27)^2}{(1/7)*4*(1/2)^2} = 5103$

   $SS(\hat{\theta}_{Age*Gender}) = \frac{(-12.29)^2}{(1/7)*4*(1/2)^2} = 1057.31$

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 ($F_{0.05,1,24} = 4.26$) and make a decision about the importance of each effect.

   $F_{Age} = \frac{6133.12}{1184.77} = 5.18$ $\quad num\ df = \underline{1}\ \ den\ df = \underline{24}$

   $F_{Gender} = \frac{5103}{1184.77} = 4.31$ $\quad num\ df = \underline{1}\ \ den\ df = \underline{24}$

   $F_{Age*Gender} = \frac{1057.31}{1184.77} = 0.89$ $\quad num\ df = \underline{1}\ \ den\ df = \underline{24}$

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?

   $$SS(Trt) = SS(\hat{\theta}_{Age}) + SS(\hat{\theta}_{Gender}) + SS(\hat{\theta}_{Age*Gender})$$

   Test Statistic is

   $$F = \frac{\frac{SS(\hat{\theta}_{Age}) + SS(\hat{\theta}_{Gender}) + SS(\hat{\theta}_{Age*Gender})}{3}}{MS(E)} = 3.46$$

   compare against $F_{0.05,3,24} = 3.009$. Reject h_0 in favor of H_A.

Notice what we've done:
Very similar to partitioning the $SS(Tot)$ into $SS(Trt)$ 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 allows for much more insight!

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. Recall: 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 coef on μ treatment coef on τ_1 coef on τ_2 coef on τ_3 coef on τ_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).

Tests of each contrast and estimate individually:

```
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

Test of the contrasts simultaneously, **same as global f-test here:

```
proc glm data=cholesterol;  class Treatment;  model Chol=Treatment;
contrast 'All three at once' intercept 0 Age 1 -1,
 intercept 0 Age 0 0 Gender 1 -1,
 intercept 0 Age 0 0 Gender 0 0 Age*Gender 0.5 -0.5 -0.5 0.5; run;
```

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
All three at once	3	12280.85714	4093.61905	3.46	0.0323

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 that will **automatically test for these contrasts of interest!**

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 $k = 1, 2, \ldots, 7$.

- 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,

$$\text{estimate is } \hat{\mu} = \bar{Y}_{\bullet\bullet\bullet}$$

- α_i represents the 'effect' for being at level i of Age,

$$\text{estimate is } \hat{\alpha}_i = \bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet\bullet\bullet}$$

- β_i represents the 'effect' for being at level j of Gender,

$$\text{estimate is } \hat{\beta}_j = \bar{Y}_{\bullet j\bullet} - \bar{Y}_{\bullet\bullet\bullet}$$

- $(\alpha\beta)_{ij}$ represents the 'joint effect' for being at level i of Age and level j of Gender,

$$\text{estimate is } \hat{(\alpha\beta)}_{ij} = \bar{Y}_{ij\bullet} - \bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet j\bullet} + \bar{Y}_{\bullet\bullet\bullet}$$

- E_{ijk} is a random error

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 **factorial 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 crossed would be a 2×3×5 factorial design.)

For level i of Age and level j of Gender we are model the mean cholesterol as

$$\mu_{ij} = E(Y_{ijk}) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

and, as you might expect, the estimate for level i of Age and level j of Gender is

$$\hat{\mu_{ij}} = \bar{Y}_{\bullet\bullet\bullet} + (\bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet\bullet\bullet}) + (\bar{Y}_{\bullet j\bullet} - \bar{Y}_{\bullet\bullet\bullet}) + (\bar{Y}_{ij\bullet} - \bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet j\bullet} + \bar{Y}_{\bullet\bullet\bullet}) = \bar{Y}_{ij\bullet}$$

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, look at *simple effects*.
   (b) If not significant, investigate the main effect p-values for significance to determine which factors are important for predicting the response - look at those *main effects*.

Types of Effects Investigated

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 $\hat{\mu}_4$)

Simple Effects

1. In a Two-way ANOVA problem, 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 $\mu_{21} - \mu_{11}$,

   Simple effect of A at level 2 of B is defined as $\mu_{22} - \mu_{12}$,

2. Define the simple effects of factor B.

   Simple effect of B at level 1 of A is defined as $\mu_{12} - \mu_{11}$,

   Simple effect of B at level 2 of A is defined as $\mu_{22} - \mu_{21}$,

3. For the Cholesterol example, estimate these four simple effects and explain what each measures.

   Simple effect of A at level 1 of B is estimated by $194.7 - 212 = -17.3$,

   Age difference for females.

   Simple effect of A at level 2 of B is estimated by $209.4 - 251.3 = -41.9$

   Age difference for males.

   Simple effect of B at level 1 of A is estimated by $251.3 - 212.0 = 39.3$

   Gender difference for old.

   Simple effect of B at level 2 of A is estimated by $209.4 - 194.7 = 14.7$

   Gender difference for young.

Main Effects

1. Main effects in a 2x2 experiment are the averages of the simple effects. Define the main effect of factor A as

$$\mu_A = \frac{1}{2}\left((\mu_{22} - \mu_{12}) + (\mu_{21} - \mu_{11})\right) = \frac{1}{2}(\mu_{22} + \mu_{21}) - \frac{1}{2}(\mu_{12} + \mu_{11})$$

   Our θ_{Age} contrast from before!

2. Define the main effect for factor B.

$$\mu_B = \frac{1}{2}\left((\mu_{12} - \mu_{11}) + (\mu_{22} - \mu_{21})\right) = \frac{1}{2}(\mu_{12} + \mu_{22}) - \frac{1}{2}(\mu_{11} + \mu_{21})$$

   *****Main effects should (usually) only be looked at when interaction effects are not significant.

3. For the Cholesterol example, estimate these two main effects and explain what each measures.

$$\mu_A = -29.6 \text{(young mean vs old mean, found earlier)}$$
$$\mu_B = 27 \text{(male mean vs female mean, negative of earlier contrast)}$$

Interaction Effects

Interaction effects in a 2x2 experiment are the average of the *difference* of simple effects. The value of this effect is not usually of interest, mostly we just want to find out if the interaction is significant.

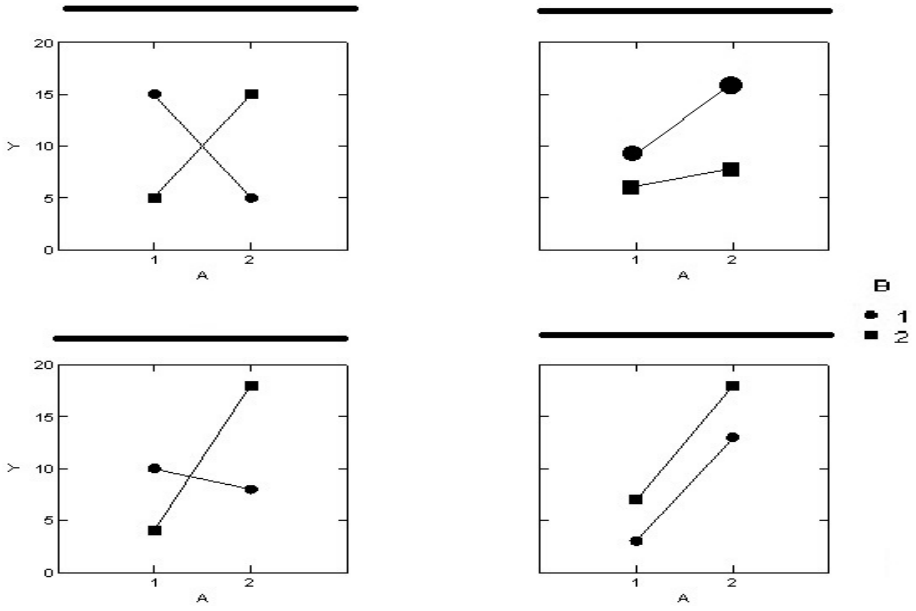
If an interaction is present, both factors are important and we look at simple effects only. Why?

If an interaction is present that implies that the effect of Age on cholesterol depends on which level of Gender you are (and vice-versa). This by itself says that both Age and Gender are affecting the response.

Visually we can investigate interactions via a 'profile' or 'interaction' plot.

1. If an interaction effect causes the relationship between the levels of a factor to change, this is called a **qualitative** interaction.
2. If the interaction effect just changes the magnitude of the relationship between the levels of a factor, this is called a **quantitative** interaction.

Label the plots below accordingly:



TL - Qualitative, TR - Quantitative, BL - Qualitative, BR - No Interaction
If interactions exist, looking at main effects is usually not necessary. For example, approximately what would the main effect for factor B be equal to for the top left plot?

Main effect is average of simple effects:

$$\mu_B = \frac{1}{2}\left((\mu_{12} - \mu_{11}) + (\mu_{22} - \mu_{21})\right) \approx \frac{1}{2}((5 - 15) + (15 - 5)) = 0$$

This would say that Gender is not important based on the main effect only, but clearly gender is important here!

To test for the Age main effect notice,

$$\begin{aligned}
\hat{\text{A main}} \text{ effect} &= \frac{1}{2}(\hat{\mu}_{22} - \hat{\mu}_{12} + \hat{\mu}_{21} - \hat{\mu}_{11}) \\
&= \frac{1}{2}(\bar{Y}_{22\bullet} - \bar{Y}_{12\bullet} + \bar{Y}_{21\bullet} - \bar{Y}_{11\bullet}) \\
&= \frac{1}{2}(\bar{Y}_{22\bullet} + \bar{Y}_{21\bullet}) - \frac{1}{2}(\bar{Y}_{12\bullet} + \bar{Y}_{11\bullet}) \\
&= \bar{Y}_{2\bullet\bullet} - \bar{Y}_{1\bullet\bullet} \\
&= \hat{\alpha}_2 - \hat{\alpha}_1
\end{aligned}$$

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.

Again the point of this section is that once we have a significant global p-value, we then want to attribute the significance to our factors and/or their interaction. This model allows us to do so by testing groups of parameters!

The general Two-Way ANOVA model:

Suppose we have a *continuous* response, Y, and two factors, A and B from a CRD.

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). Later in the semester we'll look at 'nested designs.'

1. Factor A has a levels and factor B has b levels.
2. There are n replicates for each treatment.
3. Total of $N = abn$ EUs.
4. Our main interest lies in whether or not the response differs due to the factors.

The parametrization of the Two-way ANOVA model we will use is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + E_{ijk}, \quad i = 1, ..., a, \quad j = 1, ..., b, k = 1, ..., n \text{ balanced design}$$

where $E_{ijk} \sim^{iid} N(0, \sigma^2)$ and we have sum to zero constraints on the parameters (to get a unique solution).

- α_i is an effect due to level i of factor A
- β_j is an effect due to level j of factor B
- $(\alpha\beta)_{ij}$ is an joint effect due to level i of factor A and level j of factor B

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 (using contrasts as before, just a little more complicated now):

ANOVA table for a×b (balanced) factorial experiment

Source	df	Sum of Squares (SS)	Mean Square (MS)	F-stat
A	$a-1$	$SS(A)$	MS(A)=SS(A)/(a-1)	MS(A)/MS(E)
B	$b-1$	$SS(B)$	MS(B)=SS(B)/(b-1)	MS(B)/MS(E)
AB	$(a-1)(b-1)$	$SS(AB)$	MS(AB)=SS(AB)/((a-1)(b-1))	MS(AB)/MS(E)
Error	$ab(n-1)$	$SS(E)$	MS(E)=SS(E)/(ab(n-1))	
Total	$N-1$	$SS(Tot)$		

$$SS(A) = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} (\bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet\bullet\bullet})^2$$

$$SS(B) = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} (\bar{Y}_{\bullet j\bullet} - \bar{Y}_{\bullet\bullet\bullet})^2$$

$$SS(AB) = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} (\bar{Y}_{ij\bullet} - \bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet j\bullet} + \bar{Y}_{\bullet\bullet\bullet})^2$$

$$SS(E) = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} (Y_{ijk} - \bar{Y}_{ij\bullet})^2$$

$$SS(Tot) = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} (Y_{ijk} - \bar{Y}_{\bullet\bullet\bullet})^2$$

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)

$SS(A) + SS(B) + SS(AB) + SS(E) = SS(Tot)$

$df_A + df_B + df_{AB} + df_E = df_{Tot}$.

An a×b example

An entomologist records energy expended (Y) by $N = 18$ 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	y_{111}=3.1	y_{112}=3.7	y_{113}=4.7
20	40	y_{121}=5.5	y_{122}=6.7	y_{123}=7.3
30	20	y_{211}=6	y_{212}=6.9	y_{213}=7.5
30	40	y_{221}=11.5	y_{222}=12.9	y_{223}=13.4
40	20	y_{311}=7.7	y_{312}=8.3	y_{313}=9.5
40	40	y_{321}=15.7	y_{322}=14.3	y_{323}=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

Answer the following:

1. Is anything in the model useful?
2. If so, what effects should we investigate and why?

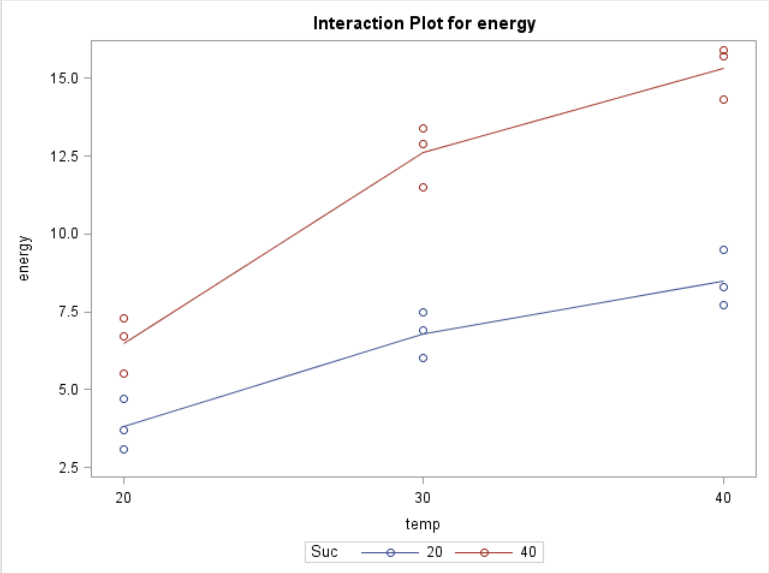
1. Yes, the global p-value is < 0.0001 which implies something in our model is useful.
2. We start with the interaction p-value which is 0.0040 and significant for most any SL. Therefore, we should investigate the simple effects of temp and the simple effects of Suc.

Unlike a 2×2 study, it is not possible to express interaction between Temp and Suc using 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. This is because no interaction implies changing the level of Temp doesn't change the effect of Suc on Energy, i.e.

$$\mu_{12} - \mu_{11} = \mu_{22} - \mu_{21} \text{ or } \mu_{Temp20,Suc40} - \mu_{Temp20,Suc20} = \mu_{Temp30,Suc40} - \mu_{Temp30,Suc20}$$

$$\mu_{12} - \mu_{11} = \mu_{32} - \mu_{31} \text{ or } \mu_{Temp20,Suc40} - \mu_{Temp20,Suc20} = \mu_{Temp40,Suc40} - \mu_{Temp40,Suc20}$$

$$\mu_{32} - \mu_{31} = \mu_{22} - \mu_{21} \text{ or } \mu_{Temp40,Suc40} - \mu_{Temp40,Suc20} = \mu_{Temp30,Suc40} - \mu_{Temp30,Suc20}$$

Notice that given any two of these contrasts (move all means to one side to see these as contrasts), we can get the third contrast. So we have 3-1=2 contrasts that are needed for testing interaction.

In terms of the plot, 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 $ab(n-1)$ denominator df.

For honeybee data,

$$F = MS(AB)/MS(E) = 6.976/0.7722 = 9.03$$

which is significant ($p = 0.0040$) at the 0.05 S.L. on 2 and 12 degrees of freedom.

As interaction is significant, both factors are important! Our next step would be to analyze the **simple effects** of each factor.

That is, we would investigate the **effects of Temperature at given levels of Sucrose** and also look at the **effect of Sucrose at given levels of Temperature**.

Inspection of main effects in not appropriate here.

To 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

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). Here we are given way more than that!

We could write estimate or contrast to get just those, but this is easier. However, as we are correcting for multiple comparisons, we may be correcting too much! In real life, you'd take the time to write the contrast and estimate statements of interest.

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

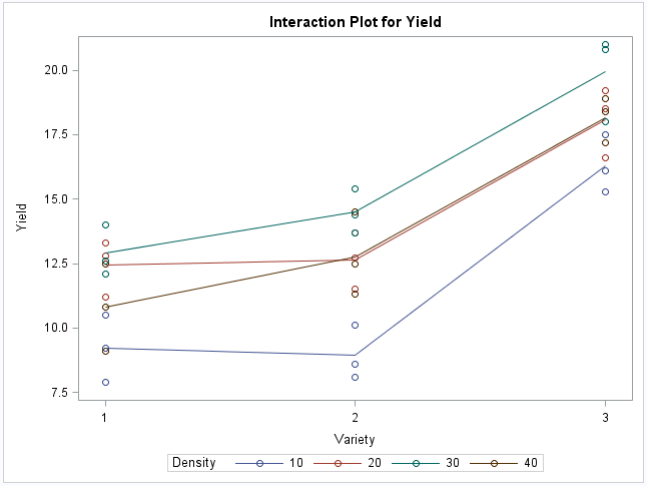
```
proc glm data=tomato; class variety density;
model Yield=Variety|Density;
lsmeans Variety Density/adjust=tukey cl; *adjust=tukey tells sas to do pdiff; run;
```

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



We proceed with our usual approach:

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, look at *simple effects*.
   (b) If not significant, investigate the main effect p-values for significance to determine which factors are important for predicting the response - look at those *main effects*.

1. Global p-value is signficant.
2. As the interaction is not significant we want to look at main effects p-values.
3. These are significant for both factors, thus both are important. We should now look at the main effect differences for both factors (we would not look at differences for a factor deemed non-significant).

Ideas for the main effect in an axb experiment:
Note that the variety main effect has 2 df. That implies that this effect comes from testing two contrasts simultaneously.

Main effects are averages of simple effects, so we must define the simple effects of A.

We have ones corresponding to switching from variety 2 to variety 1:

$$\mu_{21}-\mu_{11} \quad \mu_{22}-\mu_{12} \quad \mu_{23}-\mu_{13} \quad \mu_{24}-\mu_{14}$$

$\mu_{Var2,Den10}-\mu_{Var1,Den10}$ $\mu_{Var2,Den20}-\mu_{Var1,Den20}$ $\mu_{Var2,Den30}-\mu_{Var1,Den30}$ $\mu_{Var2,Den40}-\mu_{Var1,Den40}$
We have ones corresponding to switching from variety 3 to variety 1:

$$\mu_{31}-\mu_{11} \quad \mu_{32}-\mu_{12} \quad \mu_{33}-\mu_{13} \quad \mu_{34}-\mu_{14}$$

$\mu_{Var3,Den10}-\mu_{Var1,Den10}$ $\mu_{Var3,Den20}-\mu_{Var1,Den20}$ $\mu_{Var3,Den30}-\mu_{Var1,Den30}$ $\mu_{Var3,Den40}-\mu_{Var1,Den40}$
We have ones corresponding to switching from variety 3 to variety 2:
Notice that these could be found by subtracting above simple effects. For example,

$$(\mu_{31}-\mu_{11})-(\mu_{21}-\mu_{11})=\mu_{31}-\mu_{32}$$

So these are redundant and not needed.

Now the main effects are the averages of these groups of simple effects. That is, one component of the main effect is

$$\frac{1}{4}\left((\mu_{21}-\mu_{11})+(\mu_{22}-\mu_{12})+(\mu_{23}-\mu_{13})+(\mu_{24}-\mu_{14})\right)$$
$$=\frac{1}{4}\left(\mu_{21}+\mu_{22}+\mu_{23}+\mu_{24}\right)-\frac{1}{4}\left(\mu_{11}+\mu_{12}+\mu_{13}+\mu_{14}\right)$$

the average of the responses at level 2 of variety against the average of the responses at level 1 of variety. (Which should make sense intuitively.)

Similarly, averaging the other group of simple effects gives the average of the responses at level 3 of variety against the average of the responses at level 1 of variety.

The average of 3 vs 2 could be found by subtracting and so it is not needed. Thus, we have 2 df for testing this main effect.

First let's look at the Variety main effects (lsmeans Variety/adjust=tukey cl; output):

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

Now let's look at the Density main effects (lsmeans Density/adjust=tukey cl; output):

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

Ideas can be easily extended to an arbitrary number of factors

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 2×2×3 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, just as in the Two-Way model.

```
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

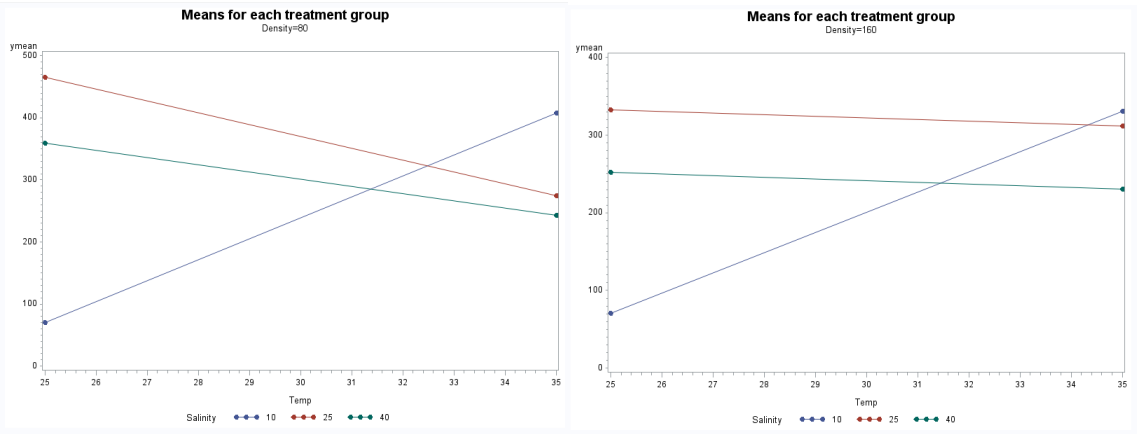
Source	DF	Type III SS	Mean Square	F Value	Pr > F
Temp	1	15376.0000	15376.0000	5.30	0.0304
Density	1	21218.7778	21218.7778	7.31	0.0124
Temp*Density	1	8711.1111	8711.1111	3.00	0.0961
Salinity	2	96762.5000	48381.2500	16.66	<.0001
Temp*Salinity	2	300855.1667	150427.5833	51.80	<.0001
Density*Salinity	2	674.3889	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 (35 degrees vs 25 degrees) at density 80 and salinity 10.
- the effect of density (160 vs 80) at temperature 25 and salinity 40.
- the effect of salinity (40 vs 10) at temperature 35 and density 80.
- Any effects like these that are of interest to the researcher. Key is that we aren't averaging over any of the variables as the three-way interaction is significant. This implies that, for instance, the way temperature effects the response depends on both density and salinity together. Therefore, it doesn't make sense to average across density, salinity, or both when looking at temperature.

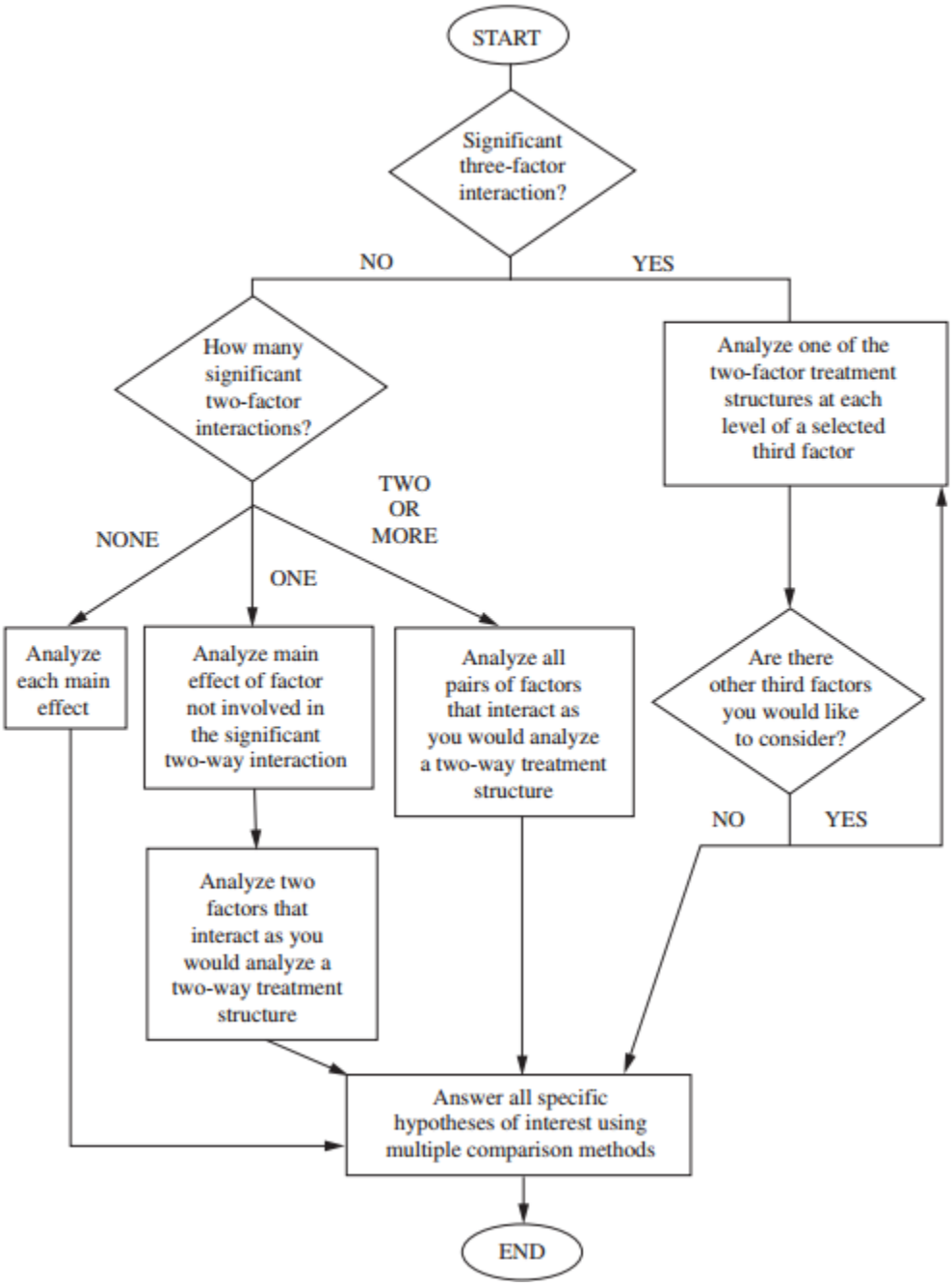
Interpretation of three-way (second order) interaction



Suppose the 3-way interaction was not significant. Then we would proceed to the 2-way interaction p-values and, if those were not significant, the main effect p-values.

