GENERALISED LINEAR MODELLING
LECTURE NOTES: ANALYSIS OF VARIANCE MODELS

## I. Introduction

The most fundamental principle at the foundation of statistical modelling is the decomposition of an observed random quantity $Y$ into a deterministic component, $\mu$, and a random, or *stochastic*, component $\epsilon$. Generally, the deterministic component will derive from some scientific theory or mathematical model of the world, while the stochastic component is just a formalisation of the generally observed phenomenon that repeated measurements on the world tend to vary somewhat from trial to trial even under ostensibly identical conditions. Typically, we will denote this latter formalisation as $\epsilon \sim F$ (read "$\epsilon$ distributed according to $F$), for the probability distribution determined by the cumulative distribution function (or $CDF$) $F$. The trick, of course, is to try and piece out what part of the observation $Y$ belongs to $\mu$ and which to $\epsilon$, since usually neither $\mu$ or $\epsilon$ is observed directly. Now, the standard linear model for an observed collection of $n$ responses, $Y = (Y_1, \ldots, Y_n)^T$, assumes an additive decomposition (i.e., that $Y = \mu + \epsilon$) with $\mu = X\beta$ for a known design matrix, $X$, containing values of explanatory variables and a vector, $\beta$, of fixed but unknown parameter values, and also assumes that $F$ is the multivariate normal distribution with mean-vector 0 and variance-covariance matrix $\sigma^2 I$, where $I$ is the $n \times n$ identity matrix (i.e., the $n$ elements of the random vector $\epsilon$ are independent and each have the same variability, $\sigma^2$). The field of generalised linear modelling examines models for similar observed data structures, but where each of these strict standard linear model assumptions are relaxed to some degree.

## II. One-way Analysis of Variance

WHAT IS GLM?

The common development of the analysis of variance (ANOVA) proceeds as an extension of the two-sample $t$-test for comparison of the means of two distinct populations or categories of response values. In this vein, the ANOVA is seen as a method of testing whether the mean responses at different values, or *levels* of a categorical predictor, or *factor*, are all equal.

### i. Model and Assumptions

Suppose that we have data values sampled from $k$ different levels of some factor, and we want to investigate the mean response within each of these factor levels. We will denote our data values as $Y_{ij}$, where $i = 1, \ldots, k$ indicates the factor level at which the observation was sampled and $j = 1, \ldots, n_i$ indicates a specific value within the group of observations at the $i^{\text{th}}$ factor level. Note that we use the notation $n_i$ to indicate that we may not necessarily sample the same number of observations at each factor level. The most straightforward model is then:

$$Y_{ij} = \mu_i + \epsilon_{ij},$$

where $\mu_i$ is the expected response at the $i^{\text{th}}$ factor level and $\epsilon_{ij}$ is the underlying error variable, which we will assume has a normal distribution with zero mean and variance $\sigma^2$. Note that the variance does not depend on $i$ or $j$, so that regardless of the factor level at which the observation was taken, the variance of the response is assumed to be the same. Furthermore, we note that we can write this model in the form $Y = \mu + \epsilon$ if we consider the observations to be collected into a vector of the form

$$Y = (Y_{11}, Y_{12}, \ldots, Y_{1n_1}, Y_{21}, \ldots, Y_{2n_2}, Y_{31}, \ldots, Y_{k-1,n_{k-1}}, Y_{k1}, \ldots, Y_{kn_k}),$$

and similarly the mean vector as $\mu = (\mu_1, \ldots, \mu_1, \mu_2, \ldots, \mu_2, \ldots, \mu_k, \ldots, \mu_k)$, where each "block" of $\mu_i$'s is of the appropiate length, $n_i$.

There are, however, other useful ways of describing this model. For example, we might write:

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij},$$

where $\mu$ is an "overall" mean (i.e., an "average" of the expected responses at all the factor levels), and $\tau_i$ is the $i^{\text{th}}$ *level effect* (also sometimes called a *treatment effect*), which indicates how the expected response of the $i^{\text{th}}$ level deviates from the overall mean. Such a re-writing is often referred to as a *reparameterisation* of the model. There is a slight problem with this model structure; namely, we cannot uniquely determine the parameter values. To see this, we note that in terms of the previous model we have $\mu_i = \mu + \tau_i$; however, we can generate the same $\mu_i$'s from the new parameter values $\mu' = \mu + c$, $\tau_i' = \tau_i - c$ for any given value $c$, since:

$$\mu' + \tau_i' = (\mu + c) + (\tau_i - c) = \mu + \tau_i = \mu_i.$$

So, the two distinct models $Y_{ij} = \mu + \tau_i + \epsilon_{ij}$ and $Y_{ij} = \mu' + \tau_i' + \epsilon_{ij}$ are indistinguishable, meaning that no observed set of data would allow us to come up with a definitive estimate of the values for $\mu$ and $\tau_i$. This is therefore a situation in which the model is *overparameterised* (i.e., in the latter model there are $k+1$ parameters, but we have only $k$ different factor levels). To combat this problem, we must place a *constraint* on the $\tau_i$'s, so that there are only $k$ "effective" parameters. In other words, we require the parameters to satisfy a specific algebraic relationship, which means that once we select values for all but one of the parameters, the remaining value is not free for us to choose since it must be such that the final collection of parameters satisfies the desired relationship. As such, there are only $k$ "free" values among the parameters, since the constraint relationship causes the $(k+1)^{\text{st}}$ value to be "locked in" once the values of the other $k$ parameters are chosen. One common constraint is to require:

$$\sum_{i=1}^{k} n_i \tau_i = 0,$$

which coincides with our initial interpretation of the $\tau_i$'s as the deviations of the expected factor effects from the overall mean $\mu$, since this constraint implies that the expectation of the overall sample average is:

$$E(\overline{Y}) = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} E(Y_{ij}) = \frac{1}{n}\sum_{i=1}^{k} n_i\mu_i = \frac{1}{n}\sum_{i=1}^{k} n_i(\mu+\tau_i) = \mu + \frac{1}{n}\sum_{i=1}^{k} n_i\tau_i = \mu,$$

where $n = \sum_{i=1}^{k} n_i$ is the total sample size for the dataset. A useful consequence of this calculation is the relationship:

$$\mu = \frac{1}{n}\sum_{i=1}^{k} n_i\mu_i$$

showing that the "overall" mean in this parameterisation is the weighted average of the individual "within factor" expected values.

Another commonly used constraint is to require that $\tau_1 = 0$. Under this constraint, the parameters take on a somewhat different interpretation; namely, $\mu = \mu_1$ becomes the expected response for the first factor level (which thus becomes a sort of "baseline" or "control group" value) and the $\tau_i$'s for $i = 2, \ldots, k$ are then the deviations of the expected responses of the rest of the factor levels from the expected response of the first level. This last model structure

may seem the least appealing, however, it turns out that this structure will demonstrate the connection between the ANOVA model and multiple regression the most clearly.

ii. Parameter Estimation

For the first model parameterisation described, it should be clear that the best estimates of the $\mu_i$'s are just the averages

$$\hat{\mu}_i = \overline{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}.$$

[NOTE: In general, we will denote a generic estimator of a particular quantity by adding a circumflex to the symbol for the quantity being estimated, so that our estimator of $\mu_i$ is here denoted by $\hat{\mu}_i$.] Indeed, if we proceed using a least-squares approach, we would want to minimise the distance function

$$d(\mu_1, \ldots \mu_k) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \mu_i)^2,$$

and we would arrive at these "within factor level" averages for our estimators.

The least-squares estimates are unbiased:

$$E(\overline{Y}_i) = E\left(\frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}\right) = \frac{1}{n_i}\sum_{j=1}^{n_i} E(Y_{ij}) = \frac{1}{n_i}\sum_{j=1}^{n_i}\mu_i = \mu_i.$$

We also note that these are the so-called *maximum likelihood estimates* (or *MLE*s) of the parameters $\mu_i$ under our assumption of normality for the error vector $\epsilon$ (we will shortly discuss *MLE*s in more detail). Moreover, we can easily calculate the variances of these estimators as:

$$Var(\hat{\mu}_i) = Var(\overline{Y}_i) = \frac{\sigma^2}{n_i}.$$

Note that the precision of our estimator for the expected response of the $i^{\text{th}}$ factor level depends (as it must) on $n_i$, the number of data points sampled at this factor level. Therefore, when we gather our data, we should make sure to obtain sufficient data at each of the factor levels. For this reason, (as well reasons of mathematical simplicity and tractability) it is common practice to use *balanced* designs (i.e., all the $n_i$'s equal) whenever possible.