For example, we might find the only significant effects were the Temp*Salinity, Salinity, and Density. What should we do here?
We should investigate the simple effects of temperature and salinity (averaged over density). That is, effects such as Temp 35 vs Temp 25 at Salinity 10 (averaged over Density). This could be found by
lsmeans Temp*Salinity/pdiff;
or by writing the appropriate contrast/estimate statements.

We would also look at the main effects of Density. That is, the effect of Density 160 vs Density 80. This could be found by
lsmeans Density/pdiff; or by writing the appropriate contrast/estimate statements.


START
Significant three-factor interaction?
NO
YES
How many significant two-factor interactions?
Analyze one of the two-factor treatment structures at each level of a selected third factor
NONE
ONE
TWO OR MORE
Analyze each main effect
Analyze main effect of factor not involved in the significant two-way interaction
Analyze all pairs of factors that interact as you would analyze a two-way treatment structure
Are there other third factors you would like to consider?
NO
YES
Analyze two factors that interact as you would analyze a two-way treatment structure
Answer all specific hypotheses of interest using multiple comparison methods
END

Chapter 4

ST 512 - Correlation

Readings for Correlation and SLR: 11.1-11.5, 11.7-11.8

Until now, we've been considering a continuous variable (our response) and at least 1 categorical variable (our predictors) measured on the same individuals. Now we'll investigate ways to analyze **two quantitative variables measured on the same units**.

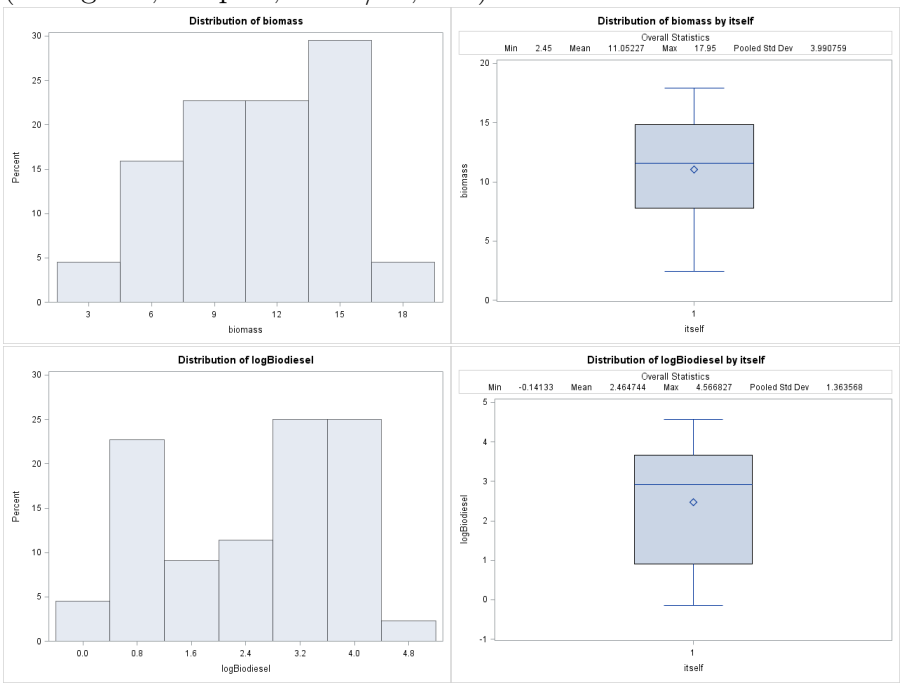
We'll start by looking at correlation and simple linear regression. Later we'll move to multiple linear regression which has one response and p quantitative predictors. Finally, we'll combine the Multi-way ANOVA (factorial effects) modeling with this regression model in what are called **General Linear Models (GLMs)**.

When is correlation appropriate?
Consider having two variables X and Y (no need to designate one as a response and one as an explanatory). We may want to look at the linear association between the two variables.

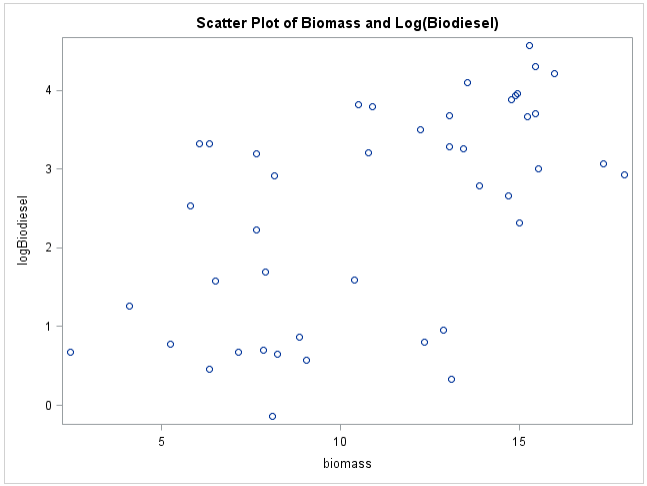
That is, attempt to investigate how the two variables vary together.

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 descriptive 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.

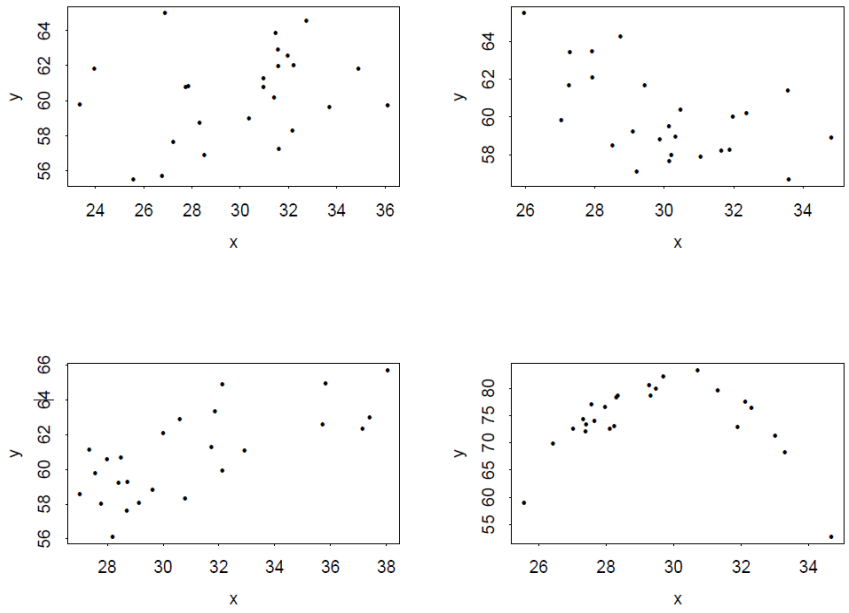


What to look for in a scatterplot?

1. **Form** - Type of overall pattern: linear, quadratic, logarithmic etc.

2. **Strength** - Point closely or loosely follow the form?

3. **Direction** - Pattern positive, negative, or NA?



With a histogram we might visually investigate the sample mean, which estimates a parameter of interest μ.

Similarly here, we can visually inspect the linear association, but there is also a statistic that measures this linear relationship.

We also have a parameter of interest here, which we label ρ and call the correlation coefficient.

The population correlation coefficient ρ.

Correlation is a unitless measure of the strength and direction of the *linear* relationship between two RVs.

ρ or (ρ_{XY} when we want to be clear which variables we are talking about) is the population parameter that measures correlation between X and Y.

The definition of ρ_{XY} is

$$\rho_{XY} = E\left[\frac{(X-\mu_X)}{\sigma_X}\frac{(Y-\mu_Y))}{\sigma_Y}\right] = \rho.$$

Here, $E(\cdot)$ denotes mathematical expectation (basically meaning the average value of the function in the parenthesis).

This is similar to the way that the true mean of a random variable is defined, $\mu = E(Y)$

The sample correlation coefficient R and r.

Given n independent pairs of quantitative data:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \qquad \text{or}$$

Quant. Var 1	Quant. Var 2
y_1	x_1
y_2	x_2
$\ldots$	$\ldots$
y_n	x_n

Sample correlation coefficient - r_{XY} of the paired data is defined by

$$r_{XY} = \frac{\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)}}{\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{(n-1)} * \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{(n-1)}}} = \frac{s_{xy}}{s_x s_y}$$

s_{xy} is called the sample covariance of X and Y:

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}.$$

At times, this may be rewritten as

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 * \sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

Just as $\bar{Y}$ provides empirical information about a population mean μ_Y, R_{XY} can be used for inference about the *population correlation coefficient* ρ.

Note, we can distinguish between the RV and the realized value, R_{XY} is an RV and r_{XY} is the observed value of that random variable

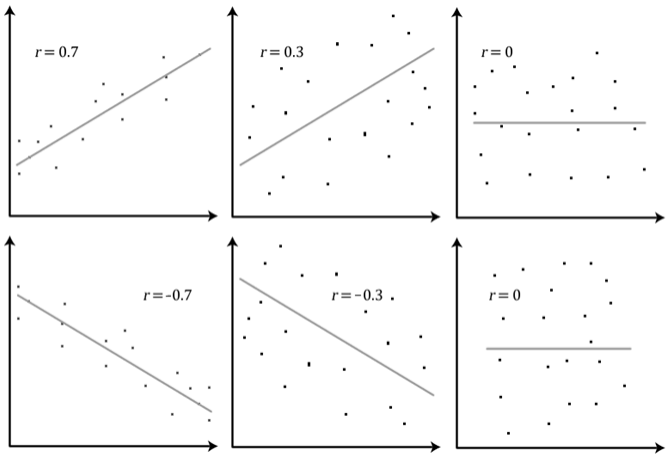
Properties of r_{XY}

- 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)
- If $r_{XY} \approx 0$, then no linear relationship. Why? (Note: We will do a test for $\rho = 0$ later).

Some example scatter plots



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

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

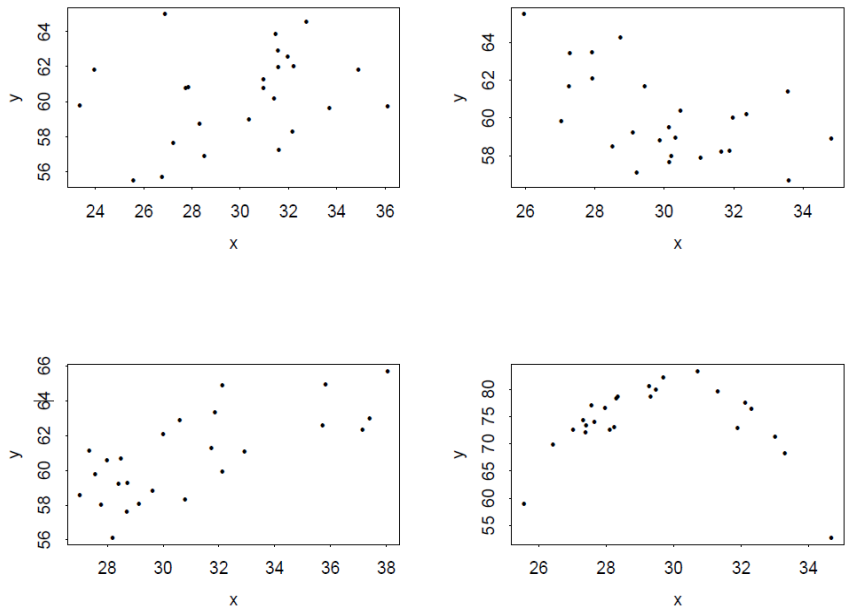
$$s_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = 3.1485$$

Applying the formula for r_{XY}, we get

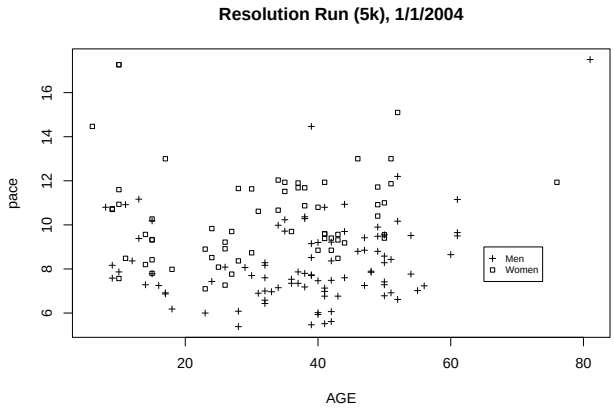
$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{3.1485}{3.9908 \times 1.3636} = 0.5786$$

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?



Inference for ρ

r is just a point estimate for ρ. If we get a new sample, r will change. We want to use the sample value to make a claim about ρ.

Relate to inference for μ. $\bar{y}$ is a point estimate for μ. To make a claim or statement about μ, we had to do a hypothesis test or a confidence interval.
This required knowledge of the **sampling distribution** of $\bar{Y}$ (or $\frac{\bar{Y}-\mu}{S/\sqrt{n}}$, a function of $\bar{Y}$).

Similarly we need to know the distribution of R or a function of R.

Much inference can somehow be related to the normal distribution.

1. R_{XY} is between -1 and 1, so the normal distribution is not ideal for this statistic.
2. Instead we look at a transformation of R that follows a normal distribution in 'large' samples. Consider 'Fisher's Transformation'
$$log\left(\frac{1+R}{1-R}\right)$$
3. When R is very close to -1, we are approaching log evaluated at 0. When R is very close to 1, we are approaching log evaluated at ∞ - so it takes on values from $(-\infty, \infty)$.
4. Thus, a test statistic useful for inference about ρ is
$$Z(\rho) = \left(\frac{1}{2}\sqrt{n-3}\right)\left(\log\left(\frac{1+R}{1-R}\right) - \log\left(\frac{1+\rho}{1-\rho}\right)\right) \sim N(0,1) \text{ for large n}$$

To perform a Hypothesis Test about ρ:

We often want to test the following hypotheses (although a one-sided test could be done),

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

Fisher's Z - Assuming H_0 is true, the observed test statistic is

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

The assumptions are that we have a large n and a random X and Y (although if X is fixed, inference still works ok).

The reference distribution is the standard normal distribution.

$$\text{Rejection Region: RR} = \left\{z_{obs} : |z_{obs}| > z_{\alpha/2}\right\}$$

The p-value is found by

$$2P(Z > |z_{obs}|)$$

T-test - Assuming H_0 is true, another test statistic is

$$t_{obs} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

The assumptions are the X and Y have a 'bivariate normal distribution' (although inference is robust to this assumption).

The reference distribution is t_{n-2}.

$$\text{Rejection Region: RR} = \left\{t_{obs} : |t_{obs}| > t_{\alpha/2,n-2}\right\}$$

The p-value is found by

$$2P(T_{n-2} > |t_{obs}|)$$

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

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

Fisher's Z gives 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 $|z_{obs}| > 1.96$.

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. (Why do we multiply by 2?)

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.

T-test gives a test statistic of

$$t_{obs} = \frac{0.5786\sqrt{44-2}}{\sqrt{1-0.5786^2}} = 4.597$$

Using an $\alpha = 0.05$ our rejection region is any $|t_{obs}| > 2.018$.

Our p-value $= 2P(T_{42} > 4.597) = 2(0.00002) = 0.00004 < \alpha = 0.05$ so we reject our null hypothesis in favor of the alternative.

To find a Confidence Interval for ρ:

An approximate $100(1-\alpha)\%$ confidence interval for ρ can be obtained by inverting the Fisher transformation hypothesis test:

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

This formula differs slightly from that in the book, but it is equivalent.

Let's find a corresponding 95% confidence interval for ρ from the biodiesel and biomass example:

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

Interpret the interval in the context of the problem. Also, state what confidence means here.

We are 95% confident that the true correlation (ρ) between log(biodiesel) and biomass is between 0.3401 and 0.7471.

When we say confident, we mean that if we did this experiment repeatedly (sample 44 plants, made measurements etc.) 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) cov;
var biomass logbiodiesel;
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 that there are also other statistics that measure the linear relationship. The sample mean is strongly affected by outliers and so r is as well. Another version of correlation is Spearman's correlation. This statistic uses the relative ranks of the data points rather than the values themselves. This makes it more robust to outliers and a good statistic to use.

Wrap up of Correlation

When we have two quantitative variables, we want to describe the relationship between them.

A basic relationship is a linear one. Correlation (ρ) is a statistical measure of this quantity.

We can estimate ρ by r, the sample correlation. A hypothesis test of interest is usually

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

The test can be carried out using a t-test or a normal based test.

Similarly a confidence interval can be created for ρ.

If the test is significant or the interval does not contain 0, then we know the two variables have a significant linear relationship. For instance, we may be able to conclude that having a high stress level is significantly associated with having high blood pressure.

Significant correlation does NOT imply causation!

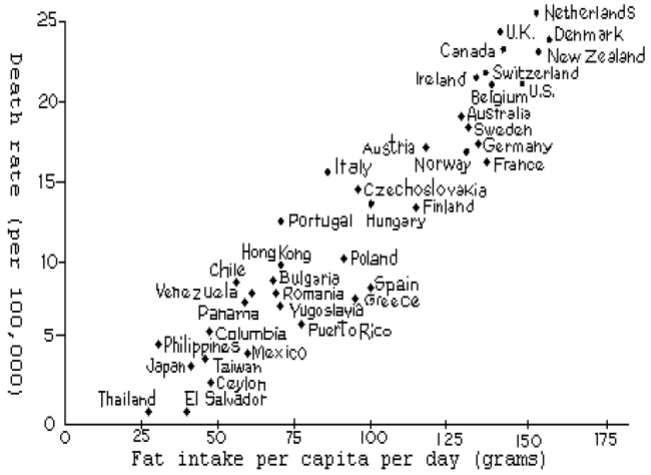
Please read the following famous examples of *spurious correlations*:

- 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. a possible 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?

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

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)

Chapter 5

ST 512 - Simple Linear Regression

Readings for Correlation and SLR: 11.1-11.5, 11.7-11.8

Correlation is one way to measure the linear association between two quantitative variables measure in pairs. However, this doesn't help us if we want to use one variable to try and predict the value of the other. Instead, we can actually fit a linear model and use the line to make inferences.

What 'data situation' and relationship are we considering here? Given n independent pairs of quantitative data measured on the same individuals:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

we may want to look at the linear relationship between X and Y.

Two of the methods for investigating the linear relationship:

1. **Correlation analysis** -
   Find the sample correlation r_{xy}, do a HT or CI to determine if ρ=0 is a reasonable value.

   Note: ALWAYS need to look at scatterplot! even if ρ=0 is reasonable, it does NOT imply that there is no *relationship* between the variables, just no linear relationship!

   SAS proc corr with the fisher option will perform the appropriate tests.

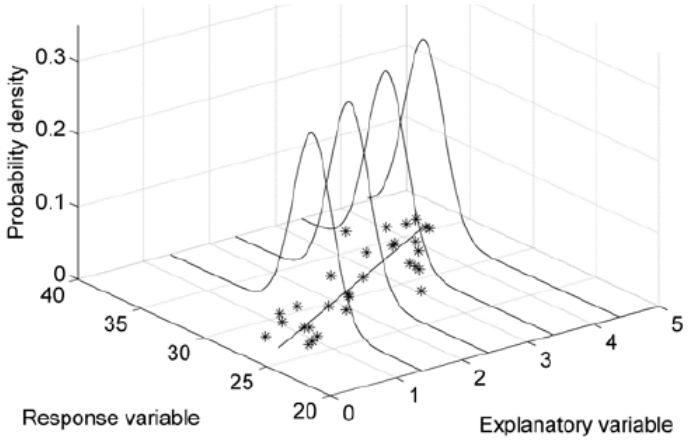
2. **Fit a linear regression model** - A probabilistic model for Y conditional on $X = x$:

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{deterministic}} + \underbrace{E_i}_{\text{random error}} \qquad i = 1, ..., n$$

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. $E_i \sim^{iid} N(0, \sigma^2)$
- $\mu(x) = E(Y|X = x) = E(\beta_0 + \beta_1 x + E) = \beta_0 + \beta_1 x$ (The line describes the mean Y for a given X.)
- $\text{Var}(Y|X = x) = Var(\beta_0 + \beta_1 x + E) = Var(E) = \sigma^2$
- σ^2 - Error variance (variance due to experimental error) (assumed constant for all x - called homoskedasticity)
- $\beta_0 = E(Y|X = 0)$ - $\mu(0) = \beta_0 + \beta_1(0) = \beta_0$ - True population intercept (average value of response when $X = 0$)
- β_1 - True population slope (average change in Y per unit increase in x)

$$\mu(x + 1) - \mu(x) = (\beta_0 + \beta_1(x + 1)) - (\beta_0 + \beta_1 x) = \beta_1$$



We assume there is a true underlying line and we observe points about that line. At every point on the line there is a normal distribution with standard deviation σ. The variation about the line is due to unidentified sources (i.e. experimental error).

Why use the line instead of just correlation?

Can use line to estimate the mean at a given x and also to find a prediction of a new response for a given x.

Parameters to be estimated and make inference on - β_0, β_1 and σ^2.
How do we determine these from the data? Start with β_0 and β_1. Denote the estimates by $\hat{\beta}_0$ and $\hat{\beta}_1$.

Choice of line determined by minimizing sum of squared residuals:

$$min_{\hat{\beta}_0,\hat{\beta}_1} \sum_1^n \Big(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\Big)^2$$

$$= \sum_1^n (y_i - \hat{y}_i)^2 = \sum_1^n e_i^2 = SS(E)$$

Thus, the resulting $\hat{\beta}_0$ and $\hat{\beta}_1$ are called 'least squares' estimates.

Calculus can show that $\hat{\beta}_0$ and $\hat{\beta}_1$ which minimize the sum of squared residuals $SS(E)$ are given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{XY}}{S_{XX}} = \frac{s_{XY}}{s_X^2} = \frac{\text{sample covariance}}{\text{sample variance}} = r_{XY}\frac{s_Y}{s_X}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

Note: what does $\hat{\beta}_1 = r\frac{s_Y}{s_X}$ say about the sign of r and $\hat{\beta}_1$?

The equation using these estimates, $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, has many names such as 'least squares regression line,' 'the fitted line,' 'the prediction line,' and 'the prediction equation'.

$\hat{\beta}_0$ = sample intercept = estimated average value of Y when $X = 0$

$$\mu(0) = \beta_0 + \beta_1 0 = \beta_0$$

$\hat{\beta}_1$ = sample slope = estimated change in y (or average y) for a unit change in x (rise/run)

Definitions:

- **Predicted value** of response Y given $X = x_0$:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

- **Residual** for the i^{th} observation:

$$e_i = y_i - \hat{y}_i = obs - pred$$

The other parameter to estimate is σ^2. An *unbiased* estimate of σ^2 for SLR is given by

$$\hat{\sigma}^2 = \frac{SS(E)}{n-2} = MS(E).$$

In lay terms, being unbiased implies that if we did the experiment hundreds of times, finding this estimate for each experiment, then the average of all the MS(E) values $\approx \sigma^2$.

For the log(Biodiesel) and Biomass example let's find our fitted line. Recall the summary stats on page 66.

$$\hat{\beta}_1 = s_{XY}/s_X^2 = 3.1485/3.9908^2 = 0.1977$$

$$\hat{\beta}_0 = 2.4647 - 11.0523 * 0.1977 = 0.2797$$

$$\hat{y} = 0.2797 + 0.1977x$$

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

How can we make inference (claims about the true values)?

Our main question is again - do we have a *significant linear relationship?*

What value of the slope do we test?

- If no linear relationship, Y won't tend to change with X

$$H_0 : \beta_1 = 0$$

- If a linear relationship, Y will tend to change with X

$$H_A : \beta_1 \neq 0$$

We have a point estimate, $\hat{\beta}_1$, we also need to know about the variability of our estimate.

$$\begin{aligned} Var(\hat{\beta}_1) &= Var\left(\frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) \\ &= Var\left(\frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2 Var(Y_i)}{\left(\sum_{i=1}^n (x_i - \bar{x})^2\right)^2} \text{ since independent Y's, covariance is 0} \\ &= \sigma^2 \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\left(\sum_{i=1}^n (x_i - \bar{x})^2\right)^2} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma^2}{S_{xx}} \end{aligned}$$

Under the normal distribution assumption, the RV $\hat{\beta}_1$ follows normal distribution as it is just a linear combination of normally distributed random variables!

$$\hat{\beta}_1 \sim N(\beta, \sigma^2/S_{xx})$$

What is the standard error of $\hat{\beta}_1$? As we don't know σ^2 we will have to estimate it. What is our estimate for σ^2? What is our estimated standard error of $\hat{\beta}_1$?

$$SE(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{S_{xx}}}$$

$$\hat{\sigma}^2 = MS(E) \implies \hat{SE}(\hat{\beta}_1) = \sqrt{\frac{MS(E)}{S_{xx}}}$$

Recall the one-sample Z and the one-sample t-test for μ. The estimated standard error of $\bar{Y}$ is $\hat{SE}(\bar{Y}) = S/\sqrt{n}$. Our test statistic was

$$Z = \frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1) \qquad or \qquad T = \frac{\bar{Y} - \mu_0}{\hat{SE}(\bar{Y})} = \frac{\bar{Y} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

The same idea applies here! Any hypothetical slope, like $H_0 : \beta_1 = \beta_{1,Null}$ may be tested using the T-statistic below with $df = n - 2$:

$$T = \frac{\hat{\beta}_1 - \beta_{1,Null}}{\widehat{SE}(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_{1,Null}}{\sqrt{MS(E)/S_{xx}}}$$

and a $100(1-\alpha)\%$ confidence interval for β_1 is

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

Note: Tests for the intercept may be performed as well (these will be a special case of testing/intervals for the mean at a given x value). It will depend on the context of the question if testing β_0=0 makes sense.