It is now straightforward to construct estimators of the parameters in the other model parameterisations. For example, in the first reparameterisation, we can estimate $\mu$ using the overall average $\hat{\mu} = \overline{Y}$ and thus the estimates of the $\tau_i$'s become

$$\hat{\tau}_i = \overline{Y}_i - \overline{Y}.$$

1st para : deviation from overall mean
2nd para: deviation from baseline mean.

Similarly, in the second reparameterisation, we would estimate $\mu$ by $\overline{Y}_1$ and then estimate each of the $\tau_i$'s by $\overline{Y}_i - \overline{Y}_1$. It is not hard to see that these estimates are all unbiased. The variances of the $\hat{\tau}_i$'s will be dealt with in the next subsection.

Finally, we need to estimate $\sigma^2$. To do this, we first note that the breakdown:

$$\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y})^2}_{SST} = \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{Y}_i-\overline{Y})^2}_{SSR} + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_i)^2}_{SSE} = \sum_{i=1}^{k} n_i(\overline{Y}_i-\overline{Y})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_i)^2$$

($k-1$; $n-k$)

holds. Indeed, the basic interpretations from multiple regression theory for each of the three pieces of this identity still apply. The left-hand side, $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y})^2$, is the total sum of

squares ($SST$) for the response and measures the total variability of the response variable. The first term on the right-hand side, $\sum_{i=1}^{k} n_i(\overline{Y}_i - \overline{Y})^2$, is the so-called *treatment* or *factor* sum of squares (which we will still denote by $SSR$ for reasons of consistency with previous regression notation) and it has $k-1$ degrees of freedom in this situation, since their are $k$ parameters in our model (note that while the latter two parameterisations actually have $k+1$ parameters, the linear constraint on the $\tau_i$'s implies that there are only $k$ effective parameters). The $SSR$ measures the variability *between* the factor levels, thus its mean square, $MSR = SSR/(k-1)$ is sometimes denote $s_b^2$.

Similarly, the second term on the right-hand side, $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)^2$, is the residual or error sum of squares ($SSE$) which measures the variability *within* the factor levels (and thus the $MSE$ in this setting is sometimes denoted by $s_w^2$). It is this term which we use to construct an estimate of the underlying error variance:

$$s^2 = \frac{SSE}{n-k} = MSE = s_w^2.$$

Note that we have divided the $SSE$ by its appropriate degrees of freedom here, $n-k$, ensuring that it is an unbiased estimator. The reason that the $SSE$ has $n-k$ degrees of freedom here again follows from the fact that our model has $k$ parameters, and we are thus simply following the general rule derived from multiple regression.

In addition, the $\overline{Y}_i$'s can be thought of as *fitted values*, $\hat{Y}_{ij}$ (since they are our best guess at the expected value of $Y_{ij}$) and thus we can define residuals as $e_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \overline{Y}_i$. Clearly, the $SSE$ is just the sum of the squared residuals as usual, and we can plot the residuals versus the fitted values and construct normal q-q plots as we did for regression to check our error structure assumptions of normality and homoscedasticity (note that we no longer have need of an assumption of "linearity"). The suitability of the homoscedasticity assumption can also be examined by comparing the values of the "within group" variance estimates,

$$s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)^2.$$

These $s_i^2$ values should all be similar if the data is truly homoscedastic. Exactly how "similar" they should be for us to feel comfortable with the homoscedasticity assumption is a subject of debate. Formal tests do exist, but they are generally thought to be unreliable, and a simple rule of thumb is to accept the assumption of homoscedasticity as reasonable provided that the ratio of the largest to the smallest "within group" variance is no more that about $k$, the number of factor levels of the categorical predictor being employed.

rule of thumb to detect homoscedasticity

iii. Confidence Intervals and Hypothesis Testing.

Once we have an estimate of $\sigma^2$, we can construct confidence intervals for the $\mu_i$'s, based on:

$$T = \frac{\overline{Y}_i - \mu_i}{s/\sqrt{n_i}} \sim t_{n-k}.$$

So, a $100(1-\alpha)\%$ confidence interval for the expected response at the $i^{\text{th}}$ level of the factor is:

$$\overline{Y}_i \pm t_{n-k}(1-\alpha/2)\frac{s}{\sqrt{n_i}},$$

CI for $\mu_i$

where $t_{n-k}(1-\alpha/2)$ is the $(1-\alpha/2)$-quantile of $t_{n-k}$, the Student's $t$-distribution with $n-k$ degrees of freedom.