Locating these tests in the SAS output:

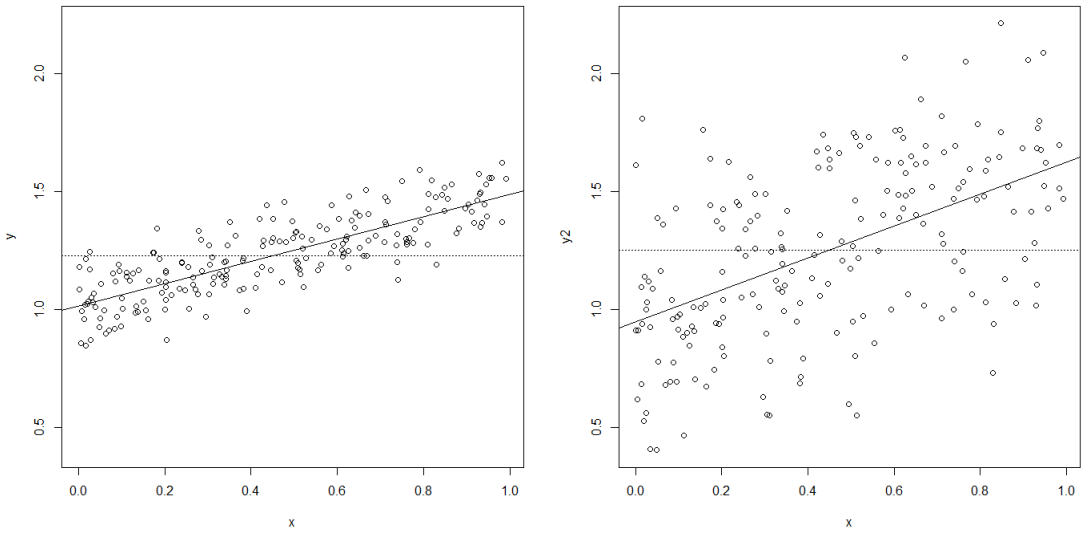
```
proc reg data=bioexp ;
model logbiodiesel=biomass/clb;
run;
```

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

The ANOVA table for SLR

An equivalent way to determine if there is a significant linear regression is to use the ANOVA table.

ANOVA = ANalysis Of VAriance. What variation are we measuring here?



Which line above gives more evidence of a significant linear relationship? Why?
The one on the left because the variability around the line is less and yet the variability from the line to the mean is the same as on the right.

This is the idea we use in the ANOVA table. How much better is the line than just using the mean? Compare that to how variable the data around the line is (and hence the line itself).

Sources of Variability
A measure of the total amount of variability in the response is the sample variance of Y. Consider the numerator, which we call the total sum of squares:

$$\begin{aligned} SS(Tot) &= \sum_{i=1}^{n}(y_i - \bar{y})^2 \text{ (variability of observations about the mean)} \\ df_{Tot} &= n - 1 \end{aligned}$$

This variability is partitioned into independent components:
Sum of squares due to regression, $SS(R)$

$$\begin{aligned} SS(R) &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \text{ (variability of the fitted line about the mean)} \\ &= \sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_i - \hat{\beta}_0 - \hat{\beta}_1 \bar{x})^2 \\ &= \hat{\beta}_1^2 S_{xx} \\ df_R &= 1 \end{aligned}$$

Sum of squares due to error, $SS(E)$

$$\begin{aligned} SS(E) &= \sum e_i^2 \\ &= \sum (y_i - \hat{y}_i)^2 \text{ (variability of observations about the fitted line)} \\ df_E &= n - 2 \end{aligned}$$

Note: $SS(Tot) = SS(R) + SS(E)$ and $df_{Tot} = df_R + df_E$ (Recall: DF = Number of independent pieces of data for the sum of squares).

In total we have n observations (assumed to be independent). For $SS(Tot)$ we have to calculate the mean. Once we have calculated the mean, we only have $n-1$ free observations left. That is, given any $n-1$ of the y_i values and $\bar{y}$, I can find the remaining y_i value. Hence, $n-1$ total df.

For $SS(R)$ we have to find $\hat{\beta}_0$ and $\hat{\beta}_1$. However, given one of these two, I can find the other ($\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$). Therefore, there are $2 - 1 = 1$ df for regression.

This leaves $n - 2$ df for error.

The ANOVA table from simple linear regression

Source	df	Sum of squares	Mean Square	F-Ratio
Regression	1	$SS(R)$	$MS(R)$	$MS(R)/MS(E) \sim^{H_0} F_{1,n-2}$
Error	$n-2$	$SS(E)$	$MS(E)$	
Total	$n-1$	$SS(Tot)$		

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

$H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$

The test statistic is

$$F = MS(R)/MS(E) \overset{H_0}{\sim} F(1, n-2).$$

The rejection region is

$$\{F_{obs} : F_{obs} > F_{\alpha,1,n-2}\}$$

The p-value is

$$P(F_{1,n-2} > F_{obs})$$

This F test is equivalent to the t test we already looked at. The relationship is that $(t_{df})^2 = F_{1,df}$.

Large values of F give evidence that the line is useful. Values near 1 imply that using the mean of the response is just as useful as the line.

How to get the ANOVA table in SAS?

For our Biodiesel and Biomass example we can get 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

Other Inferential Objectives

One reason to consider SLR instead of correlation analysis is that we can now make predictions for the response.

Let's consider a prediction made from this model. If the experiment were run again, would we get the same fitted line? The same predicted mean value?

The predicted mean is itself a RV and so we can find CI for that mean! We can also create a 'Prediction Interval' (PI) for a new x value as well.

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

$$\begin{aligned}
\text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0) &= \text{Var}(\bar{Y} - \bar{x}\hat{\beta}_1 + \hat{\beta}_1 x_0) \\
&= \text{Var}(\bar{Y} - \hat{\beta}_1(\bar{x} - x_0)) \\
&= \sigma^2/n + \sigma^2 \frac{(x_0 - \bar{x})^2}{S_{xx}} \qquad Cov(\bar{Y}, \hat{\beta}_1) = 0 \\
&= \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)
\end{aligned}$$

Again, we estimate σ^2, yielding a CI for $\mu(x_0)$ of

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

Note: We are attempting to capture **the true mean of all responses with an x-value of x_0.

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.

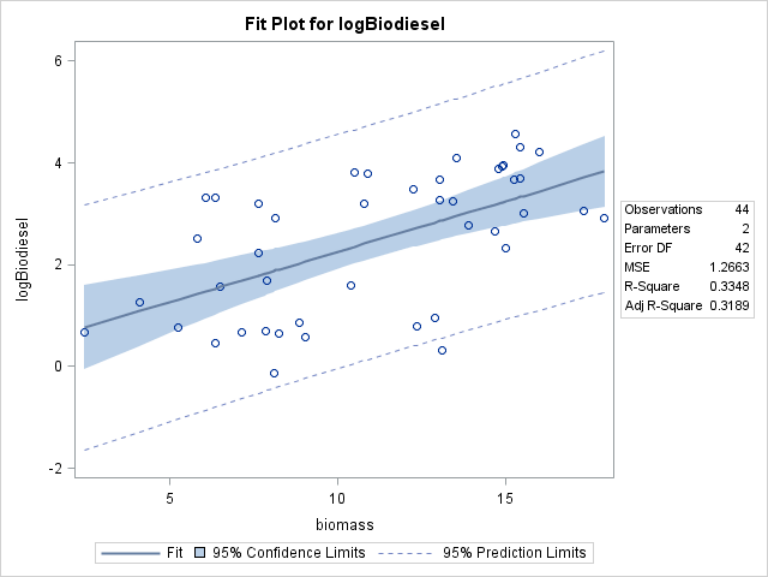
$$\begin{aligned} \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0 + E_{new} | X = x_0) &= \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) + \sigma^2 \qquad \text{As } E_{new} \text{ is independent of past Y's} \\ &= \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) \end{aligned}$$

Thus we can form a PI using

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

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.

For a SLR model, Proc Reg in SAS will also produce a very nice plot that includes *pointwise* confidence and prediction bands at all observed x's.
Notice that the bands get wider the further x_0 is from $\bar{x}$. Why? Most info at $\bar{x}$



A note of caution: We should only use our line to predict for X values within the observed range (interpolating). Extrapolating (predicting for an X outside the range of observed X's) is dangerous as we don't know if the linear pattern exists outside the observed X's.

Example: Consider measuring the heights of 4-16 year old people. There is likely to be a strong positive correlation between height and age. Consider fitting a line and using it to predict height for a 40 year old!

Using $\alpha = 0.05$, ($t_{n-2,\alpha/2} = t_{42,0.025} = 2.0181$) (Use, $SS(R) = \hat{\beta}_1^2 S_{xx} \implies S_{xx} = 26.76509/0.19769^2 = 684.8561$)

1. Find the CI for slope by hand.
2. Form a CI for the mean log of biodiesel when biomass is 12.
3. Form a PI for a future log biodiesel measurement for a biomass of 12.
   Note: In lab you will see an easy way to find the CI and PI intervals at a given x using SAS and the 'missing y' trick!

1. The CI for the slope is

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

   From the output we have $\hat{\beta}_1 = 0.1977$ and $MS(E) = 1.2663$. Thus, we are 95% confident that the true slope between biomass and log(biodiesel) (β_1) is in the interval

$$0.19769 \pm 2.0181\sqrt{1.26632/684.8561} = (0.1109, 0.2845).$$

2. The CI for a mean response is given by

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

   We already have most of the things we need from part (1). Our prediction for a biomass of 12 is

$$0.2798 + 0.1977 * 12 = 2.6522,$$

   the sample size is $n = 44$, and we can find $\bar{x}$= 11.0523 from the correlation output on page 71.
   Thus, we are 95% confident that the true mean log biodiesel for a plant with a biomass of 12 is in the interval

$$2.6522 \pm 2.0181\sqrt{1.2663(1/44 + (12 - 11.0523)^2/684.787)} = (2.3001, 3.0043).$$

3. The PI for a future response is given by

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

   Thus, we are 95% confident that a future log biodiesel measurement for a plant with a biomass of 12 is in the interval

$$2.6522 \pm 2.0181\sqrt{1.2663(1 + 1/44 + (12 - 11.0523)^2/684.787)} = (0.3541, 4.9503).$$

   Note: To make these intervals more meaningful we may want to exponentiate the end points of the intervals to put them on the scale of the original data.

R^2 - the coefficient of determination

Ideally, our model will account for most of the variation in the response.

To look at this, we create the ratio of $SS(R)$ to $SS(TOT)$

$$R^2 = \frac{SS(R)}{SS(Tot)}$$

If R^2 is close to 1, then the model fits the data very well.

R^2 is also called the *coefficient of determination*.

For the log(biodiesel) and biomass example $r^2 = 26.76509/79.95066 = 0.5786^2 = 0.335$. In simple linear regression, 'r-square' is in fact equal to R^2_{XY}. (This isn't the case in multiple regression or other models.)

The interpretation of R^2 usually goes as follows:

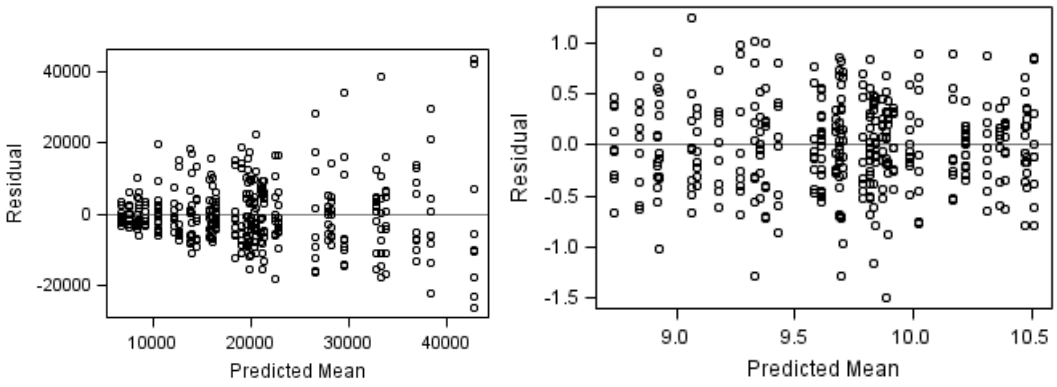
33.5% of variation in log(biodiesel) can be explained by the linear association between log(biodiesel) and biomass.

Checking Model Assumptions in SLR

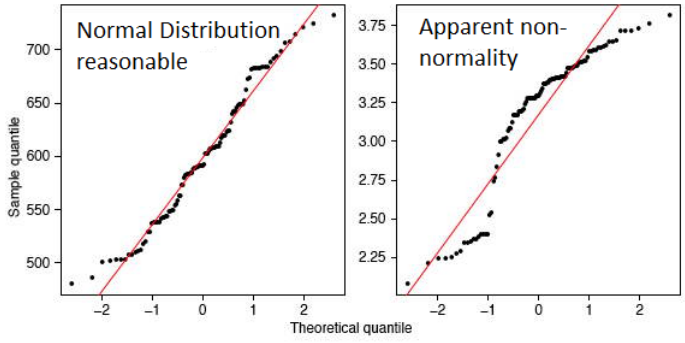
Once we have fit our model and looked at our test we need to consider the validity of our assumptions for the SLR model. Namely, that there is a true linear relationship and our errors are iid $N(0, \sigma^2)$ for all x.

Checking assumptions

1. Inspect a scatter plot to determine if the linear relationship we are assuming in our model is appropriate.
2. Assumption of *iid* $N(0, \sigma^2)$ 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. (Although, if the responses are time dependent we could plot them vs time.)
   - Constant variance - A residuals vs fitted (predicted) values plot or a residual vs independent variable plot are tools for detecting heteroskedasticity (non-constant variance). We hope to see a random scattering of points with no clear trends.



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



More on qq-plots
The p^{th} quantile (or percentile) of a distribution is the value that has p % of the distribution below it.

If we have n observations we can get observed quantiles for the data and create the qq-plot.

1. Sort the data from smallest to largest.
2. The i^{th} ordered data point is the observed $\frac{(i-0.5)}{n}$ quantile (There are other formulas used for this as well.)
3. Find the corresponding theoretical quantiles.
4. Plot the pairs of points.

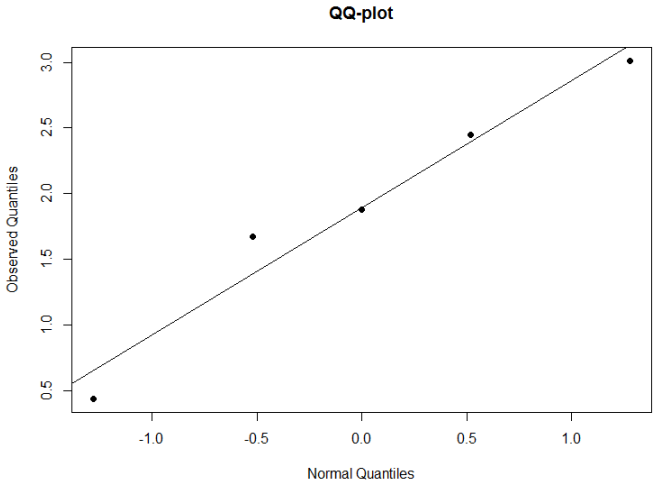
The idea of a qq-plot is to compare the observed quantiles to theoretical ones from a particular distribution (often the normal). If the distribution is a good fit, the observed and theoretical quantiles will be spread out in the same manner and should roughly fall in a line.

Ex: Consider the data set consisting of the following (ordered) observations:

$$0.44, 1.67, 1.88, 2.45, 3.01$$

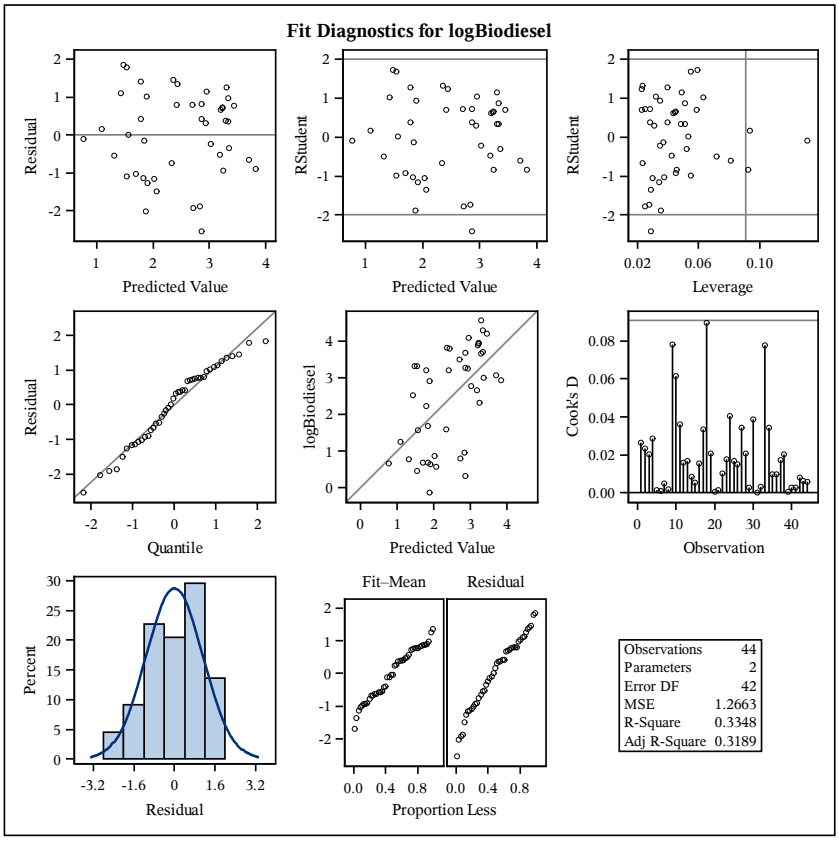
Is the normal distribution a good fit for this data?

Ordered Observed	i	Observed Quantile (i-0.5)/n	Corresponding Normal Quantile
0.44	1	0.1	-1.28
1.67	2	0.3	-0.52
1.88	3	0.5	0.00
2.45	4	0.7	0.52
3.01	5	0.9	1.28
		'y-values'	'x-values'

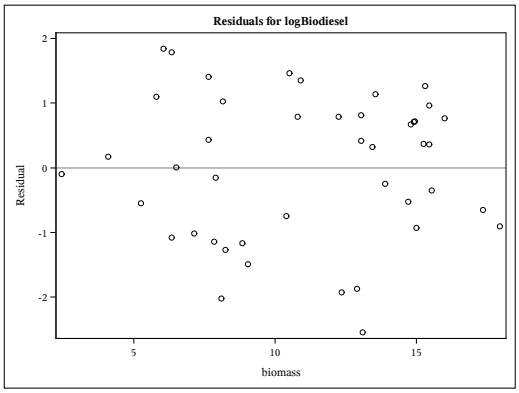


We can inspect the diagnostic plots that SAS produces when the reg procedure is used:

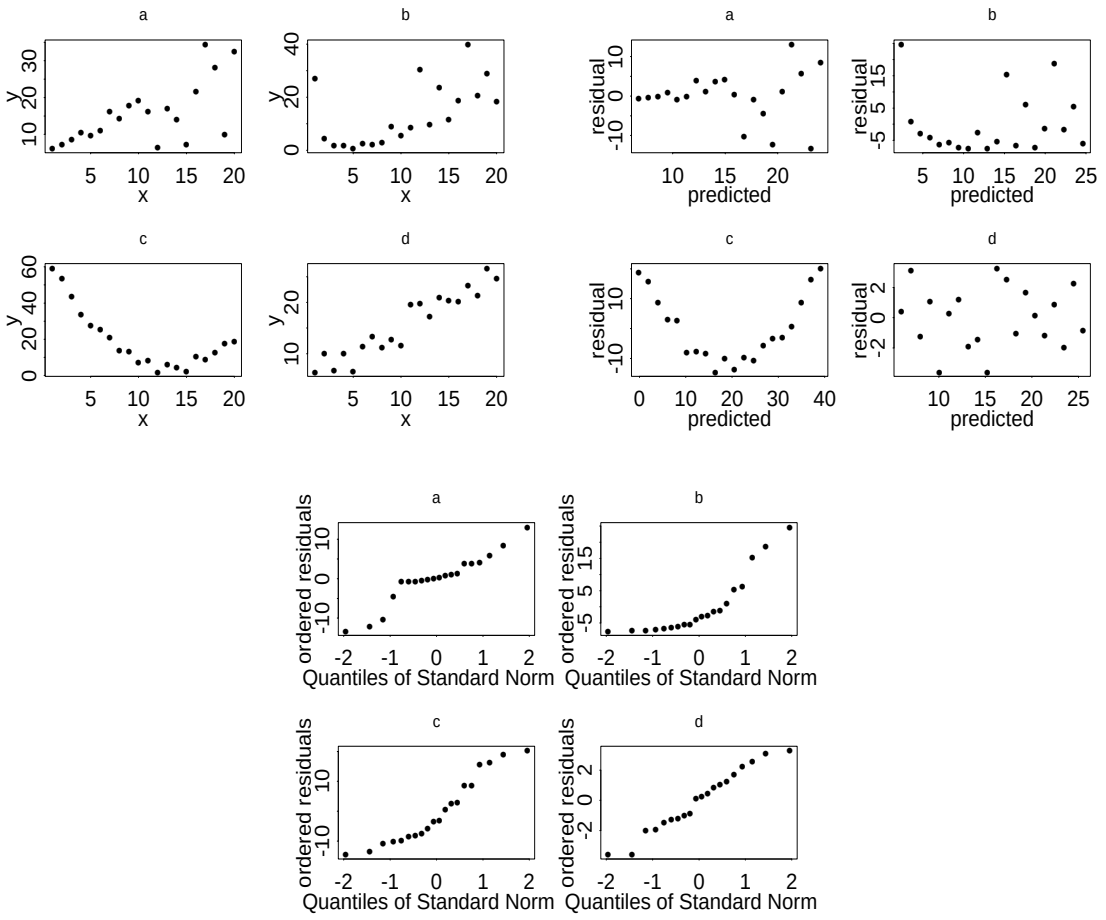
The REG Procedure
Model: MODEL1
Dependent Variable: logBiodiesel



The REG Procedure
Model: MODEL1
Dependent Variable: logBiodiesel



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



a=Heteroskedasticity, b = Nonnormality, c = nonlinearity, d=model fits

What to do when assumptions are violated? - Transformations of the data can often resolve fit issues. (For instance, we have been working with log(biodiesel) for just that reason.) You will take up transformations of data in lab!

This wraps up the discussion of Correlation and SLR. We will now consider expanding the ideas used in SLR to the case when we have multiple quantitative explanatory variables. Then things get really exciting when we use these ideas to then also include categorical explanatory variables in what are called general linear models (GLMs)

Chapter 6

ST 512 - Multiple Linear Regression

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

The majority of time we will have more than 1 predictor variable we want to use to try and model or explain our response variable. We may also want to use only one of our variables but allow for a quadratic or other relationship (curvilinear regression). Multiple Linear regression will give us the tools to fit and evaluate these models. Luckily, most of the ideas for inference and estimation are straight forward extensions of SLR!

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	Label	Aluminum	Label	Iron	Label
4	y_1	13	$x_{1,1}$	61	$x_{1,2}$
18	y_2	21	$x_{2,1}$	175	$x_{2,2}$
14	y_3	24	$x_{3,1}$	111	$x_{3,2}$
18	y_4	23	$x_{4,1}$	124	$x_{4,2}$
26	y_5	64	$x_{5,1}$	130	$x_{5,2}$
26	y_6	38	$x_{6,1}$	173	$x_{6,2}$
21	y_7	33	$x_{7,1}$	169	$x_{7,2}$
30	y_8	61	$x_{8,1}$	169	$x_{8,2}$
28	y_9	39	$x_{9,1}$	160	$x_{9,2}$
36	y_{10}	71	$x_{10,1}$	244	$x_{10,2}$
65	y_{11}	112	$x_{11,1}$	257	$x_{11,2}$
62	y_{12}	88	$x_{12,1}$	333	$x_{12,2}$
40	y_{13}	54	$x_{13,1}$	199	$x_{13,2}$

We could fit a SLR with either Aluminum or Iron separately, but both could have an association with adsorption and in factor may have a combined (**interaction**) effect on adsorption.

Additive (No Interaction) MLR model for two quantitative explanatory variables:

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

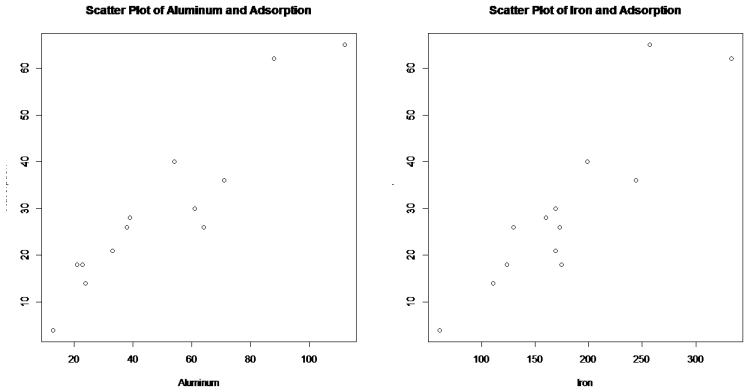
Generally our model is

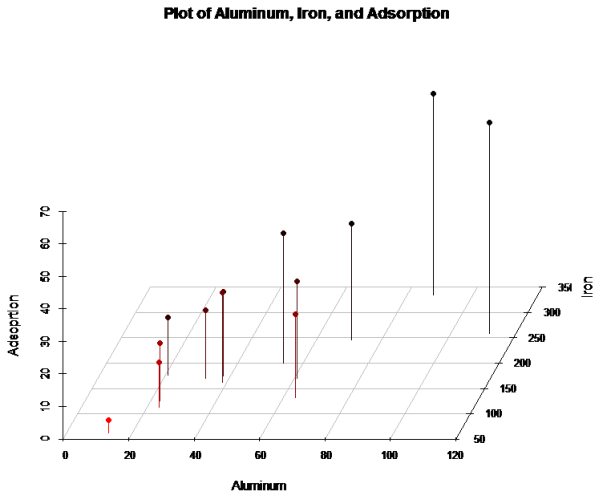
$$Adsorption = \beta_0 + \beta_1 Aluminum + \beta_2 Iron + Experimental\ Error$$

For observation i we write

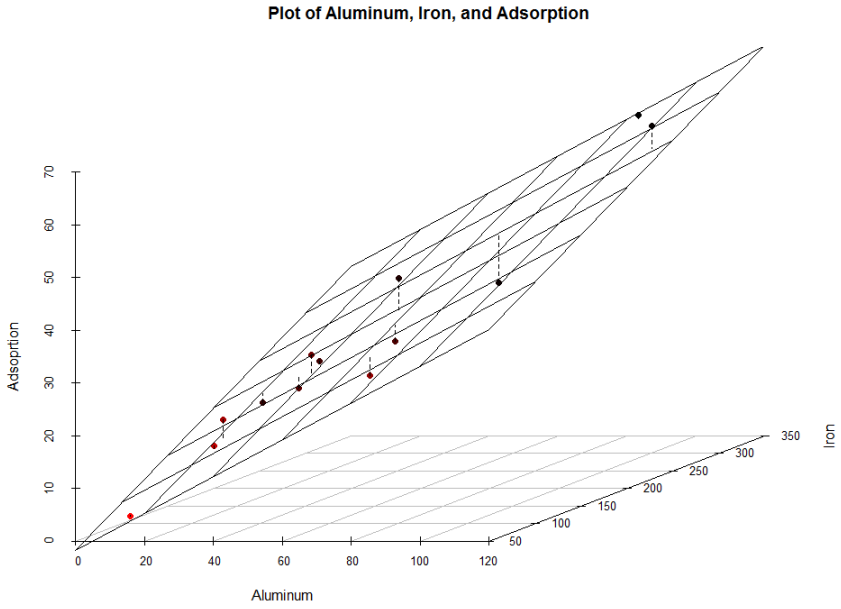
$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + E_i$$

Visualizing the data

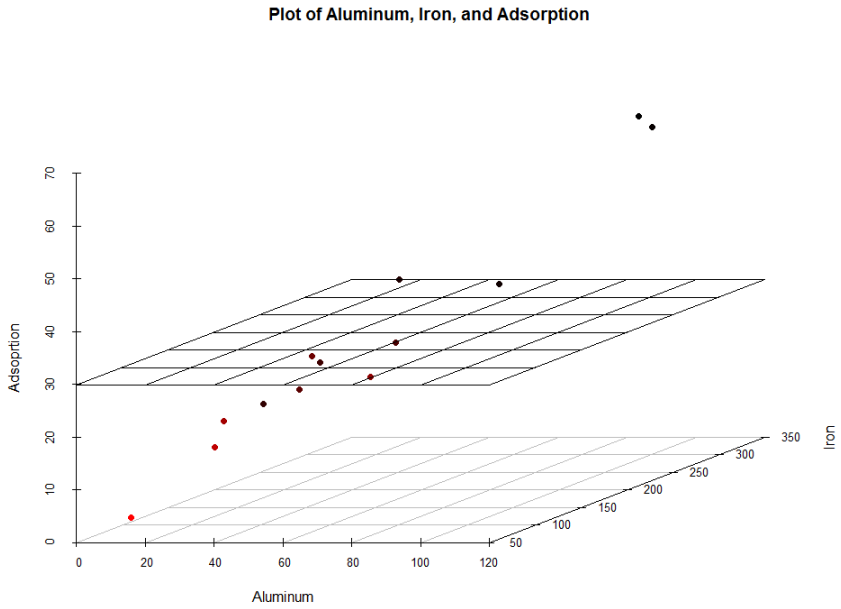




In this case, we don't want to find the best fitting line, but rather the best fitting *plane* (the one the minimizes the squared distances between the plane and the data points).



In SLR we are testing if a 'flat' line is reasonable given the data, with 2 predictors we are testing for a 'flat' plane.



Inferential Objective:

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 (called the global test)

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

This is the hypothesis tested by the ANOVA table p-value.

To fit this model in SAS

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

Output From Proc Reg for Adsorption Example **1**

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

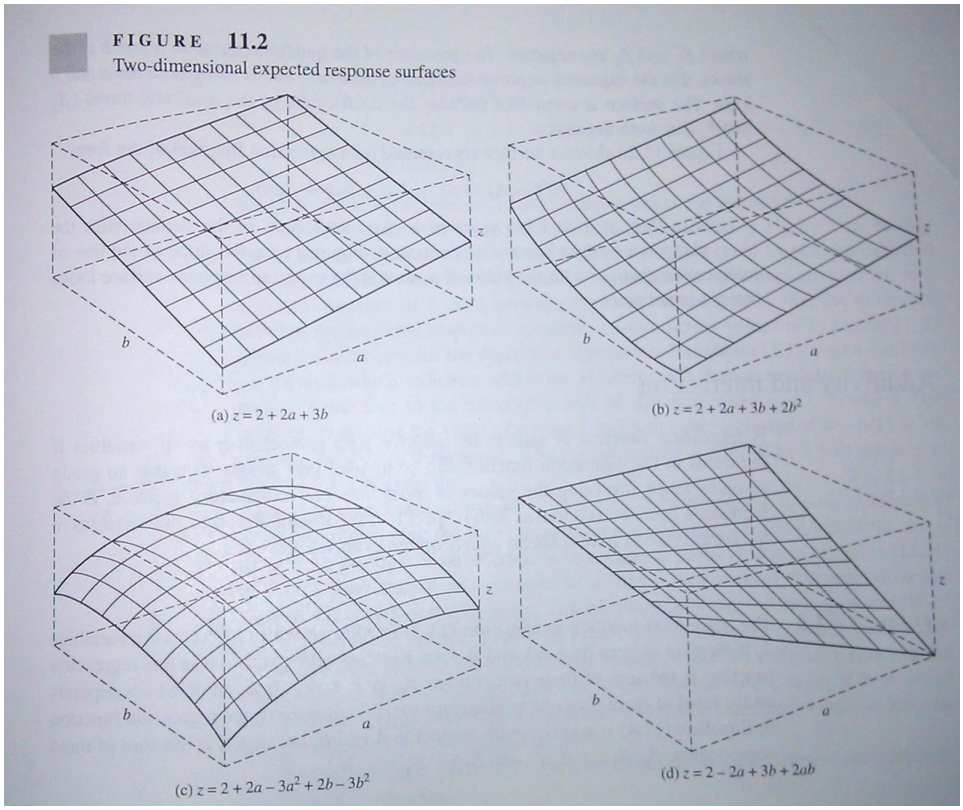
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-7.35066	3.48467	-2.11	0.0611	-15.11498	0.41366
aluminum	1	0.34900	0.07131	4.89	0.0006	0.19012	0.50788
iron	1	0.11273	0.02969	3.80	0.0035	0.04658	0.17889

Fitted model:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 = -7.3507 + 0.3490 x_1 + 0.1127 x_2$$

The surface we fit can be very flexible. All we need to do is include quadratic terms and interaction terms.

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`

These types of models are very important! For instance, if you'd like to optimize your response over two predictors, you might fit a quadratic model in both as in the bottom left plot.

Interaction MLR model for two quantitative explanatory variables:

$$Adsorption = \beta_0 + \beta_1 Aluminum + \beta_2 Iron + \beta_3 Aluminum * Iron + Experimental\ Error$$

For observation i we write

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + E_i$$

(This surface for would be similar to the bottom right plot above.)

Our global test would now be

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \qquad vs \qquad H_A : \text{ at least 1 not zero}$$

(no predictors important vs something in the model is important).

To fit this model in SAS

```
proc glm data=adexp ;
model adsorp=aluminum iron aluminum*iron/solution clparm;
run;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	3542.592385	1180.864128	59.34	<.0001
Error	9	179.099923	19.899991		
Corrected Total	12	3721.692308			

R-Square	Coeff Var	Root MSE	adsorp Mean
0.951877	14.94645	4.460941	29.84615

Source	DF	Type I SS	Mean Square	F Value	Pr > F
aluminum	1	3253.412031	3253.412031	163.49	<.0001
iron	1	276.491053	276.491053	13.89	0.0047
aluminum*iron	1	12.689301	12.689301	0.64	0.4451

Source	DF	Type III SS	Mean Square	F Value	Pr > F
aluminum	1	54.91884691	54.91884691	2.76	0.1310
iron	1	58.76710458	58.76710458	2.95	0.1198
aluminum*iron	1	12.68930127	12.68930127	0.64	0.4451

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	-2.367840436	7.17890749	-0.33	0.7491	-18.60765744	13.871976567
aluminum	0.245970673	0.14806386	1.66	0.1310	-0.088973050	0.580914397
iron	0.082788344	0.04817575	1.72	0.1198	-0.026192776	0.191769465
aluminum*iron	0.000527835	0.00066101	0.80	0.4451	-0.000967466	0.002023136

Note: None of the parameter estimates have significant p-values. We'll talk about this soon.

General MLR model for p predictors $x_1, x_2, \ldots, x_p$ and response Y

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + E_i$$

Assumptions

- The true underlying relationship exists and $\mu_(x_1, ..., x_p) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$.
- Observations observed with random error E_i - where $E_i \sim^{iid} N(0, \sigma^2)$.

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
- with two predictors, we are fitting the best plane in a 3-d space.

Note: The model is written for p predictors but x_2 could really be x_1^2 and x_3 could really be $x_1 x_2$.

ANOVA Table
Generally, the global test is

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

Again, this is the hypothesis tested by the ANOVA table p-value.

The ANOVA table for MLR follows the same ideas as in SLR.

Take the total amount of variation in the response, $SS(Tot)$, and partition it

- into a part due to the model, $SS(R)$
- into 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	df	Sum of squares	Mean Square	F-Ratio
Regression	p	$SS(R)$	$MS(R)$	$MS(R)/MS(E)$
Error	$n-p-1$	$SS(E)$	$MS(E)$	
Total	$n-1$	$SS(Tot)$		

Fitting the Model

Regression parameters (β's) estimated using least squares as in SLR

$$min_{\hat{\beta}'s} SS(E)$$

where

$$SS(E) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip})^2$$

and our estimate for the error variance is similarly defined as

$$\hat{\sigma}^2 = MS(E) = \frac{SS(E)}{df_E} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - p - 1}$$

There are formulas for the estimates of our regression parameters, however, they are cumbersome to write out the way we have been doing.

We will also want to make inference, which implies that we will need to know about the variability of the estimates.

It will be much easier to write all of these things out using matrix notation.

Interpretations of regression parameters:

- σ^2 = unknown **error variance** parameter (measure of variability due to experimental error).
- $\beta_0, \beta_1, \ldots, \beta_p$ are $p+1$ unknown regression parameters:
  - β_0 = average response when $x_1 = x_2 = \ldots = x_p = 0$

$$\mu(x_1, ..., x_p) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \rightarrow \mu(0, ..., 0) = \beta_0$$

  - β_i is called the **partial slope** for x_i.
    - For the additive model linear with unique $x's$, β_i represents the estimated change in mean of Y per unit increase in x_i **with all other independent variables held fixed**.

$$\mu(x_1, ..., x_i + 1, ..., x_p) - \mu(x_1, ..., x_i, ..., x_p) = \beta_i$$

Inference for regression parameters:

Similar to the SLR case, we still use t-tests to test our hypotheses of interest

$$H_0 : \beta_i = 0 \qquad vs \qquad H_A : \beta_i \neq 0$$

Test Statistic:

$$T = \frac{\hat{\beta}_i}{\hat{SE}(\hat{\beta}_i)} \sim^{H_0} t_{n-p-1}$$

To make a conclusion we use

$$RR = \left\{ t_{obs} : |t_{obs}| > t_{\alpha/2, n-p-1} \right\}$$

or

$$p - value = 2 * P(T_{n-p-1} > |t_{obs}|)$$

To actually know about the SE, to find a CI for $\mu(x_1, ..., x_p)$, or to find a PI for a future observation at $x_1, ..., x_p$ we will need to look at the matrix formulation of this model.

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):

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

One of the most important uses of matrices is to represent systems of equations. For instance say we take a subset of the Adsorption data and look at the SLR model between adsorption and aluminum (x):

$$\begin{aligned}
Y_1 &= \beta_0 + \beta_1 * x_1 + E_1 \\
Y_2 &= \beta_0 + \beta_1 * x_2 + E_2 \\
Y_3 &= \beta_0 + \beta_1 * x_3 + E_3
\end{aligned}$$

We can write this in matrix form as

$$\mathbf{y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \qquad \mathbf{e} = \begin{pmatrix} E_1 \\ E_2 \\ E_3 \end{pmatrix}$$

The system is now

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

Recall Some Basic Matrix Operations

1. Dimension of a matrix = # of rows × # of columns
2. Multiplication - requires *conformability* of matrices
   (i.e. for **AB** to be defined - # of columns of A must be the same as the # of rows of B)

   AB matrix has (i, j) element given by 'dot-product' of i^{th} row of **A**, j^{th} column of **B**:

$$\mathbf{AB} = \begin{pmatrix} 7 & 5 \\ 5 & 2 \\ 3 & 2 \end{pmatrix} \begin{pmatrix} 4 & 2 & 1 \\ 3 & 1 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 7\cdot 4+5\cdot 3, & 7\cdot 2+5\cdot 1, & 7\cdot 1+5\cdot 1 \\ 5\cdot 4+2\cdot 3, & 5\cdot 2+2\cdot 1, & 5\cdot 1+2\cdot 1 \\ 3\cdot 4+2\cdot 3, & 3\cdot 2+2\cdot 1, & 3\cdot 1+2\cdot 1 \end{pmatrix}$$

$$= \begin{pmatrix} 43 & 19 & 12 \\ 26 & 12 & 7 \\ 18 & 8 & 5 \end{pmatrix}$$

   Note: Unlike scalar multiplication, the product **DE** is not usually equal to **ED**.

For our system, we have

$$\mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \beta_0 + \beta_1 x_3 \end{pmatrix}$$

Note: A special matrix is called the **Identity Matrix** and is denoted by **I**. It is a *square*, *symmetric*, and *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**: $\mathbf{AI}_2 = \mathbf{A} = \mathbf{I}_3\mathbf{A}$

3. Addition - performed element-wise, matrices must have same *dimension* (same # of rows and columns)

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

Subtraction, same deal

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

For our system we have

$$\mathbf{X}\boldsymbol{\beta} + \mathbf{E} = \begin{pmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \beta_0 + \beta_1 x_3 \end{pmatrix} + \begin{pmatrix} E_1 \\ E_2 \\ E_3 \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_1 x_1 + E_1 \\ \beta_0 + \beta_1 x_2 + E_2 \\ \beta_0 + \beta_1 x_3 + E_3 \end{pmatrix}$$

Thus, we have represented our system of equations using these matrices!

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

Other properties we need:

4. Inverse - The *inverse* $\mathbf{M}^{-1}$ of a *square* $(r \times r)$ matrix $\mathbf{M}$, if it exists, satisfies $\mathbf{M}\mathbf{M}^{-1} = \mathbf{I}_r$ (similar to the reciprocal of real number giving us 1).
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 $\mathbf{C}^{-1}$.

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

Now

$$\mathbf{C}\mathbf{C}^{-1} = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 1/2 & -1/2 \\ 1/2 & 1/2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

5. Transposition - swap rows for columns, columns for rows (turn matrix on its side):

$$t(\mathbf{A}) = \mathbf{A}\text{`} = \mathbf{A}^T = \begin{pmatrix} 7 & 5 & 3 \\ 5 & 2 & 2 \end{pmatrix} \quad \text{“transpose of } \mathbf{A}\text{”}$$

How can we use this matrix representation?

Using our example, say we've observed the first three observations:

$$4 = \beta_0 + \beta_1 13$$

$$18 = \beta_0 + \beta_1 21$$

$$14 = \beta_0 + \beta_1 24$$

Now our goal is to estimate the β parameters here. We can use matrices to represent this scenario!

$$\mathbf{y} = \begin{pmatrix} 4 \\ 18 \\ 14 \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & 13 \\ 1 & 21 \\ 1 & 24 \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

The system is now

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$$

and we want to find the optimal $\boldsymbol{\beta}$ - by optimal the $\boldsymbol{\beta}$ that minimizes

$$SS(E) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

We could optimize $\boldsymbol{\beta}$ via calculus yielding the 'normal equations'

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y}$$

and our estimates written as

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

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

$$E(\mathbf{a}^T\mathbf{Y}) = \mathbf{a}^T E(\mathbf{Y})$$

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

Now, when we want to make inference about a given β element (or a combination of them) we can use Expected Value and Variance properties of this matrix.

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= E((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E(\mathbf{Y}) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta} \end{aligned}$$

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}) &= \text{Var}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Var(\mathbf{Y})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma^2\mathbf{I}_n\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} = \boldsymbol{\Sigma} \end{aligned}$$

The variance involves the true error variance so we will estimate it by

$$\widehat{\boldsymbol{\Sigma}} = \widehat{\text{Var}}(\hat{\boldsymbol{\beta}}) = MS(E)(\mathbf{X}^T\mathbf{X})^{-1}$$

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

General 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**:

$$\mathbf{E} = \begin{pmatrix} E_1 \\ E_2 \\ \vdots \\ E_n \end{pmatrix}$$

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

$$\mathbf{E} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n) \text{ (multivariate normal distribution)}$$

where $\sigma^2\mathbf{I}_n$ is called the variance-covariance matrix of the errors:

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

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).

Later in the semester, we will consider cases (such as block designs and split plots) where this variance-covariance matrix will not be diagonal.

Note: Similarly, we have the variance-covariance matrix of our $\boldsymbol{\beta}$ vector:

$$\boldsymbol{\Sigma} = Var(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \begin{pmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) & Cov(\hat{\beta}_0, \hat{\beta}_2) & \cdots & Cov(\hat{\beta}_0, \hat{\beta}_p) \\ & Var(\hat{\beta}_1) & Cov(\hat{\beta}_1, \hat{\beta}_2) & \cdots & Cov(\hat{\beta}_1, \hat{\beta}_p) \\ & & Var(\hat{\beta}_2) & \cdots & Cov(\hat{\beta}_2, \hat{\beta}_p) \\ & & & \vdots & \vdots \\ & & & & Var(\hat{\beta}_p) \end{pmatrix}$$

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

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

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 13 & 641 & 2305 \\ 641 & 41831 & 133162 \\ 2305 & 133162 & 467669 \end{pmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 0.633138 & 0.002477 & -0.003826 \\ 0.002477 & 0.000265 & -0.000088 \\ -0.003826 & -0.000088 & 0.000046 \end{pmatrix}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{pmatrix} -7.3507 \\ 0.3490 \\ 0.1127 \end{pmatrix}$$

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

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 fitted equation?
   $\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 = -7.3507 + 0.3490 x_1 + 0.1127 x_2$

2. What is the estimate for β_2? What is its interpretation?
   $\hat{\beta}_2 = 0.1127$, which represents the estimated change in adsorption for a one unit increase in extractable iron while holding the amount of extractable aluminum constant.

3. What is the estimated standard error of $\hat{\beta}_2$, $\hat{SE}(\hat{\beta}_2)$?
   $\sqrt{0.00088} = 0.0297$ (square root of (3,3) element of $\widehat{\mathbf{\Sigma}}$)

4. Conduct a test to determine if $\beta_2 = 0$ is plausible (once accounting for X_1). This is a test for the importance of extractable iron once the relationship between extractable aluminum and adsorption index is accounted for (called a type III test, more on this later). Note: $t_{0.025,10} = 2.228$

   $H_0 : \beta_2 = 0 \;\; vs \;\; H_A : \beta_2 \neq 0$, T-statistic: $t = (\hat{\beta}_2 - 0)/SE(\hat{\beta}_2) = 0.1127/0.0297 = 3.795$
   Since our observed test statistic is greater than 2.228, we reject H_0 in favor of H_A, that is, at the 5% significance level, extractable iron has a significant linear association with adsorption (even after accounting for extractable aluminum).

Fitted Values and Residuals

The predicted values can be written as

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

- $\hat{\mathbf{y}}$ 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 $\mathbf{y}$

The residuals as

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

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

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

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \begin{pmatrix} 4.0610 \\ 19.7008 \\ 13.5350 \\ 14.6511 \\ 29.6363 \\ 25.4084 \\ 23.2126 \\ 32.9846 \\ 24.2923 \\ 44.9271 \\ 60.7012 \\ 60.8904 \\ 33.9226 \end{pmatrix} \quad \mathbf{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \begin{pmatrix} 4 \\ 18 \\ 14 \\ 18 \\ 26 \\ 26 \\ 21 \\ 30 \\ 28 \\ 36 \\ 65 \\ 62 \\ 40 \end{pmatrix} - \begin{pmatrix} 4.0610 \\ 19.7008 \\ 13.5350 \\ 14.6511 \\ 29.6363 \\ 25.4084 \\ 23.2126 \\ 32.9846 \\ 24.2923 \\ 44.9271 \\ 60.7012 \\ 60.8904 \\ 33.9226 \end{pmatrix} = \begin{pmatrix} -0.0610 \\ -1.7008 \\ 0.4650 \\ 3.3489 \\ -3.6363 \\ 0.5916 \\ -2.2126 \\ -2.9846 \\ 3.7077 \\ -8.9271 \\ 4.2988 \\ 1.1096 \\ 6.0774 \end{pmatrix}$$

$$SS(E) = \mathbf{e}'\mathbf{e} = 191.7897$$

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

Inference about a mean or a future value

To make inference about some linear combination of $\hat{\beta}$'s such as

$$\hat{\mu}(x_1, x_2, ...x_p) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p$$

we can again use matrix/vector notation to make things easy.

Define

$$\mathbf{a} = \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ \ldots \\ x_p \end{pmatrix}$$

The mean above can be rewritten as

$$(1 \quad x_1 \quad x_2 \quad \ldots \quad x_p)\hat{\boldsymbol{\beta}} = \mathbf{a}^T\hat{\boldsymbol{\beta}}$$

Then we have the point estimate of

$$E(\mathbf{a}^T\hat{\boldsymbol{\beta}}) = \mathbf{a}^T\boldsymbol{\beta}$$

the measure of variability is

$$\text{Var}(\mathbf{a}^T\hat{\boldsymbol{\beta}}) = \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a}$$

estimated as

$$\widehat{\text{Var}}(\mathbf{a}^T\hat{\boldsymbol{\beta}}) = \mathbf{a}^T\hat{\boldsymbol{\Sigma}}\mathbf{a}$$

Now we can use the t-distribution (n-p-1 df) to perform Hypothesis Tests and form Confidence Intervals very similar to the SLR case.

If we want the variance of a future observation, we just add σ^2 to the variance (and thus add $MS(E)$ to the estimated variance).

Note we are making the assumption that $\mathbf{X}^T\mathbf{X}$ matrix has an inverse. This is true if $\mathbf{X}$ is full 'rank'. If this is not true, generalized inverses can be used, but we won't worry about any of that.

1. Estimate the mean adsorption index for soil with extractable aluminum = 100 and extractable iron = 150.
   Unknown population mean: $\theta = \beta_0 + \beta_1(100) + \beta_1(150)$
   Estimate : $\hat{\theta} = (1, 100, 150) * \hat{\beta} = 44.454$

2. Report a standard error for the estimate above.
   To find the standard error, find the variance and take the square root:

$$\text{Var}((1, 100, 150) * \hat{\boldsymbol{\beta}}) = (1, 100, 150)\text{Var}(\hat{\boldsymbol{\beta}})(1, 100, 150)'$$

   estimated as

$$= (1, 100, 150)\widehat{\boldsymbol{\Sigma}}(1, 100, 150)' = 19.832$$

   Which implies $SE(\hat{\theta}) = \sqrt{19.832} = 4.453$.

3. Find and interpret a 95% confidence interval for the population of ALL soil with extractable aluminum = 100 and extractable iron = 150. Note: $t_{0.025,10} = 2.228$
   We are 95% confident that the true mean adsorption index among the population of ALL soil with extractable aluminum = 100 and extractable iron = 150 is between

$$(44.454 - 2.228(4.453), 44.454 - 2.228(4.453)) = (34.533, 54.375)$$

4. Find a 95% prediction interval for a future observation with extractable aluminum = 100 and extractable iron = 150.
   The variance of a future value is $Var(\hat{\mu}(100, 150)) + Var(E_{new})$ which is estimated as 19.832+19.17897=39.01097.
   The estimated SE is the square root = 6.2459.
   Therefore, we are 95% confident that a future absorption index for soil with extractable aluminum = 100 and extractable iron 150 is between

$$(44.454 - 2.228(6.2459), 44.454 - 2.228(6.2459)) = (30.538, 58.370)$$

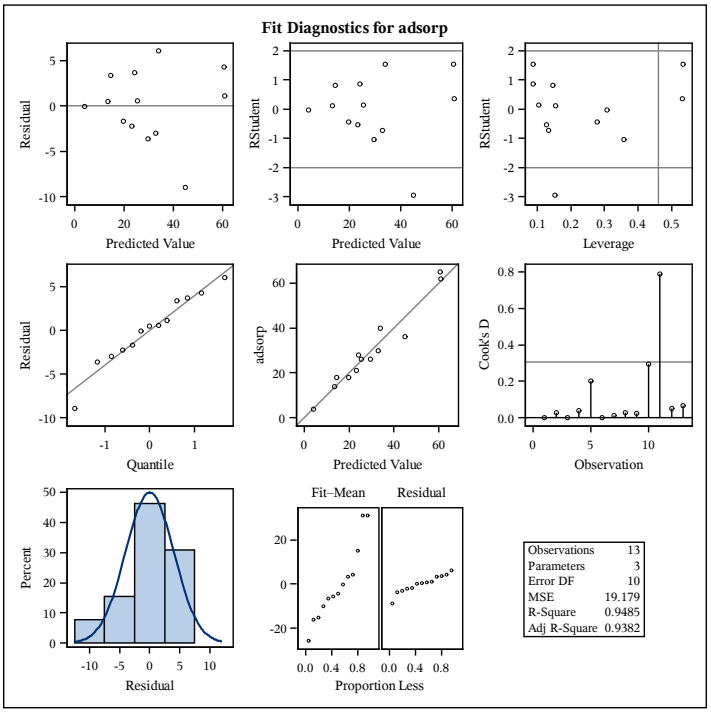
How to get these intervals in SAS?

The following code will produce output appropriate for analysis: (more practice with this in lab)

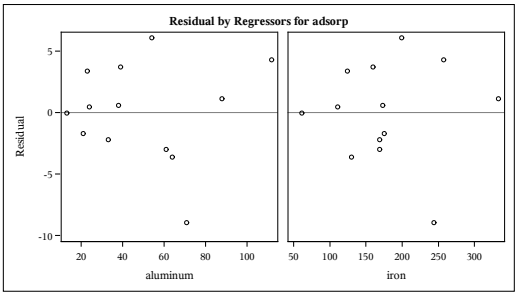
```
data newad;
input adsorp aluminum iron;
datalines;
. 100 150
;

proc datasets;
append base=adexp data=newad;
run;

proc glm data=adexp plots=all;
* can only run one of clm or cli at once;
model adsorp=aluminum iron/solution clparm clm;
*model adsorp=aluminum iron/solution clparm cli;
run;
```

Observation		Observed	Predicted	Residual	95% Confidence Limits for Mean Predicted Value		95% Confidence Limits for Individual Predicted Value	
1		4.00000000	4.06305233	-0.06305233	-1.34831689	9.47442155	-7.09484388	15.22094854
2		18.00000000	19.70660766	-1.70660766	14.56029175	24.85292357	8.67482184	30.73839348
3		14.00000000	13.53870170	0.46129830	9.70330512	17.37409827	3.05414172	24.02326167
4		18.00000000	14.65522935	3.34477065	10.93881551	18.37164320	4.21360769	25.09685101
5		26.00000000	29.64063944	-3.64063944	23.79465905	35.48661983	18.26561339	41.01566550
6		26.00000000	25.41414703	0.58585297	22.28050346	28.54779061	15.16546648	35.66282758
7		21.00000000	23.21821382	-2.21821382	19.74301371	26.69341393	12.85999260	33.57643503
8		30.00000000	32.99022240	-2.99022240	29.42658878	36.55385603	22.60199755	43.37844725
9		28.00000000	24.29761938	3.70238062	21.41005273	27.18518603	14.12148199	34.47375677
10		36.00000000	44.93519447	-8.93519447	41.12614474	48.74424420	34.46024383	55.41014511
11		65.00000000	60.70973500	4.29026500	53.57543490	67.84403510	48.62197388	72.79749612
12		62.00000000	60.90142956	1.09857044	53.78797684	68.01488229	48.82596101	72.97689811
13		40.00000000	33.92920786	6.07079214	31.05872844	36.79968727	23.75790592	44.10050979
14	*	.	44.45930888	.	34.53093041	54.38768735	30.53851760	58.38010016

The REG Procedure
Model: MODEL1
Dependent Variable: adsorp



The REG Procedure
Model: MODEL1
Dependent Variable: adsorp



Model Selection Ideas and Methods

As a researcher, there will often be some models that you want to fit with preselected variables. You may want to leave these variables in regardless of whether or not they significantly improve the model fit.

Other times, you may be unsure of the variables to include and you will want to let the data help you decide. This idea is called **model selection** and there are many, many ways to do this. We will talk about only a few methods.

Consider having three predictors that we'd like to determine the usefulness of: x_1, x_2, x_3

Several models of interest (although we could also include quadratic, interaction terms, etc.)

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. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_2 x_2 + \beta_3 x_3$

Nested Models: - model A is nested in model B implies that model A can be obtained by restricting parameter values (e.g. setting to 0 or setting β's equal) in model B.

True or false:

- Model 1 nested in Model 4 T — Model 1 nested in Model 5 T
- Model 2 nested in Model 4 T — Model 4 nested in Model 1 F
- Model 3 nested in Model 4 T — Model 5 nested in Model 4 T
- Model 3 nested in Model 7 T — Model 1 nested in Model 7 F

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

In comparing two models, suppose

$$\beta_1, \ldots, \beta_q \text{ are in the reduced model } A$$

$$\beta_1, \ldots, \beta_q, \beta_{q+1}, \ldots, \beta_p \text{ are in the full model } B$$

Comparison of models A and B amounts to testing

$$H_0 : \beta_{q+1} = \beta_{q+2} = \ldots = \beta_p = 0 \text{ (model } A \text{ ok)}$$

$$H_1 : \beta_{q+1}, \beta_{q+2}, \ldots, \beta_p \text{ not all 0 (model } B \text{ adds something)}$$

To test this hypothesis we can use the F distribution with $p-q$ numerator df and $n-p-1$ denominator df

$$F = \frac{(SS(E)_r - SS(E)_f)/(p-q)}{MS(E)_f} = \frac{(SS(R)_f - SS(R)_r)/(p-q)}{MS(E)_f}$$

(r and f abbreviate *reduced* and *full*, respectively.)

Difference in the numerator called an **extra regression sum of squares** - sometimes denoted by:

$$R(\beta_{q+1}, \beta_{q+2}, \ldots, \beta_p | \beta_1, \beta_2, \ldots, \beta_q) = SS(R)_f - SS(R)_r.$$

Consider why this test stat makes sense:

$SS(R)_f - SS(R)_r$ can be thought of as the amount of variation in Y (or part of SS(Tot)) that can be attributed to the variables in the alternative hypothesis.

If the variables in the alternative are really meaningful, this should a relatively large quantity compared to MS(E).

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

Consider comparing the simple model that Y depends only on x_1 (triceps) versus the full model that it depends on all three. (Output given on the following page.)

We can construct the F-test for nested models (lack of fit test, LOF):

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

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

(with x_1 in both models)

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

At $\alpha = 0.05$, the critical value is $F(0.05, 2, 16) = 3.63$.

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).

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

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: 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

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		

Test for midarm circumference: - Let's also consider a test that midarm circumference is useful once triceps thickness and thigh circumference are in the model.

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

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

We can construct the F-test for nested models (lack of fit test):

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

or the null hypothesis

$$H_0: \beta_3 = 0 \quad \text{vs} \quad H_1: \beta_3 \neq 0$$

(after accounting for x_1 and x_2.)

$$F = \frac{(396.9 - 385.4)/1}{6.15} = \frac{11.5}{6.15} = 1.87$$

At $\alpha = 0.05$ the critical value is $F(0.05, 1, 16) = 4.49$.

Our conclusion about the hypotheses?
That is, after accounting for the linear dependence between triceps/thigh circumference and bodyfat, there is no evidence for a linear association between mean bodyfat and midarm circumference.

What do we notice about our F statistic calculated and the test for x_3 in the parameter estimates table of the full model?

$$t = -1.37 \qquad t^2 = (-1.37)^2 = 1.8769$$

*****The tests in the parameter estimates table are really equivalent to doing the nested model selection tests removing only that predictor (called type III tests).

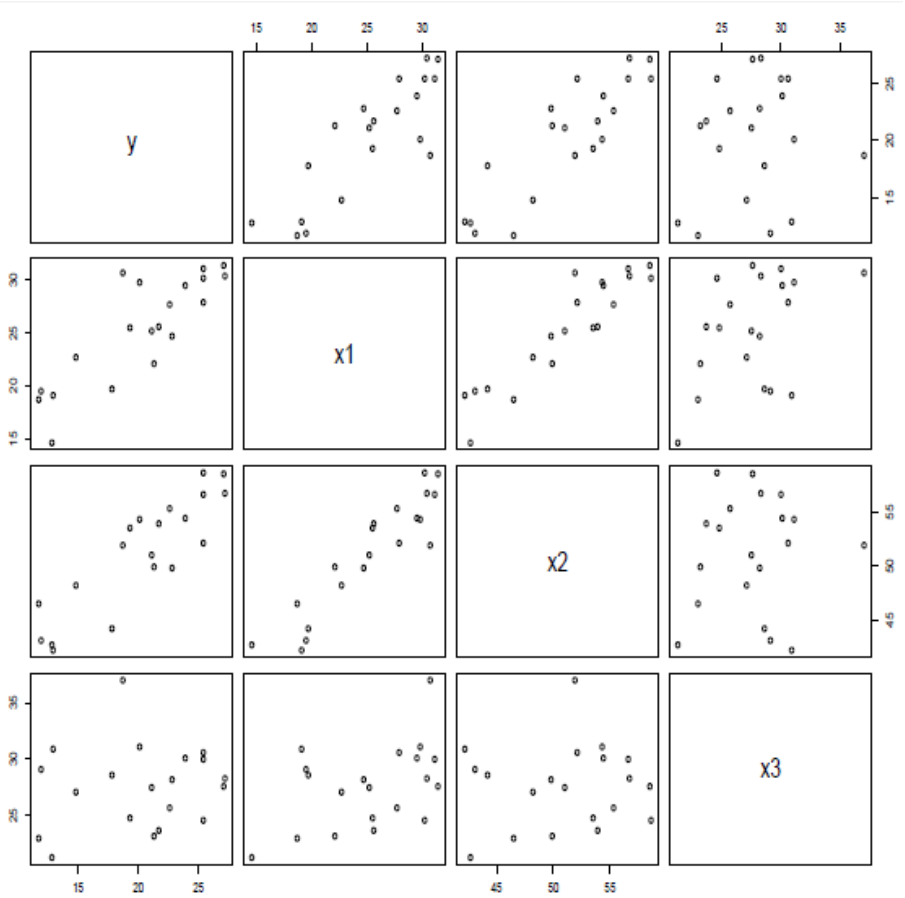
Full model questions
Full model - Global p-value is significant

None of the individual parameters have a significant p-value.

Why?
Tests in parameter estimates tables are type III tests and so they are testing if each variable adds something once the others are all in the model.

Multicollinearity: linear associations among the independent variables; causes problems such as inflated sampling variances for $\hat{\boldsymbol{\beta}}$. (Which makes our t-stats smaller!)

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



Consider all the 'pairwise' correlations:

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

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

Looking at the scatter plots and the correlation output:

- marginal (individual) associations between y and x_1 and between y and x_2 are highly significant
- provides evidence of a strong ($r \approx 0.85$) linear association between average bodyfat and triceps skinfold
- provides evidence of a strong ($r \approx 0.88$) linear association 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 thickness and thigh circumference are giving some of the same information. That is, if you know triceps skinfold thickness is large then you know thigh circumference is large too... This can lead to issues when fitting a model.

In the previous output we saw that a model with all three variables shows that once any two predictors are included, the third predictor does not add anything to the model. This is most likely due to this multicollinearity.

So what model to use?
General suggestion - Use the model that fits best, makes most sense scientifically, and also has the fewest number of predictors.

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).

Hypothesis Testing Methods:

1. **Forward Selection** - Start with nothing and work forward.
   (a) Begin with a model with only β_0
   (b) Fit SLR 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 (x_j) was added in the last step, repeat above process with added predictor. That is, calculate all MLR models with x_j and one more predictor.
   (e) Stop when no new 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.

Fit Criterion Method

3. **Subset Selection** - Compute all possible models, pick 'best' according to a criterion.
   (a) Fit all models under consideration.
   (b) Compare each of the models using a criterion. Choose 'best' model.
   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 and model fit)

Penalization Methods

4. More recent literature has focuses on penalization methods such as the LASSO, Elastic Net, SCAD, etc. We will not discuss these in class.

How to do these model selection methods in SAS?

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

Subset Selection using Mallow's Cp

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

Subset Selection using Adjusted R^2

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

Backward Selection

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			

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

Forward Selection

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

Note: the models selected by each procedure are not necessarily the same. You should bring some subject matter knowledge into play here.

Types of Sums of Squares

If you notice, we now really have multiple tests for a given slope term. A different test for each set of variables already being accounted for. Let's discuss this idea in more detail.

Given that we have 4 predictors, x_1, x_2, x_3, x_4, we really can have a number of tests based on nested models for each parameter.

For instance, we can test

$$H_0 : \beta_4 = 0 \quad vs \quad H_A : \beta_4 \neq 0$$

with

- No other variables in the model (SLR test)
- After accounting for x_1 only
- After accounting for x_2 only
- After accounting for x_1 and x_2
- After accounting for x_1 and x_3
- 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*.

- Type I SS - **sequential** - test for adding the variable after all *previous* variables are accounted for.
  The order the variables are entered into the model determines the tests.
  ***** If you add up the type I SS for all parameters, you get the regression sum of squares (ignore the intercept type I SS if that is given).
- Type II sums of squares - **partial** - test for adding the variable once all other terms not containing a function of that variable are accounted for (i.e. interactions/quadratics/etc). (We won't use these much.)
- Type III sums of squares - **partial** - test for adding the variable after *all* other terms in the model are accounted for.

The tests given for the parameter estimates are all type III tests.

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: (could be done in proc reg as well)

```
proc glm data=bodyfat;
   model y=x1 x3 x2;   *order of variables important for type I SS!;
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

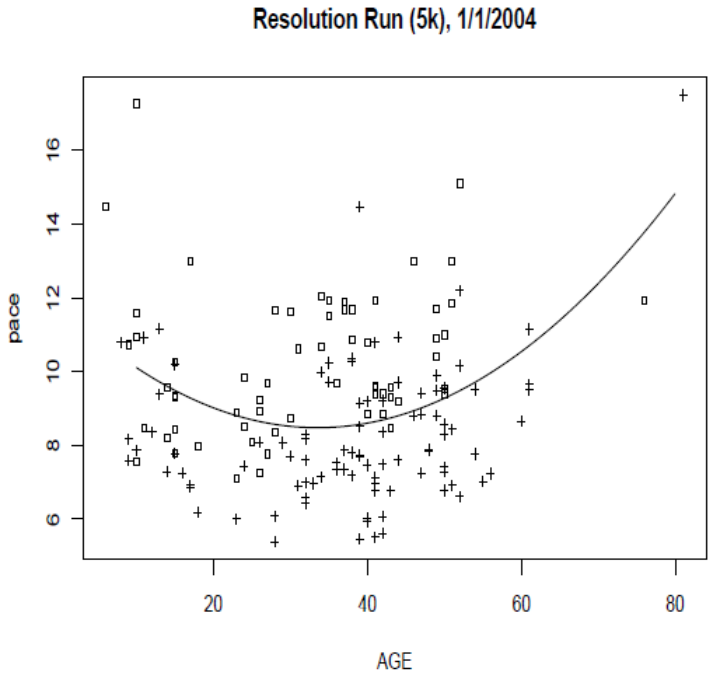
Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	117.0846948	99.78240295	1.17	0.2578
x1	4.3340920	3.01551136	1.44	0.1699
x3	-2.1860603	1.59549900	-1.37	0.1896
x2	-2.8568479	2.58201527	-1.11	0.2849

These tests corresponding to the Type I SS use the *full model* MS(E) rather than the MS(E) for the full model 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.

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

What does σ^2 represent in the model?
The unknown error variance of paces given age x.

What do the parameters mean, i.e. what is their interpretation?
This is a quadratic equation and it is difficult to talk about the parameters separately. You will usually want to discuss the entire curve at once. However, we do know that the β_2 coefficient will shrink or expand the parabola as well as control concavity and the β_1 coefficient will move the parabola around. Also, recall the minimum of the parabola will be at $-\frac{\beta_1}{2\beta_2}$.

We may want to do a LOF test with the SLR model

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

```
proc glm data=running;
   model pace=age age*age ;
run;
```

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

Note with a little work we could get the exact SLR test given above (we'd have to solve for the MS(E) of that model for use in testing).

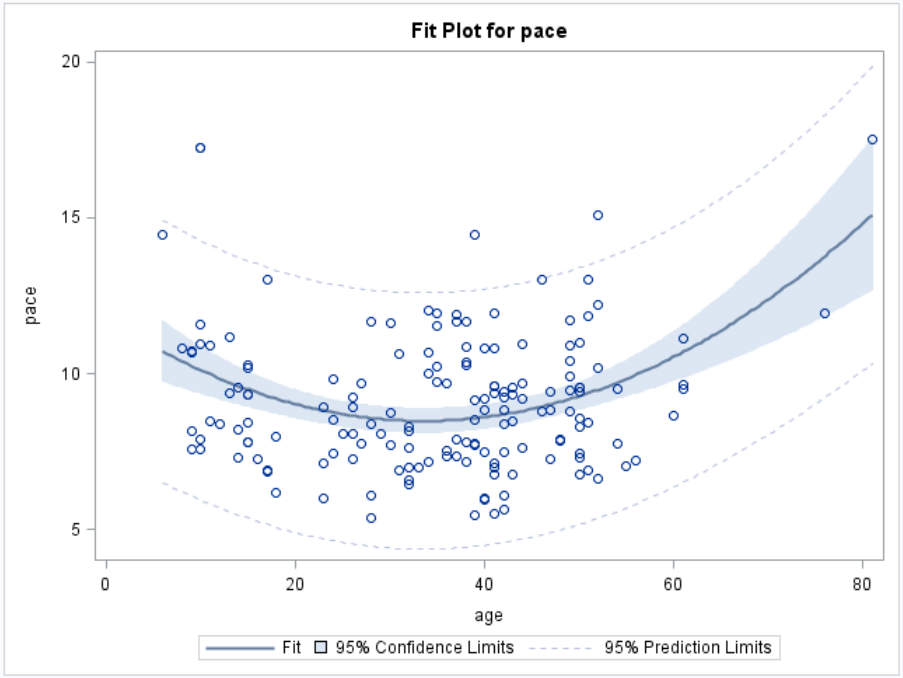
Fitted models are:

$$\begin{array}{ll} \text{Model 1:} & \hat{\mu}(age) = 8.923 + 0.0056age \\ \text{Model 2:} & \hat{\mu}(age) = 11.785 - 0.197age + 0.00294age^2 \end{array}$$

$$\begin{aligned} F &= \frac{(SS(R)_{full} - SS(R)_{red})/1}{MS(E)_{full}} \\ &= \frac{(113.6 - 1.1)/1}{4.3} \\ &= \frac{(SS(E)_{red} - SS(E)_{full})/1}{MS(E)_{full}} \\ &= \frac{(787.0 - 674.4)/1}{4.3} \\ &= 26.2 \end{aligned}$$

Note: $\left(\frac{\hat{\beta}_2}{\hat{SE}}\right)^2 = (5.12)^2$

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 age*age!



Note: The estimate of the minimum average pace is $-\hat{\beta}_1/(2\hat{\beta}_2) = -(-0.197)/(2 * 0.003) = 32.83$.

We may want to make inference for the mean pace time of all individuals that are 32.83 years old.

We can use the same ideas as before. We really want to make inference for

$$\hat{\mu}(32.83) = \hat{\beta}_0 + \hat{\beta}_1 32.83 + \hat{\beta}_2 32.83^2$$

which can be written in vector (linear combination) form as

$$\mathbf{a}^T\hat{\boldsymbol{\beta}}$$

where

$$\mathbf{a}^T = (1 \quad 32.83 \quad 32.83^2)$$

We can get the $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ matrix by adding in 'inverse' to the model statement

```
model pace=age age*age/inverse ;
```

X'X Inverse Matrix

	Intercept	age	age*age	pace
Intercept	0.1147677219	-0.006172105	0.0000746932	11.78503486
age	-0.006172105	0.0003938825	-5.287862E-6	-0.196993007
age*age	0.0000746932	-5.287862E-6	7.6641472E-8	0.0029369853
pace	11.78503486	-0.196993007	0.0029369853	674.44971502

A 95% confidence interval for the mean pace time is given by

$$\mathbf{a}^T\hat{\boldsymbol{\beta}} \pm t_{157,0.025}\sqrt{\mathbf{a}^T\hat{\boldsymbol{\Sigma}}\mathbf{a}} \qquad \textit{or} \qquad \mathbf{a}^T\hat{\boldsymbol{\beta}} \pm t_{157,0.025}\sqrt{MS(E)\mathbf{a}^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{a}}$$

$$(1 \quad 32.83 \quad 32.83^2)\begin{pmatrix} 11.785 \\ -0.197 \\ 0.003 \end{pmatrix} \pm$$

$$1.975\sqrt{4.296(1 \quad 32.83 \quad 32.83^2)\begin{pmatrix} 0.115 & -0.006 & 7.469E-5 \\ -0.006 & 3.94E-4 & -5.29E-6 \\ 7.47E-5 & -5.29E-6 & 7.66E-8 \end{pmatrix}\begin{pmatrix} 1 \\ 32.83 \\ 32.83^2 \end{pmatrix}}$$

We are 95% confident that the mean pace time for people that are 32.83 years old is between 8.077 and 8.890 minutes.

Note: Residual diagnostic shows slight non-normality.

An Interaction Example

A random sample of students taking the same exam yielded:

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

We could consider an *additive* model

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

$$Grade = \beta_0 + \beta_1 IQ + \beta_2 Time + Error$$

However, the grade on the exam might depend on how IQ and Study Time together. We may want to investigate this **interaction** effect - IQ*Time.

Recall: Interaction implies that the way one variable effects the response depends on the level of the other variable.

Here, what does that mean?

The effect IQ has on your Grade differs depending on your Study Time. Or equivalently, the effect Study Time has on Grade differs depending on your IQ.

Consider the *interaction* model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + E_i$$

$$Grade = \beta_0 + \beta_1 IQ + \beta_2 Time + \beta_3 Time * IQ + Error$$

The following models were fit in SAS:

```
proc glm data=test;
model Grade=IQ|Time;
run;
```

```
proc glm data=test;
model Grade=Time|IQ;
run;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	610.8103257	203.6034419	26.22	0.0043
Error	4	31.0646743	7.7661686		
Corrected Total	7	641.8750000			

R-Square	Coeff Var	Root MSE	GRADE Mean
0.951603	3.424620	2.786785	81.37500

Source	DF	Type I SS	Mean Square	F Value	Pr > F
IQ	1	15.9392992	15.9392992	2.05	0.2252
TIME	1	580.1758161	580.1758161	74.71	0.0010
IQ*TIME	1	14.6952103	14.6952103	1.89	0.2410

Source	DF	Type III SS	Mean Square	F Value	Pr > F
IQ	1	0.64458880	0.64458880	0.08	0.7876
TIME	1	6.41230324	6.41230324	0.83	0.4149
IQ*TIME	1	14.69521035	14.69521035	1.89	0.2410

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	72.20607616	54.07277588	1.34	0.2527
IQ	-0.13117020	0.45529954	-0.29	0.7876
TIME	-4.11107221	4.52430095	-0.91	0.4149
IQ*TIME	0.05307053	0.03858059	1.38	0.2410

Source	DF	Type I SS	Mean Square	F Value	Pr > F
TIME	1	474.8767361	474.8767361	61.15	0.0014
IQ	1	121.2383792	121.2383792	15.61	0.0168
TIME*IQ	1	14.6952103	14.6952103	1.89	0.2410

Notice that our explanatory variables do not have an independent effect on the response.

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

Now if IQ = 100 we get

$$\widehat{MeanGrade} = (72.21 - 13.1) + (-4.11 + 5.31) * Time = 59.11 + 1.20 * Time$$

and if IQ = 120 we get

$$\widehat{MeanGrade} = (72.21 - 15.7) + (-4.11 + 6.37) * Time = 56.51 + 2.26 * Time$$

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
- $\beta_3 x_2$ - 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 $\beta_3 x_2$ is the amount the slope for x_1 changes per unit change in x_1 for a fixed constant value of x_2.

Note: You may want to center your predictors to make the parameter interpretations more meaningful.

The global p-value is significant, but none of our individual terms are. We can use the type I Sums of Squares to do model building!! What model would you choose and why?

Polynomial Contrasts (Optional Reading)

Often in One-Way ANOVA with a single factor, the factor is not categorical - rather it is numeric but observed at only a few levels.

If this is the case we can test for polynomial relationships as opposed to the One-Way ANOVA model.

With t levels, we can fit any polynomial of degree $t-1$ or less. (Polynomial of degree $t-1$ is equivalent to fitting the One-Way 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 numeric but measured at only 4 levels.

Consider the One-way ANOVA model using diet and the Linear Regression model cubic in protein:

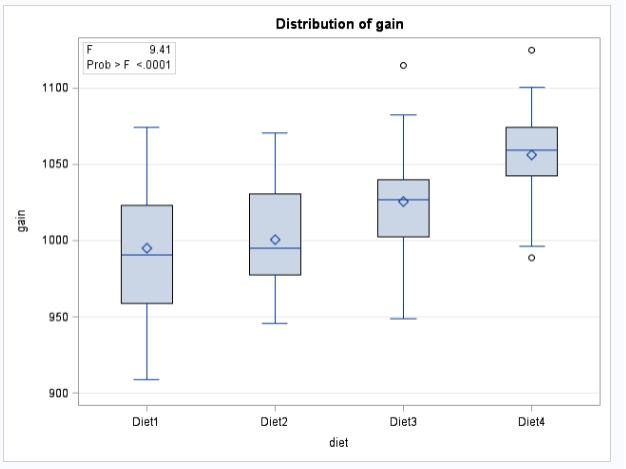
```
proc glm data=chickens; class diet;
model gain=diet; run;

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



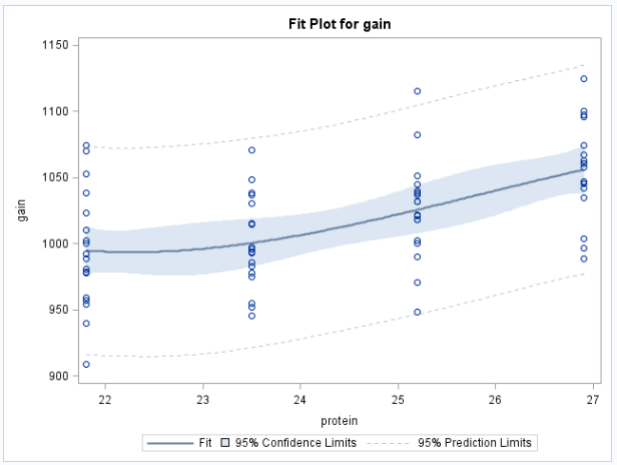
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	42020.4921	14006.8307	9.41	<.0001
Error	68	101214.9702	1488.4554		
Corrected Total	71	143235.4623			

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.

$$SS(\text{lack-of-fit}) = 42020.49 - (39129.40 + 2700.51) = 190.58$$
$$MS(E)_{full} = 1488.46$$

This implies

$$F = \frac{SS(\text{lack of fit})/(4-1-2)}{MS(E)_{full}} = 190.58/1488.46 = 0.128$$

Fail to reject that the full ANOVA model is necessary. That is, the quadratic model does not suffer from lack of fit.

Chapter 7

ST 512 - General Linear Models

Readings: 12.1-12.6 pg 540 - 583

Now that we've seen how to include multiple quantitative predictors, the question becomes how do we include categorical variables such as gender? If we just wanted to compare race times based on the genders only, what model would we employ?

Two-sample t-test or more generally an ANOVA model.

The next topic we'll cover is called the general linear model or GLM. GLMs will allow us to employ both categorical and quantitative predictors simultaneously to model our response.

- (factorial effects) Multi-way ANOVA is useful when we have a quantitative response and qualitative explanatory predictors
- MLR is useful when we have a quantitative response and quantitative explanatory predictors
- GLMs are useful when we have a quantitative response and both categorical and quantitative predictors

To understand how GLMs incorporate qualitative variables we will need to see how to write the one-way ANOVA model in terms of a regression model. This will lead to a straightforward way to include both types of predictors. We will start with a very brief review of one-way ANOVA.

How to write qualitative variables in a GLM format?

One-Way ANOVA 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
- $j = 1, ..., n$ represents the number of observations we have in each treatment group.

We will consider 'balanced' designs only.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)

Other form of the One-Way ANOVA model is

$$Y_{ij} = \mu_i + E_{ij}$$

Unknown parameters:

- μ_i - treatment mean for population i
- 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 each other.

Recall Binding Fraction of Antibiotic One-Way ANOVA example:
An experiment was done to determine if there was a difference between antibiotic types in terms of their mean 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.

Antibiotic	True Trt Mean	Sample Mean
Chloramphenicol	$\mu_1 = \mu + \tau_1$	$\bar{y}_{1\bullet} = 27.8$
Erythromycin	$\mu_2 = \mu + \tau_2$	$\bar{y}_{2\bullet} = 19.1$
Penicillin G	$\mu_3 = \mu + \tau_3$	$\bar{y}_{3\bullet} = 28.6$
Streptomycin	$\mu_4 = \mu + \tau_4$	$\bar{y}_{4\bullet} = 7.8$
Tetracyclin	$\mu_5 = \mu + \tau_5$	$\bar{y}_{5\bullet} = 31.4$

Writing categorical predictors as dummy variables

Can do all of this using the linear model framework we did in the MLR chapter.

The ANOVA model can be fit using MLR with 4 **indicator variables** (or dummy variables) $x_1, \ldots, x_4$ for the 5 antibiotics. Let $i = 1, ..., n$ where n is now the total number of observations again. For observation i

$$x_{i1} = \begin{cases} 1 & \text{if treatment 1} \\ 0 & \text{else} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if treatment 2} \\ 0 & \text{else} \end{cases}$$

$$x_{i3} = \begin{cases} 1 & \text{if treatment 3} \\ 0 & \text{else} \end{cases}$$

$$x_{i4} = \begin{cases} 1 & \text{if treatment 4} \\ 0 & \text{else} \end{cases}$$

The GLM model is then

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + E_i \quad i = 1, \ldots, 20$$

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

Note: The x's here are not quantitative variables, but are numerically labeled nominal variables. Hence 'dummy variables'.

What do the parameters represent in this model?

GLM model	ANOVA model
β_0	$\mu_5 = \mu + \tau_5$
β_1	$\mu_1 - \mu_5 = (\mu + \tau_1) - (\mu + \tau_5) = \tau_1 - \tau_5$
β_2	$\mu_2 - \mu_5 = \tau_2 - \tau_5$
β_3	$\mu_3 - \mu_5 = \tau_3 - \tau_5$
β_4	$\mu_4 - \mu_5 = \tau_4 - \tau_5$

The last treatment is being used as a *reference treatment*. Any of the levels could be used as the reference, SAS uses the last (alphabetically).

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

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

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

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

Estimates of the β's still found by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{pmatrix} 31.375 \\ -3.575 \\ -12.300 \\ -2.775 \\ -23.550 \end{pmatrix}$$

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

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

$$\begin{aligned}
\text{Trt 1 estimate} = \hat{\mu}(1,0,0,0) &= \hat{\beta}_0 + \hat{\beta}_1 = 27.800 \\
\text{Trt 2 estimate} = \hat{\mu}(0,1,0,0) &= \hat{\beta}_0 + \hat{\beta}_2 = 19.075 \\
\text{Trt 3 estimate} = \hat{\mu}(0,0,1,0) &= \hat{\beta}_0 + \hat{\beta}_3 = 28.600 \\
\text{Trt 4 estimate} = \hat{\mu}(0,0,0,1) &= \hat{\beta}_0 + \hat{\beta}_4 = \ 7.825 \\
\text{Trt 5 estimate} = \hat{\mu}(0,0,0,0) &= \hat{\beta}_0 \qquad\quad = 31.375
\end{aligned}$$

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

$$\hat{\boldsymbol{\Sigma}} = MS(E)(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 2.264 & -2.264 & -2.264 & -2.264 & -2.264 \\ & 4.527 & 2.264 & 2.264 & 2.264 \\ & & 4.527 & 2.264 & 2.264 \\ & & & 4.527 & 2.264 \\ & & & & 4.527 \end{pmatrix}$$

Note the pattern, what is the reason for it?

The intercept represents a treatment mean, whereas all the other β's represent differences between treatment means.

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

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

Then

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

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

$$\mathbf{a}^T\hat{\mathbf{\Sigma}}\mathbf{a} = \mathbf{b}^T\hat{\mathbf{\Sigma}}\mathbf{b} = \mathbf{c}^T\hat{\mathbf{\Sigma}}\mathbf{c} = \mathbf{d}^T\hat{\mathbf{\Sigma}}\mathbf{d} = \mathbf{f}^T\hat{\mathbf{\Sigma}}\mathbf{f} = \widehat{\mathrm{Var}}(\hat{\beta}_0) = \widehat{\mathrm{Var}}(\hat{\beta}_0 + \hat{\beta}_j) = 2.264$$

Note:

$\widehat{\mathrm{Var}}(\hat{\beta}_0+\hat{\beta}_j) = \widehat{\mathrm{Var}}(\hat{\beta}_0)+\widehat{\mathrm{Var}}(\hat{\beta}_j)+2\widehat{Cov}(\hat{\beta}_0,\hat{\beta}_j) = 2.264+4.527+2*(-2.264) = 2.263$ $(rounding)$

so the estimated SE for any sample treatment mean is

$$\sqrt{2.264} = 1.505$$

Recall from one-way ANOVA that

$$\widehat{SE}(\bar{Y}_{i\bullet}) = \sqrt{\frac{MS(E)}{n}} = \sqrt{\frac{9.055}{4}} = \sqrt{2.264} = 1.505$$

and that for differences between treatment means

$$\widehat{SE}(\bar{Y}_{i\bullet} - \bar{Y}_{j\bullet}) = \sqrt{\frac{2MS(E)}{n}} = \sqrt{4.528} = 2.128$$

which is $\sqrt{4.527}$!

Using dummy variables, we can now easily include 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 $\mu(x_1) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$ to predict the mean pace for runners. Consider modeling gender (a categorical variable with 2 levels, implying we need 2-1=1 dummy 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:

   $$\mu(x_1, x_3) = \beta_0$$

   Just use the overall mean pace to predict for any age and any gender.

2. The One-Way ANOVA model in GLM form:

   $$\mu(x_1, x_3) = \beta_0 + \beta_3 x_3$$

   A model that uses the mean pace for men to predict for all men and the mean pace for women to predict for all women.

3. The SLR model using Age:

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

   Pace has a linear relationship with age, same relationship for both men and women.

4. The GLM model using Age and Gender:

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

   Pace has a linear relationship with age, different intercept for men and women.

5. The GLM model using Age, Gender, and their Interaction:

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

   Pace has a linear relationship with age, different intercept and slope for men and women.

6. The MLR model quadratic in Age:

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

Pace has a quadratic relationship with age, same parabola for both men and women.

7. The GLM model quadratic in Age with a gender main effect:

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

Pace has a quadratic relationship with age, different intercepts for the men and women parabolas.
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: $(\beta_0 + \beta_3) + \beta_1 x_1 + \beta_2 x_1^2$

8. A GLM model quadratic in Age with all gender interactions:

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

Pace has a quadratic relationship with age, different intercepts and different shapes of the parabolas.
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)

```
*Note, you would need to run a different proc glm for each model;
proc glm;
title 'Model 2'; class sex; model pace=sex/solution clparm;
title 'Model 3';            model pace=age/solution clparm;
title 'Model 4'; class sex; model pace=age sex/solution clparm;
title 'Model 5'; class sex; model pace=age sex age*sex/solution clparm;
title 'Model 6';            model pace=age age*age/solution clparm;
title 'Model 7'; class sex; model pace=age age*age sex/solution clparm;
title 'Model 8'; class sex; model pace=age age*age sex sex*age sex*age*age/solution clparm;
run;
```

Model 2 11:08 Monday, January 12, 2015 **2**

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

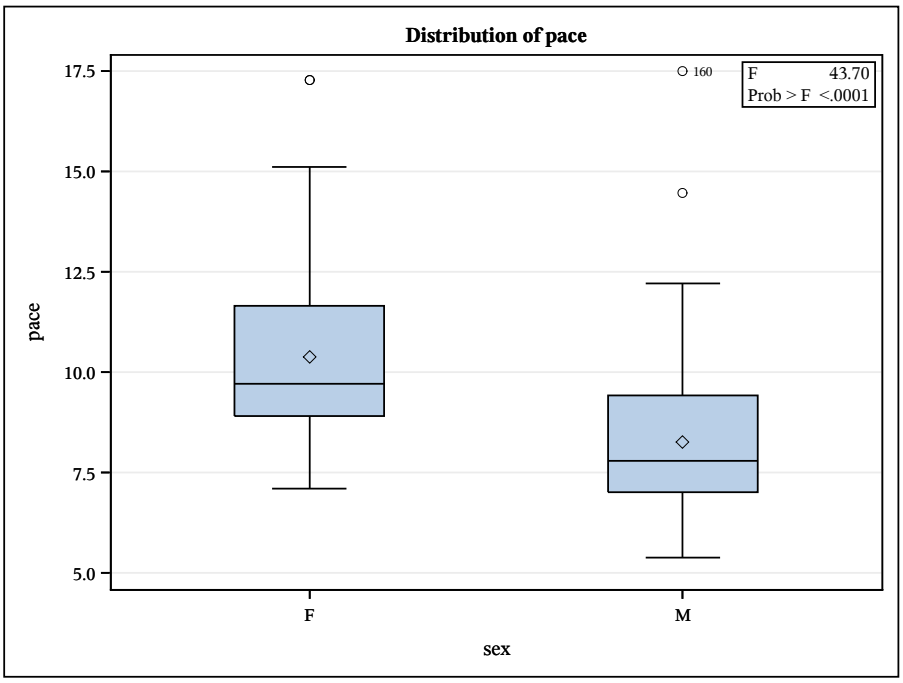
Source	DF	Type III SS	Mean Square	F Value	Pr > F
sex	1	170.7413698	170.7413698	43.70	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	8.266140351	B	0.20280402	40.76	<.0001	7.865583741	8.666696961
sex F	2.103346829	B	0.31818512	6.61	<.0001	1.474901915	2.731791743
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

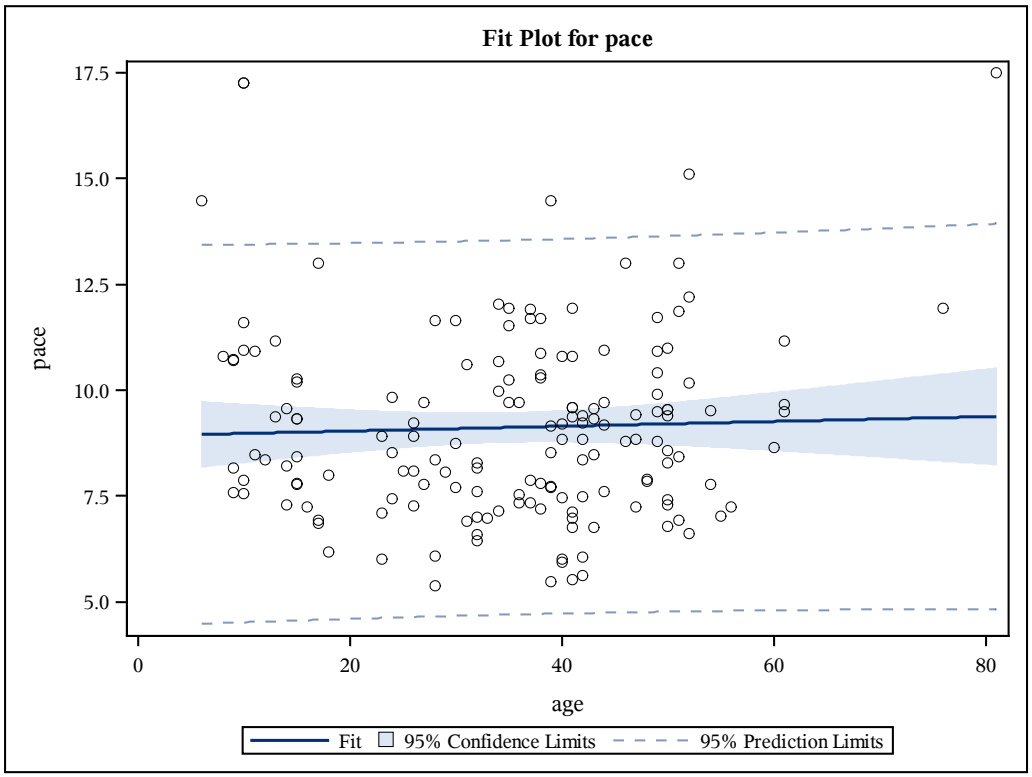
Source	DF	Type I SS	Mean Square	F Value	Pr > F
age	1	1.09650427	1.09650427	0.22	0.6396

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|	95% Confidence Limits	
Intercept	8.922709126	0.45724042	19.51	<.0001	8.019617196	9.825801055
age	0.005643654	0.01202856	0.47	0.6396	-0.018113853	0.029401160

The GLM Procedure

Dependent Variable: pace



The GLM Procedure

Dependent Variable: pace

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	187.6011336	93.8005668	24.52	<.0001
Error	157	600.4935816	3.8247999		
Corrected Total	159	788.0947153			

R-Square	Coeff Var	Root MSE	pace Mean
0.238044	21.44271	1.955710	9.120625

Source	DF	Type I SS	Mean Square	F Value	Pr > F
age	1	1.0965043	1.0965043	0.29	0.5931
sex	1	186.5046294	186.5046294	48.76	<.0001

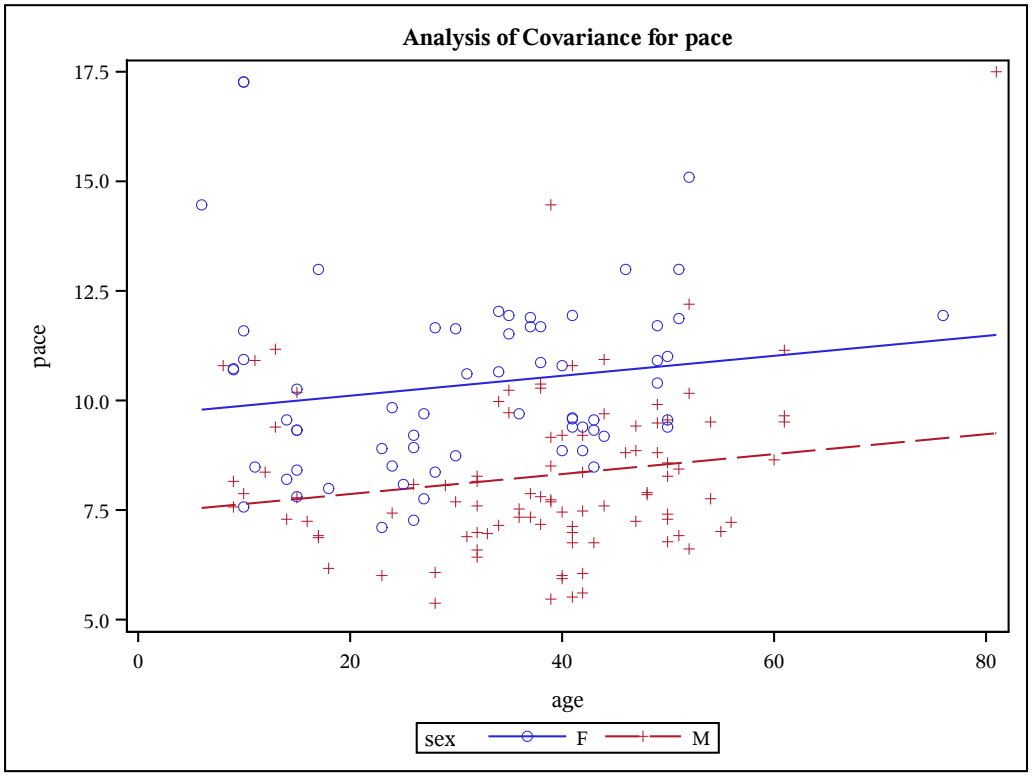
Source	DF	Type III SS	Mean Square	F Value	Pr > F
age	1	16.8597638	16.8597638	4.41	0.0374
sex	1	186.5046294	186.5046294	48.76	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	7.407176454	B	0.45567795	16.26	<.0001	6.507126301	8.307226606
age	0.022717586		0.01082034	2.10	0.0374	0.001345373	0.044089800
sex F	2.256667543	B	0.32316709	6.98	<.0001	1.618351401	2.894983685
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	3	190.8725998	63.6241999	16.62	<.0001
Error	156	597.2221155	3.8283469		
Corrected Total	159	788.0947153			

R-Square	Coeff Var	Root MSE	pace Mean
0.242195	21.45265	1.956616	9.120625

Source	DF	Type I SS	Mean Square	F Value	Pr > F
age	1	1.0965043	1.0965043	0.29	0.5933
sex	1	186.5046294	186.5046294	48.72	<.0001
age*sex	1	3.2714661	3.2714661	0.85	0.3567

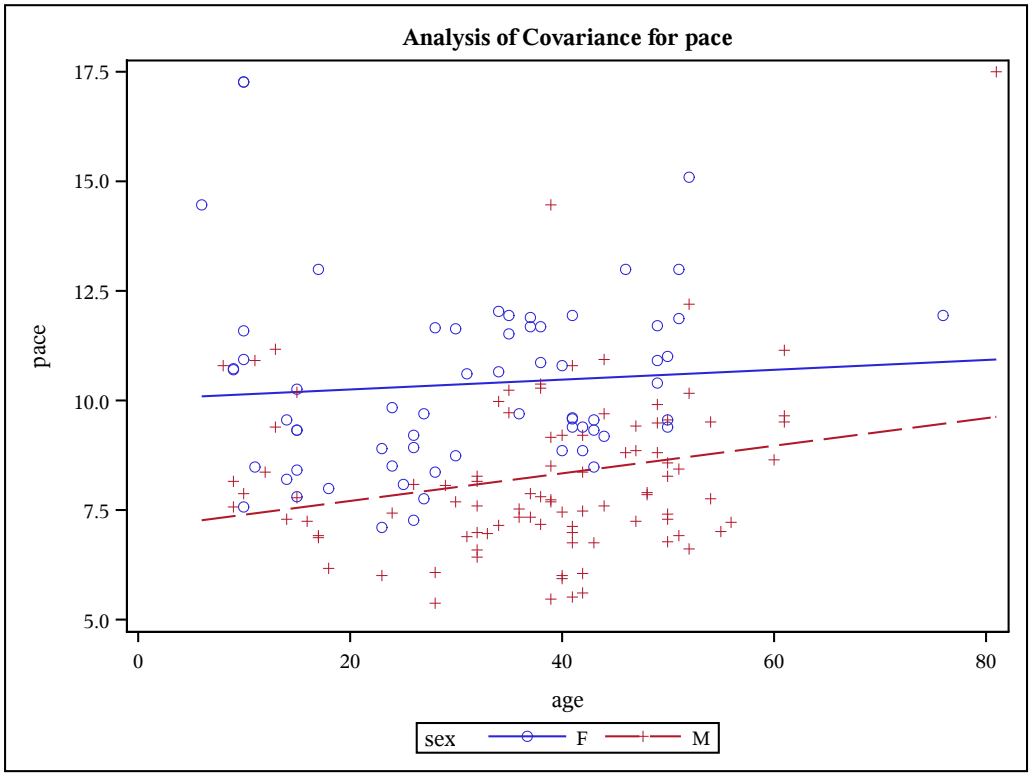
Source	DF	Type III SS	Mean Square	F Value	Pr > F
age	1	14.48792920	14.48792920	3.78	0.0535
sex	1	50.54054069	50.54054069	13.20	0.0004
age*sex	1	3.27146615	3.27146615	0.85	0.3567

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	7.079392510	B	0.57755215	12.26	<.0001	5.938560959	8.220224061
age	0.031386705	B	0.01432253	2.19	0.0299	0.003095594	0.059677816
sex F	2.943260487	B	0.81005480	3.63	0.0004	1.343169360	4.543351615
sex M	0.000000000	B	.	.	.	.	.
age*sex F	-0.020220672	B	0.02187409	-0.92	0.3567	-0.063428289	0.022986946
age*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	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

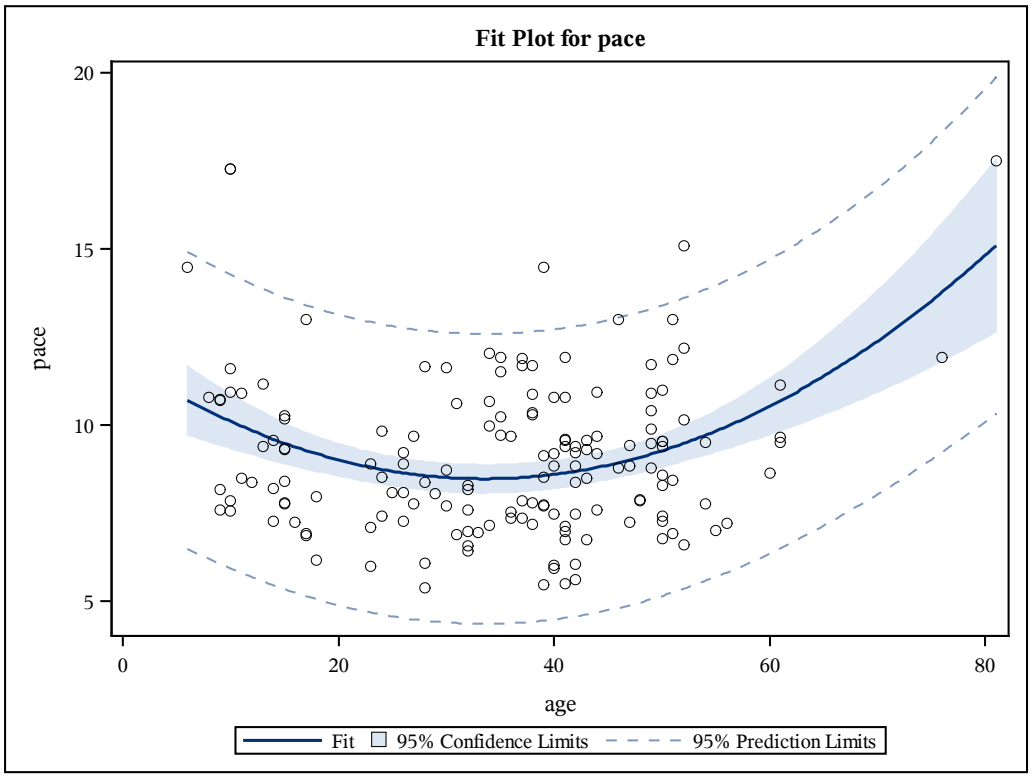
Source	DF	Type I SS	Mean Square	F Value	Pr > F
age	1	1.0965043	1.0965043	0.26	0.6141
age*age	1	112.5484960	112.5484960	26.20	<.0001

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|	95% Confidence Limits	
Intercept	11.78503486	0.70215799	16.78	<.0001	10.39814001	13.17192971
age	-0.19699301	0.04113470	-4.79	<.0001	-0.27824181	-0.11574420
age*age	0.00293699	0.00057380	5.12	<.0001	0.00180363	0.00407034

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

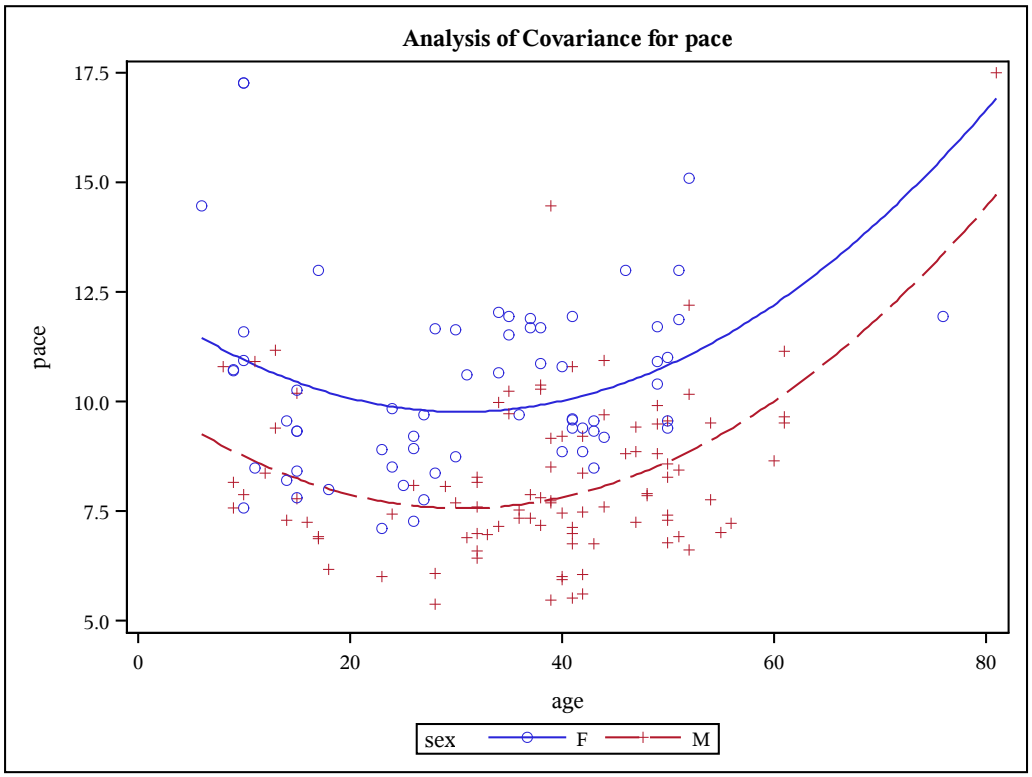
Source	DF	Type III SS	Mean Square	F Value	Pr > F
age	1	73.9438080	73.9438080	23.17	<.0001
age*age	1	102.7473738	102.7473738	32.20	<.0001
sex	1	176.7035071	176.7035071	55.38	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	10.18316690	B	0.64227743	15.85	<.0001	8.91448431	11.45184948
age	-0.17145849		0.03561638	-4.81	<.0001	-0.24181108	-0.10110590
age*age	0.00280792		0.00049481	5.67	<.0001	0.00183052	0.00378531
sex F	2.19792213	B	0.29534621	7.44	<.0001	1.61452846	2.78131581
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



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

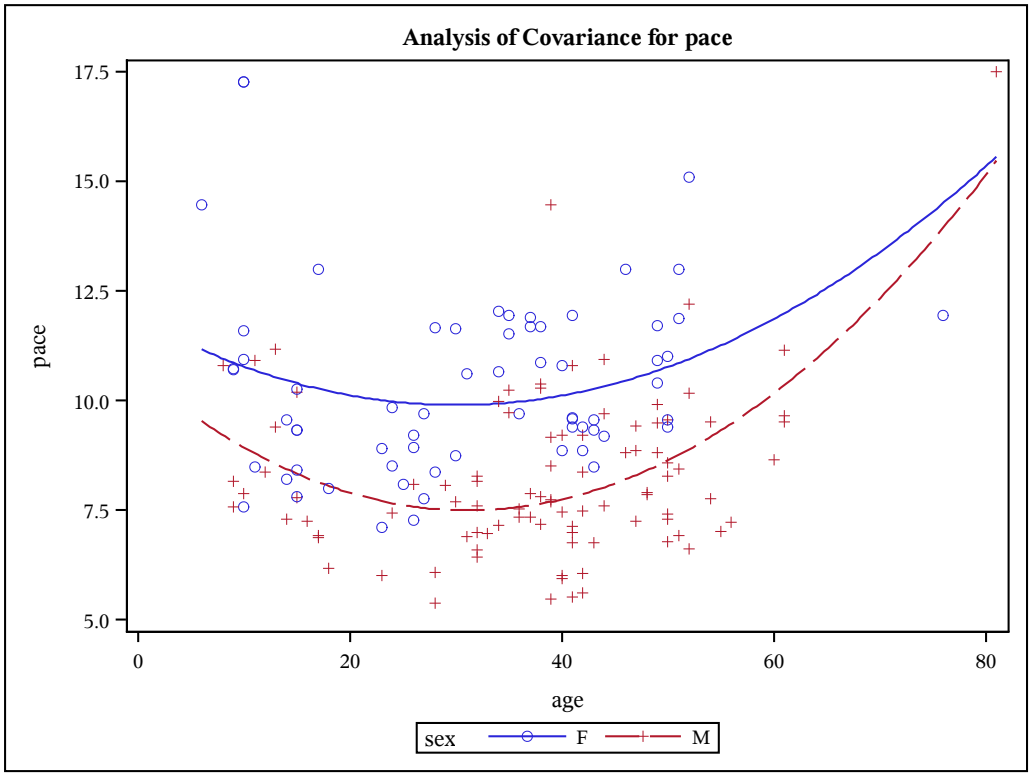
Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	10.60848468	B	0.88640608	11.97	<.0001	8.85740007	12.35956930
age	-0.19985505	B	0.04841621	-4.13	<.0001	-0.29550069	-0.10420940
age*age	0.00320665	B	0.00064628	4.96	<.0001	0.00192993	0.00448336
sex F	1.25727925	B	1.23237262	1.02	0.3092	-1.17725815	3.69181664
sex M	0.00000000	B	.	.	.	.	.
age*sex F	0.06882008	B	0.07297821	0.94	0.3471	-0.07534750	0.21298766
age*sex M	0.00000000	B	.	.	.	.	.

The GLM Procedure

Dependent Variable: pace

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
age*age*sex F	-0.00102594	B	0.00103197	-0.99	0.3217	-0.00306459	0.00101271
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.



For model 7 the fitted model is

$$\hat{\mu}(x_1, x_3) = \begin{cases} \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_1^2 + \hat{\beta}_3(0) & \text{for men} \\ \boxed{= 10.18 - 0.17x + 0.0028x^2} & \\ \hat{\beta}_0 + \hat{\beta}_3 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_1^2 & \text{for women} \\ \boxed{= (10.18 + 2.20) - 0.17x_1 + 0.0028x_1^2} & \end{cases}$$

For model 8 the fitted model is

$$\hat{\mu}(x_1, x_3) = \begin{cases} \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_1^2 + \hat{\beta}_3(0) + \hat{\beta}_4(0) + \hat{\beta}_5(0) & \text{men} \\ \boxed{10.61 - 0.20x_1 + 0.0032x_1^2} & \\ \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_1^2 + \hat{\beta}_3(1) + \hat{\beta}_4(x_1) + \hat{\beta}_5(x_1^2) & \text{women} \\ (\hat{\beta}_0 + \hat{\beta}_3) + (\hat{\beta}_1 + \hat{\beta}_4)x_1 + (\hat{\beta}_2 + \hat{\beta}_5)x_1^2 & \\ (10.61 + 1.26) + (-0.20 + 0.07)x_1 + (0.0032 - 0.0010)x_1^2 & \\ \boxed{11.87 - 0.13x_1 + 0.0022x_1^2} & \end{cases}$$

Which model is "better"? What do we mean by "better?" Is there a test we can use to compare these models?

Yes! F-test for nested models can be used (LOF test).

Comparison of models 7 and 8:

$$\begin{aligned} \text{reduced: } \mu(x_1, x_3) &= \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_3 \\ \text{full: } \mu(x_1, x_3) &= \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_3 + \beta_4 x_1 x_3 + \beta_5 x_1^2 x_3 \end{aligned}$$

Extra Regression sum of squares:

$$SS(R)_f - SS(R)_r = 293.5 - 290.3 = 3.2$$

The F-ratio

$$F = \frac{(SS(R)_f - SS(R)_r)/(p-q)}{MS(E)_f} = \frac{3.2/2}{3.21} = \frac{1.6}{3.21} = 0.5$$

The observed F-ratio is not significant for an F with $df = 2, 154$ $F_{0.05,2,154} = 3.05$.

As the age squared term and sex term are both significant using a type III test, the model quadratic in Age that has separate intercepts for males and females seems appropriate (model 7 seems good).

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, allowing for clearer estimates of treatment effects.

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	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	$\bar{y}_{1\bullet} = 63.0$	2	$\bar{y}_{2\bullet} = 57.4$	3	$\bar{y}_{3\bullet} = 65.2$	4	$\bar{y}_{4\bullet} = 48.8$
1	$s_1 = 12.3$	2	$s_2 = 14.3$	3	$s_3 = 9.7$	4	$s_4 = 10.9$

Main question: Is there evidence of a vitamin supplement effect?

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

equivalently in the glm format $H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \quad vs \quad H_A : \text{at least 1 not 0}$

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

Why might we want to incorporate the caloric intake?

Weight gain may be affected by the the caloric intake. If the different diet groups ate different amounts, then this may be masking or enhancing the diets' effects.

How can we incorporate the caloric intake into the model?

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

(Diet 4 is the baseline here.)

ANCOVA analysis:

```
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			

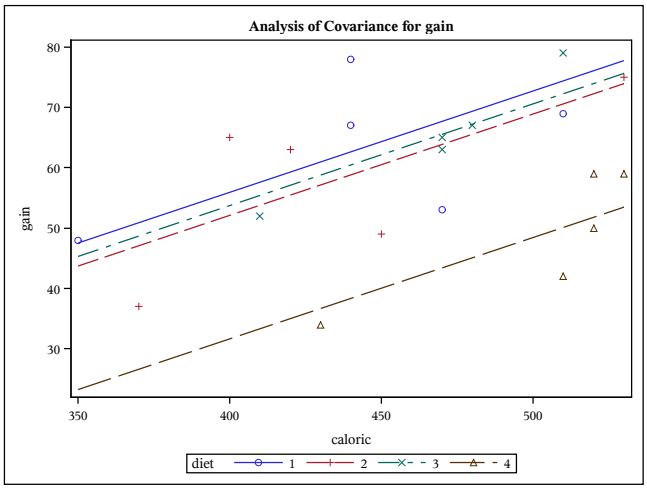
R-Square	Coeff Var	Root MSE	gain Mean
0.623102	15.11308	8.871376	58.70000

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



What type of test should we do if we want to test for a diet effect *once caloric intake has been accounted for*? Type I or Type III?
Type III as we want to see, after we have taken into account how much they ate, whether or not diet had an effect.

To test for a diet effect once caloric intake has been accounted for:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

we compare 6.51 to an F distribution with 3 numerator and 15 denominator degrees of freedom. (Note that this is a comparison of nested models!)

Conclusion?

As $p - value = 0.0049 < 0.05$, reject H_0 in favor of H_A. At the 5% significance level, there is enough evidence to conclude that a diet effect exists once caloric intake has been accounted for.

What now? We know there is a difference due to diet. If we look at the sample means of the weight gains for the diets we can compare their effectiveness. However, the weight gain is also affected by caloric intake...

LSMeans!

LSMEANS - Least Squares Means or Adjusted Means

Estimation of Mean Weight Gains for Diets, Taking into Account 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?

Some of the weight gain for diet 4 may be due to the fact that the caloric intake was much higher for this group. Similarly, lack of weight gain in diet 2 may be due to low caloric intake

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.

Inference for Lsmeans

We have our point estimate, $\bar{y}_{i,a}$.

Now we need to know the **standard errors of** $\bar{Y}_{i,a}$

Consider $\bar{Y}_{2,a}$. What vector c is needed so that $\mathbf{c}^T\hat{\boldsymbol{\beta}} = \bar{Y}_{2,a}$?

$$\mathbf{c}^T = (1 \quad 0 \quad 1 \quad 0 \quad 461.5)$$

What is the standard error of $\mathbf{c}^T\hat{\boldsymbol{\beta}}$?

$$Var(\mathbf{c}^T\hat{\boldsymbol{\beta}}) = \mathbf{c}^T\hat{\boldsymbol{\Sigma}}\mathbf{c} = 17.198$$

This implies

$$SE(\mathbf{c}^T\hat{\boldsymbol{\beta}}) = \sqrt{17.198} = 4.147$$

By the normality, we can now form a CI for the lsmean using the usual $Estimate \pm t\ \hat{SE}(Estimate)$.

Here a 95% CI for the true diet 2 mean weight gain when caloric intake is 461.5 is

$$\bar{Y}_{2,a} \pm t_{20-4-1,0.05/2}\hat{SE}(\bar{Y}_{2,a}) = 62.4 \pm 2.131(4.147) = (53.56, 71.24)$$

We are 95% confident that the true mean weight gain for diet 2 when caloric intake is 461.5 is between 53.56 and 71.24 grams.

Inference for Lsmeans in SAS

```
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

In One-Way Anova with a balanced design, all of the standard errors for the treatment means are equal. Why do they differ here?

These standard errors take into account the covariate, whose mean differed for each treatment group!

Investigating Pairwise Differences of LSMeans

We can now look at all pairwise differences of the lsmeans to see which levels differ significantly. Inference for these differences of lsmeans can again be followed through using the MLR material.

Consider testing the lsmean for group 1 vs that of group 4.

Need to know **standard error of** $\bar{Y}_{1,a} - \bar{Y}_{4,a}$

Consider $\bar{Y}_{1,a}$. From above we know we can find a vector c_1 so that $\mathbf{c}_1^T\hat{\boldsymbol{\beta}} = \bar{Y}_{1,a}$

$$\mathbf{c}_1^T = (1 \quad 1 \quad 0 \quad 0 \quad 461.5)$$

Likewise, we can find c_2 so that $\mathbf{c}_2^T\hat{\boldsymbol{\beta}} = \bar{Y}_{4,a}$

$$\mathbf{c}_2^T = (1 \quad 0 \quad 0 \quad 0 \quad 461.5)$$

Now we can find the estimate of

$$\bar{y}_{1,a} - \bar{y}_{4,a} = \mathbf{c}_1^T\hat{\boldsymbol{\beta}} - \mathbf{c}_2^T\hat{\boldsymbol{\beta}}$$

We simply subtract the vectors elementwise and get

$$\bar{y}_{1,a} - \bar{y}_{4,a} = (0 \quad 1 \quad 0 \quad 0 \quad 0)\hat{\boldsymbol{\beta}} = \mathbf{c}_3^T\hat{\boldsymbol{\beta}}$$

Note: We could have plugged in any common value of the covariate, z, and it would cancel out!

What is the standard error of $\mathbf{c}_3^T\hat{\boldsymbol{\beta}}$?

$$Var(\mathbf{c}_3^T\hat{\boldsymbol{\beta}}) = \mathbf{c}_3^T\hat{\boldsymbol{\Sigma}}\mathbf{c}_3 = 38.4313$$

$$\text{This implies } SE(\mathbf{c}_3^T\hat{\boldsymbol{\beta}}) = \sqrt{38.4313} = 6.1993$$

We can now form a CI for this difference in lsmeans using the usual $Estimate \pm t\ \hat{SE}(Estimate)$.

Here a 95% interval is:

$$\bar{y}_{1,a} - \bar{y}_{4,a} \pm t_{20-4-1,0.05/2}\hat{SE}(\bar{y}_{1,a} - \bar{y}_{4,a}) = 24.295 \pm 2.131(6.1993) = (11.084, 37.506)$$

As 0 is not in this interval, we are 95% confident that the two treatment group means differ significantly if they have the same caloric intake.

Investigation of Pairwise Differences of LSMeans using SAS

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

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

diet	gain LSMEAN	LSMEAN Number
1	66.2809372	1
2	62.4269627	2
3	64.1063543	3
4	41.9857458	4

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

i/j	1	2	3	4
1		0.9010	0.9806	0.0067
2	0.9010		0.9912	0.0265
3	0.9806	0.9912		0.0083
4	0.0067	0.0265	0.0083	

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

Least Squares Means for Effect diet

i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	3.853974	-12.348360	20.056309
1	3	2.174583	-14.327874	18.677040
1	4	24.295191	6.428260	42.162122
2	3	-1.679392	-18.413474	15.054691
2	4	20.441217	2.120450	38.761983
3	4	22.120608	5.386526	38.854691

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 - Random Effects Models

Readings: 17.1-17.2

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 weight 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	X	
- finite number of possible levels		X
Another expt - would use same levels		X
- would involve new levels sampled from same population	X	
Goal - estimate varcomps	X	
- estimate longrun means		X
Inference - for these levels used in this expt		X
- for the population of levels	X	

Contrast this situation with the binding fractions example. Why not model antibiotic effects random? Why fixed?
We care about those particular antibiotics and have an interest in their means.

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:

$$H_0 : \sigma_T^2 = 0 \quad vs. \quad H_A : \sigma_T^2 > 0$$

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

$$\hat{\mu} = \bar{y}_{\bullet\bullet}$$

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, $\bar{y}_{\bullet\bullet} = 82.6$ 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 in birthweight?

   $Var(Y_{ij}) = \sigma_T^2 + \sigma^2$

   $\hat{Var}(Y_{ij}) = \hat{\sigma}_T^2 + \hat{\sigma}^2 = 117 + 464 = 581$

2. Estimate the proportion of this variation that is due to the sire effect.

$$\frac{\hat{\sigma}_T^2}{\hat{\sigma}_T^2 + \hat{\sigma}^2} = \frac{117}{581} = 0.2$$

3. Estimate the proportion of this variation that is not due to the sire effect.

$$\frac{\hat{\sigma}^2}{\hat{\sigma}_T^2 + \hat{\sigma}^2} = \frac{464}{581} = 0.8$$

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 derived using normality and the t distribution:

$$\begin{aligned}\bar{Y}_{\bullet\bullet} &= \frac{1}{N}\sum_{i=1}^{t}\sum_{j=1}^{n} Y_{ij} \\ &= \frac{1}{N}\sum_{i=1}^{t}\sum_{j=1}^{n}(\mu + T_i + E_{ij}) \\ &= \mu + \bar{T}_{\bullet} + \bar{E}_{\bullet\bullet}\end{aligned}$$

where $\bar{T}_{\bullet} = (T_1 + \cdots + T_t)/t$ and $\bar{E}_{\bullet\bullet} = (\sum\sum E_{ij})/N$, so that

$$\begin{aligned}\mathrm{Var}(\bar{Y}_{\bullet\bullet}) &= \mathrm{Var}(\bar{T}_{\bullet} + \bar{E}_{\bullet\bullet}) \\ &= \frac{\sigma_T^2}{t} + \frac{\sigma^2}{nt} \\ &= \frac{1}{nt}(n\sigma_T^2 + \sigma^2) \\ &= \frac{1}{nt}E(MS[T]).\end{aligned}$$

If the data are normally distributed, then

$$\frac{\bar{Y}_{\bullet\bullet} - \mu}{\sqrt{\frac{MS(T)}{nt}}} \sim t_{t-1}$$

and a $100(1-\alpha)\%$ confidence interval for μ given by

$$\bar{Y}_{\bullet\bullet} \pm t(\alpha/2, t-1)\sqrt{\frac{MS[T]}{nt}}$$

Sires data: $\bar{y}_{\bullet\bullet} = 82.6, MS(T) = 1398, nt = 40$. 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

$$\left(\frac{(N-t)MS[E]}{\chi^2_{\alpha/2,N-t}}, \frac{(N-t)MS[E]}{\chi^2_{1-\alpha/2,N-t}} \right).$$

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. This means SAS is not estimating by equating the mean squares, but rather is using the normal distribution to find estimates.

```
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, $\rho_I, \sigma^2, \sigma_T^2, \mu$:

$$\bar{Y}_{\bullet\bullet} \pm t(0.025, t-1)\sqrt{\frac{MS[T]}{nt}}$$

$$\left(\frac{F_{obs} - F_{1-\alpha/2}}{F_{obs} + (n-1)F_{1-\alpha/2}}, \frac{F_{obs} - F_{\alpha/2}}{F_{obs} + (n-1)F_{\alpha/2}} \right)$$

$$\left(\frac{(N-t)MS[E]}{\chi^2_{\alpha/2,N-t}}, \frac{(N-t)MS[E]}{\chi^2_{1-\alpha/2,N-t}} \right)$$

$$\left(\frac{\widehat{df}\hat{\sigma}_T^2}{\chi^2_{\alpha/2,\widehat{df}}}, \frac{\widehat{df}\hat{\sigma}_T^2}{\chi^2_{1-\alpha/2,\widehat{df}}} \right)$$

Chapter 9

ST 512 - Nested Designs

Readings: 17.6

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 the levels of factor B differ for at least one 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 = $\bar{y}_{\bullet\bullet\bullet} = 22.3$). $\bar{y}_{(i)\bullet}$
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 $b_{(3)2}$ 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.

  Factors are Brand and Flavor. Brand has 3 levels, flavor has 2 levels and the levels depend on Brand, so flavor is nested in brand. Flavor(Brand).

- Gum arabic is used to lengthen the shelf life 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.

  Factors are Variety and Treatment. Variety has 2 levels, treatment has 2 levels (trt or control). These are all observed together so this is a crossed experiment.

- 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.

  Factors are setting and story. Setting has 2 levels and Story has 3 levels. The levels of the story depend on the setting so story is nested in setting. Story(Setting).

  Note: Story would probably be considered random. More on this later.

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:

$$Y_{ijk} = \mu + \tau_i + \beta_{(i)j} + E_{ijk}, \quad E_{ijk} \sim^{iid} N(0, \sigma^2)$$

$i = 1, 2, \ldots, a \quad j = 1, 2, \ldots, b_i \quad k = 1, 2, \ldots, n_{ij}$ with some sum to zero constraints

What are the interpretations of each parameter in the above model?

- μ - overall mean
- τ_i - effect of level i of factor A
- $\beta_{(i)j}$ - effect of level j of B for level i of A

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)	$MS(A) = \frac{SS(A)}{a-1}$	$\sigma^2 + nb\psi_A^2$	F= $\frac{MS(A)}{MS(E)}$
B(A)	$\sum_{i=1}^a b_i - a$	SS(B(A))	$MS(B(A)) = \frac{SS(B(A))}{\sum_{i=1}^a b_i - a}$	$\sigma^2 + n\psi_{B(A)}^2$	F= $\frac{MS(B(A))}{MS(E)}$
Error	$N - \sum_{i=1}^a b_i$	SS(E)	$MS(E) = \frac{SS(E)}{N - \sum_{i=1}^a b_i}$	σ^2	

The null and alternative hypotheses of interest for a nested experiment are

Global test $H_0 : \tau_i = 0$ for all i and $\beta_{(i)j} = 0$ for all i, j H_A : At least 1 differs

Test for effect of A $H_0 : \tau_i = 0$ for all i H_A : At least 1 differs

Test for nested factor B $H_0 : \beta_{(i)j} = 0$ for all i, j H_A : At least 1 differs

and these hypotheses jointly (the global test). In words, what are each of these hypotheses testing?

Global: Is anything in the model useful?
τ_i all 0: does factor A have an effect?
$\beta_{(i)j} = 0$ for all i,j: do any of the levels of factor B differ from each other?

Assuming the above null hypotheses are true, what value is expected for each F in the above ANOVA table? What values give evidence against these H_0's?

F's should be approximately 1 under H_0. F's larger than 1 give evidence for H_A.

There is no interaction being tested here, why do you think that is not something we look at for a nested design? Can't inspect interaction! No level of B at each level of A!

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?

Time, Technician. Technician(Time)

(Let's consider technician as fixed for now.) Write out the statistical model for these data and interpret μ, τ_i and $\beta_{(i)j}$ in terms of the experiment.

$$Y_{ijk} = \mu + \tau_i + \beta_{(i)j} + E_{ijk} \qquad E_{ijk} \sim^{iid} N(0, \sigma^2)$$

$\mu =$ overall mean amount of readily soluble phosphorus

$\tau_i =$ effect of time (AM=1 or PM=2)

$\beta_{(i)j} =$ effect of technician j at time i

The model is fit in SAS using the code here:

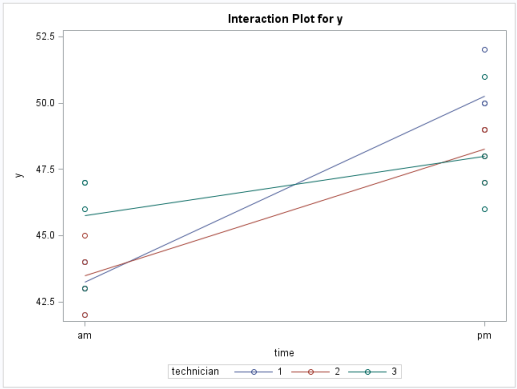
```
proc glm data=phosph;
class time technician;
model y=time technician(time);
run;
```

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?

   The global test in the ANOVA table tests this hypothesis. P-value<0.0001 so reject H_0 : τ_i all 0 and $\beta_{(i)j}$ all 0 in favor of H_A : at least one of those parameters is not 0.

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.

   p-value=0.0429 <0.05, Reject H_0 : $\beta_{(i)j}$ all 0 in favor of H_A that at least one is not 0.
   Conclusion: Technicians vary in at least one of the times.

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.

   p-value< 0.0001 <0.05, Reject H_0 : τ_i all 0 in favor of H_A that at least one is not 0.
   Conclusion: Time has an effect on the response.

4. Is the interaction plot given by SAS meaningful? Why?

   No, technician 1 in the AM doesn't have any relationship with technician 1 in the PM. The labels are arbitrary so the connected lines are not meaningful.

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 10

ST 512 - Mixed Models

Readings: 17.1-17.8

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: Temp
   - Second factor: Sucrose
   - Fixed or random? Both fixed
   - Crossed or nested? Crossed
   - Model: $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + E_{ijk}$

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:Drug
- Second factor:Admin
- Fixed or random?Both fixed
- Crossed or nested?Nested, Admin(Drug)
- Model: $Y_{ijk} = \mu + \alpha_i + \beta_{(i)j} + E_{ijk}$

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:Lab
- Second factor:Sample
- Fixed or random?Both random
- Crossed or nested?Crossed
- Model: $Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + E_{ijk}$

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:Location
   - Second factor:Day
   - Fixed or random?Loacation fixed, Day random
   - Crossed or nested?Crossed
   - Model: $Y_{ijk} = \mu + \alpha_i + B_j + (\alpha B)_{ij} + E_{ijk}$

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:Plant
   - Second factor:Leaf
   - Fixed or random?Both random
   - Crossed or nested?Nested, Leaf(Plant)
   - Model: $Y_{ijk} = \mu + A_i + B_{(i)j} + E_{ijk}$

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:Treatment
- Second factor:Pot
- Fixed or random?Treatment fixed, Pot random
- Crossed or nested?Nested, Plot(Treatment)
- Model: $Y_{ijk} = \mu + \alpha_i + B_{(i)j} + E_{ijk}$

Recap: Six types of two-factor models possible with fixed and/or random effects that are either crossed or nested.

Experiment Number	Model	Identifier
3	$Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + E_{ijk}$	crossed/random
2	$Y_{ijk} = \mu + \alpha_i + \beta_{(i)j} + E_{ijk}$	nested/fixed
5	$Y_{ijk} = \mu + A_i + B_{(i)j} + E_{ijk}$	nested/random
4	$Y_{ijk} = \mu + \alpha_i + B_j + (\alpha B)_{ij} + E_{ijk}$	crossed/mixed
6	$Y_{ijk} = \mu + \alpha_i + B_{(i)j} + E_{ijk}$	nested/mixed
1	$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + E_{ijk}$	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.

$$\begin{aligned}
\psi_A^2 &= \frac{1}{a-1}\sum_1^a \alpha_i^2 \quad \text{effect size of factor } A \\
\psi_B^2 &= \frac{1}{b-1}\sum_1^b \beta_i^2 \quad \text{effect size of factor } B \\
\psi_{AB}^2 &= \frac{1}{(a-1)(b-1)}\sum_{i=1}^a\sum_{j=1}^b (\alpha\beta)_{ij}^2 \quad \text{effect size of interaction} \\
\psi_{B(A)}^2 &= \frac{1}{a(b-1)}\sum_{i=1}^a\sum_{j=1}^b \beta_{(i)j}^2 \quad \text{effect size of factor } B \\
\sigma_A^2 &= \text{Var}(A_i) \quad \text{variance component for factor } A \\
\sigma_B^2 &= \text{Var}(B_i) \quad \text{variance component for factor } B \\
\sigma_{AB}^2 &= \text{Var}((AB)_{ij}) \quad \text{variance component for interaction} \\
\sigma_{B(A)}^2 &= \text{Var}(B_{(i)j}) \quad \text{variance component for factor } B \\
\sigma^2 &= \text{Var}(E_{ijk)}) \quad \text{error variance}
\end{aligned}$$

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

$$F = MS(A)/MS(AB) \quad vs \quad F(\alpha, a-1, (a-1)(b-1))$$

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.

ex: Suppose you have 3 factors with factor B fixed and factors A and C random, factors crossed. To find the expected mean squares:

$$Y_{ijkl} = \mu + A_i + \beta_j + C_k + (A\beta)_{ij} + (AC)_{ik} + (\beta C)_{jk} + (A\beta C)_{ijk} + E_{ijkl}$$

Then

$$E(MS(B)) = \sigma^2 + n\sigma_{A\beta C}^2 + na\sigma_{\beta C}^2 + nc\sigma_{A\beta}^2 + nac\psi_\beta^2$$

$$E(MS(ABC)) = \sigma^2 + n\sigma_{A\beta C}^2$$

$$E(MS(A)) = \sigma^2 + n\sigma_{A\beta C}^2 + nb\sigma_{AC}^2 + nc\sigma_{A\beta}^2 + nbc\sigma_A^2$$

Sometimes you have to solve for things to get correct error term. Error term to use for testing of factor B's importance is

$$MS(AB) + MS(BC) - MS(ABC)$$

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.

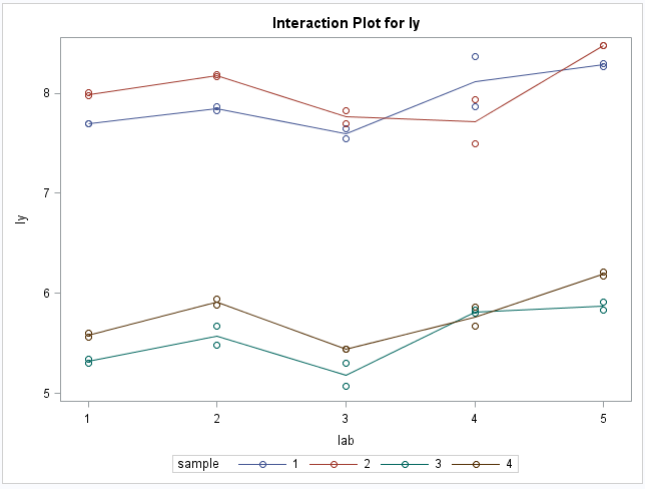
```
proc glm; class lab sample;
model ly=sample|lab;
random sample lab sample*lab; run;
```

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, $F \sim F(a-1, (a-1)(b-1)) = F(4, 12)$, which has $F(0.05, 4, 12) = 3.25$, yielding the $\alpha = 0.05$ critical region reject H_0 if $F_{obs} > 3.25$ 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: $F(0.05, 3, 12) = 3.49$)

$$F = \frac{MS(B)}{MS(AB)} = \frac{17.75}{0.045} = 391.94 > 3.49 \text{ so Reject } H_0 : \sigma_B^2 = 0$$

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 $\bar{Y}_{\bullet\bullet\bullet}$ can be derived by

$$\begin{aligned} \bar{Y}_{\bullet\bullet\bullet} &= \mu + \bar{A}_{\bullet} + \bar{B}_{\bullet} + \overline{(AB)}_{\bullet\bullet} + \bar{E}_{\bullet\bullet\bullet} \\ \text{Var}(\bar{Y}_{\bullet\bullet\bullet}) &= \text{Var}(\bar{A}_{\bullet}) + \text{Var}(\bar{B}_{\bullet}) + \text{Var}(\overline{(AB)}_{\bullet\bullet}) + \text{Var}(\bar{E}_{\bullet\bullet\bullet}) \\ &= \frac{\sigma^2_A}{a} + \frac{\sigma^2_B}{b} + \frac{\sigma^2_{AB}}{ab} + \frac{\sigma^2}{abn} \end{aligned}$$

Estimation of standard error and approximation of *df*:

The standard error

$$SE(\bar{Y}_{\bullet\bullet\bullet}) = \sqrt{\frac{\sigma^2_A}{a} + \frac{\sigma^2_B}{b} + \frac{\sigma^2_{AB}}{ab} + \frac{\sigma^2}{abn}}$$

can be estimated by substitution of estimated variance components ($\hat{\sigma}^2$), which leads to

$$\begin{aligned} \widehat{SE}(\bar{Y}_{\bullet\bullet\bullet}) &= \sqrt{\frac{\hat{\sigma}^2_A}{a} + \frac{\hat{\sigma}^2_B}{b} + \frac{\hat{\sigma}^2_{AB}}{ab} + \frac{\hat{\sigma}^2}{abn}} \\ &= \text{lots of algebra and cancellations} \\ &= \sqrt{\frac{1}{nab}\left(MS[A] + MS[B] - MS[AB]\right)} \end{aligned}$$

For the milk data, we have

$$\widehat{SE}(\bar{Y}_{\bullet\bullet\bullet}) = \sqrt{\frac{1}{40}(0.58 + 17.73 - 0.05)} = 0.6757$$

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 $\widehat{SE}(\bar{Y}_{\bullet\bullet\bullet})$ 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

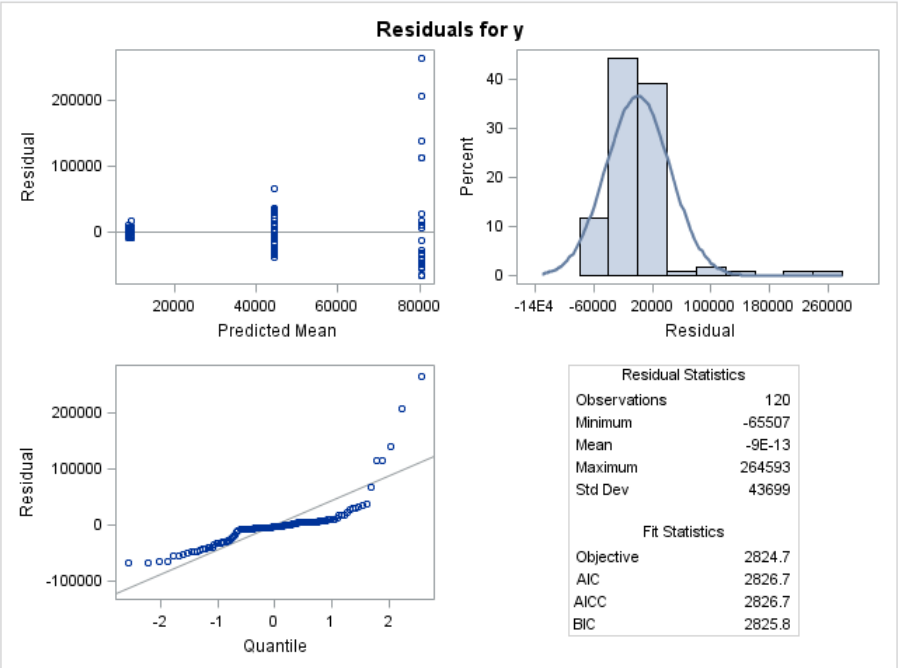
$$Y_{ijk} = \mu + \alpha_i + B_j + (\alpha B)_{ij} + E_{ijk}$$

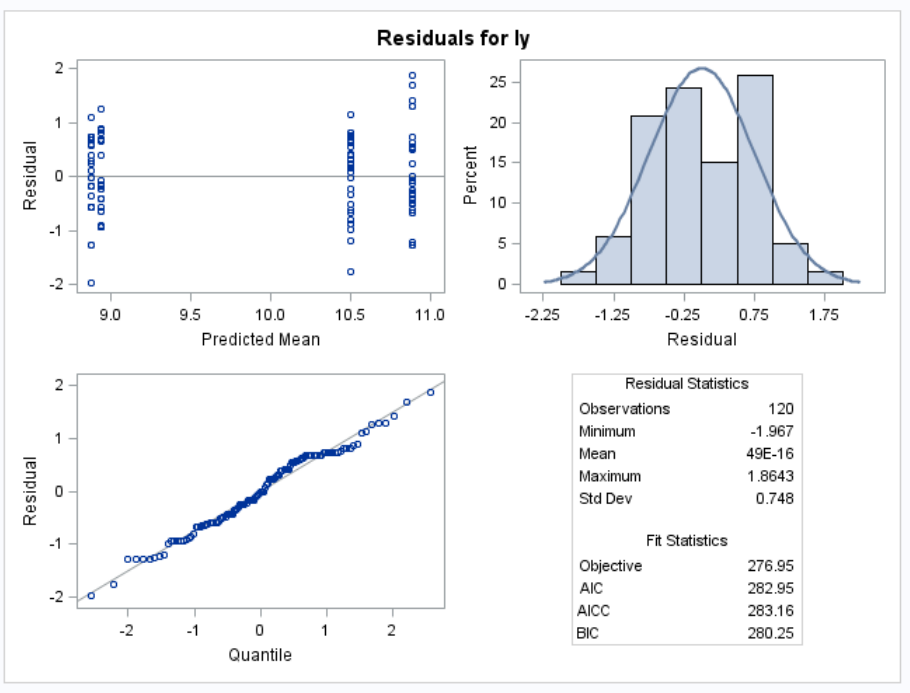
with variance components $\sigma^2_B, \sigma^2_{\alpha B}, \sigma^2$.

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

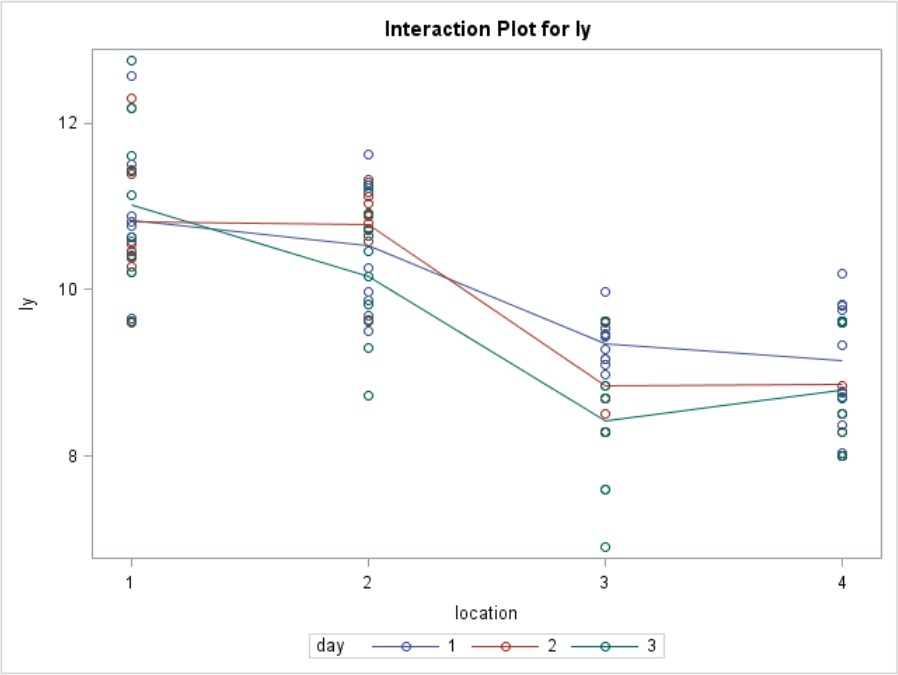
Cov Parm	Estimate
day	0.01595
day*location	0.02069
Residual	0.5487

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? same location, same day
- at different locations? different location, same day

Recall that $Y_{ijk} = \mu + \alpha_i + B_j + (\alpha B)_{ij} + E_{ijk}$.

$$\begin{aligned}
Corr(Y_{ijk_1}, Y_{ijk_2}) &= \frac{\text{Cov}(Y_{ijk_1}, Y_{ijk_2})}{\sigma^2 + \sigma^2_B + \sigma^2_{\alpha B}} \\
&= \frac{\text{Cov}(B_j, B_j) + \text{Cov}((\alpha B)_{ij}, (\alpha B)_{ij})}{\sigma^2 + \sigma^2_B + \sigma^2_{\alpha B}} \\
&= \frac{\sigma^2_B + \sigma^2_{\alpha B}}{\sigma^2 + \sigma^2_B + \sigma^2_{\alpha B}} \\
Corr(Y_{1jk_1}, Y_{2jk_2}) &= \frac{\text{Cov}(Y_{1jk_1}, Y_{2jk_2})}{\sigma^2 + \sigma^2_B + \sigma^2_{\alpha B}} \\
&= \frac{\text{Cov}(B_j, B_j)}{\sigma^2 + \sigma^2_B + \sigma^2_{\alpha B}} \\
&= \frac{\sigma^2_B}{\sigma^2 + \sigma^2_B + \sigma^2_{\alpha B}}
\end{aligned}$$

Estimates of these correlations are

- $\frac{0.016+0.021}{0.016+0.021+0.55} = \frac{0.037}{.587} = 0.063$
- $\frac{0.016}{0.016+0.021+0.55} = \frac{0.016}{.587} = 0.027$

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,

$$F_A = \frac{MS[A]}{MS[AB]} = \frac{32.6}{0.76} = 43.2$$

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

$$\hat{\theta} = \bar{y}_{4\bullet\bullet} - \bar{y}_{3\bullet\bullet} = 8.940 - 8.875 = -0.065$$

Note that

$$\mathrm{Var}(\bar{Y}_{4\bullet\bullet} - \bar{Y}_{3\bullet\bullet}) \neq \sigma^2(\frac{1}{nb} + \frac{1}{nb})$$

Since our Y's are not independent anymore!

What is $SE(\hat{\theta})$ and how can it be estimated?

$$\begin{aligned}
\hat{\theta} &= \bar{Y}_{4\bullet\bullet} - \bar{Y}_{3\bullet\bullet} \\
&= \mu + \alpha_4 + \bar{B}_\bullet + \overline{\alpha B}_{4\bullet} + \overline{E}_{4\bullet\bullet} - (\mu + \alpha_3 + \bar{B}_\bullet + \overline{\alpha B}_{3\bullet} + \overline{E}_{3\bullet\bullet}) \\
&= \alpha_4 - \alpha_3 + \overline{\alpha B}_{4\bullet} - \overline{\alpha B}_{3\bullet} + \overline{E}_{4\bullet\bullet} - \overline{E}_{3\bullet\bullet}
\end{aligned}$$

which has variance

$$\begin{aligned}
\mathrm{Var}(\hat{\theta}) &= \mathrm{Var}(\overline{\alpha B}_{4\bullet}) + \mathrm{Var}(\overline{\alpha B}_{3\bullet}) + \mathrm{Var}(\overline{E}_{4\bullet\bullet}) + \mathrm{Var}(\overline{E}_{3\bullet\bullet}) \\
&= 2\frac{\sigma^2_{\alpha B}}{b} + 2\frac{\sigma^2}{nb} \\
&= \frac{2}{nb}(\sigma^2 + n\sigma^2_{\alpha B})
\end{aligned}$$

which can be estimated nicely on $(a-1)(b-1)=6$ df by

$$\hat{\mathrm{Var}}(\hat{\theta}) = \frac{2}{nb} MS[AB]$$

for the chickens, where $\overline{y}_{4\bullet\bullet} - \overline{y}_{3\bullet\bullet} = -0.06$ 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:

$$\begin{aligned}
\overline{Y}_{i\bullet\bullet} &= \mu + \alpha_i + \overline{B}_\bullet + \overline{\alpha B}_{i\bullet} + \overline{E}_{i\bullet\bullet} \\
\text{Var}(\overline{Y}_{i\bullet\bullet}) &= \text{Var}(\overline{B}_\bullet) + \text{Var}(\overline{\alpha B}_{i\bullet}) + \text{Var}(\overline{E}_{i\bullet\bullet}) \\
&= \frac{\sigma_B^2}{b} + \frac{\sigma_{\alpha B}^2}{b} + \frac{\sigma^2}{nb} \\
&= \frac{1}{nb}(n\sigma_B^2 + n\sigma_{\alpha B}^2 + \sigma^2) \\
&\quad \text{estimated by} \\
\widehat{\text{Var}}(\overline{Y}_{i\bullet\bullet}) &= \frac{1}{nb}(n\hat{\sigma}_B^2 + n\hat{\sigma}_{\alpha B}^2 + \hat{\sigma}^2) \\
&= \text{algebra yields a linear combo of multiple EMS terms} \\
&= \frac{1}{nab}\{(a-1)MS[AB] + MS[B]\}
\end{aligned}$$

The standard error is estimated easily enough:

$$\begin{aligned}
\widehat{SE}(\overline{Y_{i\bullet\bullet}}) &= \sqrt{\frac{1}{nab}\{(a-1)MS[AB] + MS[B]\}} \\
&= \sqrt{\frac{1}{120}\{(4-1)0.76 + 1.39\}} \\
&= \sqrt{0.03} = 0.175
\end{aligned}$$

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 11

ST 512 - Block Designs

Readings: 15.1-15.2, 15.4

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!

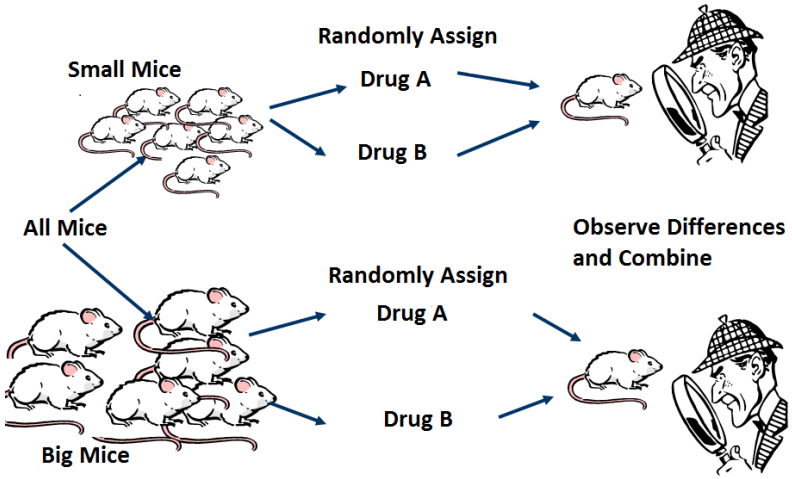
Recall:

- 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 affect 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	$nt\psi_B^2 + \sigma^2$	$nt\sigma_B^2 + n\sigma_{B*Trt}^2 + \sigma^2$
Treatment	t-1	$nb\psi_{Trt}^2 + \sigma^2$	$nb\psi_{Trt}^2 + n\sigma_{B*Trt}^2 + \sigma^2$
Block*Treatment	(b-1)(t-1)	$n\psi_{B*Trt}^2 + \sigma^2$	$n\sigma_{B*Trt}^2 + \sigma^2$
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.

We have no replication so we can't estimate this!

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	$nt\psi_B^2 + \sigma^2$	$nt\sigma_B^2 + n\sigma_{B*Trt}^2 + \sigma^2$
Treatment	t-1	$nb\psi_{Trt}^2 + \sigma^2$	$nb\psi_{Trt}^2 + n\sigma_{B*Trt}^2 + \sigma^2$
Block*Treatment	(b-1)(t-1)	$n\psi_{B*Trt}^2 + \sigma^2$	$n\sigma_{B*Trt}^2 + \sigma^2$
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 ($\psi_{B*Trt}^2 = 0$) 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	$nt\psi_B^2 + \sigma^2$	$nt\sigma_B^2 + n\sigma_{B*Trt}^2 + \sigma^2$
Treatment	t-1	$nb\psi_{Trt}^2 + \sigma^2$	$nb\psi_{Trt}^2 + n\sigma_{B*Trt}^2 + \sigma^2$
Block*Treatment	(b-1)(t-1)	$n\psi_{B*Trt}^2 + \sigma^2$	$n\sigma_{B*Trt}^2 + \sigma^2$
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 the 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.

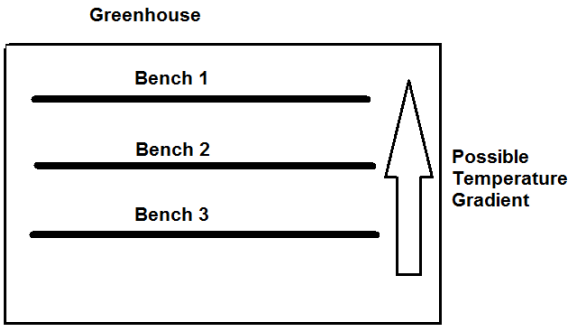
Blocks random:

- $\bar{y}_{.j}$ 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 $MS(Block * Trt)$

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	.	.	.

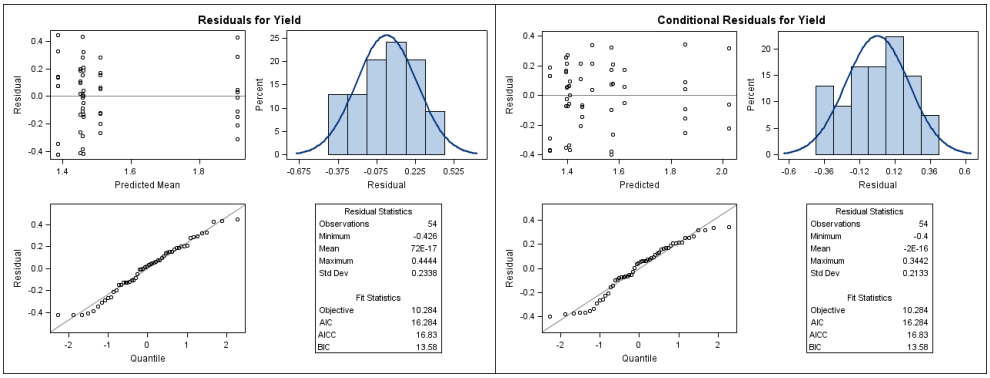
We can see we have a significant Fertilizer effect. We don't care about the block effect significance. Even if not significant we would leave this effect in!

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 12

ST 512 - Split Plot Designs (A Repeated Measures Model)

Readings: 18.1-18.7

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

$$Y_{ijk} = \mu + \alpha_i + S_{(i)k} + \beta_j + (\alpha\beta)_{ij} + E_{ijk}$$

- $i = 1, 2, 3$ pesticides (generally, $i = 1, ..., a$)
- $j = 1, 2, 3, 4$ irrigation (generally, $j = 1, ..., b$)
- $k = 1, 2$ plots (generally $k = 1, ..., r$)
- $S_{(i)j} \sim^{iid} N(0, \sigma_S^2)$ '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! One possible way to make inference? Look at expected mean squares!

$$Y_{ijk} = \underbrace{\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}}_{\mu_{ij}:(\text{fixed component})} + \overbrace{S_{(i)k} + E_{ijk}}^{\text{random error components}} .$$

Here, $i = 1, \ldots, a$ and $j = 1, \ldots, b$ and $k = 1, \ldots, r$ where r denotes the number of plots treated with level i of factor a. (Note: Ott and Longnecker uses different subscripts.)

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) (Subplot error)	$(b-1)(r-1)a = 9$	σ^2
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: $\bar{y}_{i_1\bullet\bullet} - \bar{y}_{i_2\bullet\bullet}$
2. Main effects of B: $\bar{y}_{\bullet j_1\bullet} - \bar{y}_{\bullet j_2\bullet}$
3. Simple effects of A: $\bar{y}_{i_1 j\bullet} - \bar{y}_{i_2 j\bullet}$
4. Simple effects of B: $\bar{y}_{i j_1\bullet} - \bar{y}_{i j_2\bullet}$
5. Interaction effects: $\bar{y}_{i_1 j_1\bullet} - \bar{y}_{i_2 j_2\bullet}$

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*
$\bar{Y}_{i_1\bullet\bullet} - \bar{Y}_{i_2\bullet\bullet}$	$\frac{2}{rb}(\sigma^2 + b\sigma_s^2)$	$\frac{2}{rb}MS[S(A)]$	$(r-1)a$
$\bar{Y}_{\bullet j_1\bullet} - \bar{Y}_{\bullet j_2\bullet}$	$\frac{2}{ra}\sigma^2$	$\frac{2}{ra}MS[E]$	$(r-1)(b-1)a$
$\bar{Y}_{i_1 j\bullet} - \bar{Y}_{i_2 j\bullet}$	$\frac{2}{r}(\sigma^2 + \sigma_s^2)$	$\frac{2}{r}(\hat{\sigma}^2 + \hat{\sigma}_s^2)$	messy
$\bar{Y}_{i j_1\bullet} - \bar{Y}_{i j_2\bullet}$	$\frac{2}{r}\sigma^2$	$\frac{2}{r}MS[E]$	$(r-1)(b-1)a$
$\bar{Y}_{i_1 j_1\bullet} - \bar{Y}_{i_2 j_2\bullet}$	$\frac{2}{r}(\sigma^2 + \sigma_s^2)$	$\frac{2}{r}(\hat{\sigma}^2 + \hat{\sigma}_s^2)$	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.