Now, if we want to find confidence intervals for the $\tau_i$'s, we must first determine which constraint has been employed to define them, since this will change the interpretation (and thus the estimate) of these parameters. If we are using the second reparameterisation, in which the constraint is just $\tau_1 = 0$, then $\tau_i = \mu_i - \mu_1$, and

$$Var(\hat{\tau}_i) = Var(\overline{Y}_i - \overline{Y}_1) = Var(\overline{Y}_i) + Var(\overline{Y}_1) = \sigma^2\left(\frac{1}{n_i} + \frac{1}{n_1}\right), \qquad i = 2, \ldots, k$$

where we have used the fact that $\overline{Y}_{i_1}$ and $\overline{Y}_{i_2}$ are independent as long as $i_1 \neq i_2$, since distinct "within factor" averages are formed from disjoint (and therefore independent) subsets of the observed responses. Using this variance, a $100(1-\alpha)\%$ confidence interval for $\mu_i - \mu_1$ is:

$$(\overline{Y}_i - \overline{Y}_1) \pm t_{n-k}(1-\alpha/2)s\sqrt{\frac{1}{n_i} + \frac{1}{n_1}}.$$

Similarly, in the other parameterisation, where $\tau_i = \mu_i - \mu$, it can be shown that:

$$Var(\hat{\tau}_i) = Var(\overline{Y}_i - \overline{Y}) = \sigma^2\left(\frac{1}{n_i} - \frac{1}{n}\right),$$

so that a $100(1-\alpha)\%$ confidence interval for $\mu_i - \mu$ is:

CI for $\tau_i$

$$(\overline{Y}_i - \overline{Y}) \pm t_{n-k}(1-\alpha/2)s\sqrt{\frac{1}{n_i} - \frac{1}{n}}.$$

General Testing

In fact, we can find a $100(1-\alpha)\%$ confidence interval for any *linear combination* of the $\mu_i$'s, say $h_1\mu_1 + \ldots + h_k\mu_k$, for any vector of constants $h = (h_1, \ldots, h_k)$. Such a linear combination is often called a *contrast*. Using the independence of the $\overline{Y}_i$'s shows:

Contrast

$$Var\left(\sum_{i=1}^{k} h_i\overline{Y}_i\right) = \sum_{i=1}^{k} h_i^2 Var(\overline{Y}_i) = \sigma^2\sum_{i=1}^{k}\frac{h_i^2}{n_i},$$

and thus the desired interval would be

$$\left(\sum_{i=1}^{k} h_i\overline{Y}_i\right) \pm t_{n-k}(1-\alpha/2)s\sqrt{\sum_{i=1}^{k}\frac{h_i^2}{n_i}}.$$

The preceding results can also be used to test hypotheses of the form:

$$H_0 : \sum_{i=1}^{k} h_i\mu_i = c_0 \qquad \text{versus} \qquad H_A : \sum_{i=1}^{k} h_i\mu_i \neq c_0.$$

Using the test statistic:

$$T = \frac{\sum_{i=1}^{k} h_i\overline{Y}_i - c_0}{s\sqrt{\sum_{i=1}^{k}\frac{h_i^2}{n_i}}}$$

we would reject the null hypothesis at significance level $\alpha$ if $|T| > t_{n-k}(1-\alpha/2)$. theoretical

Perhaps more usefully, however, we can use the sum of squares breakdown to test the hypothesis that all the factor levels have the same mean response. The null hypothesis for such a test can be written as $H_0 : \mu_1 = \ldots = \mu_k$ or as $H_0 : \tau_1 = \ldots = \tau_k = 0$. Just as we did for the regression setting, we construct an ANOVA table

| Source | *df* | Sum of Squares | Mean Squares |
|---|---|---|---|
| Factor (or Treatment) | $k-1$ | $SSR = \sum_{i=1}^{k} n_i(\overline{Y}_i - \overline{Y})^2$ | $MSR = SSR/(k-1) = s_b^2$ |
| Residual (or Error) | $n-k$ | $SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)^2$ | $MSE = SSE/(n-k) = s_w^2$ |
| Total | $n-1$ | $SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y})^2$ | |

We then note that the $MSE$ is an unbiased estimate of $\sigma^2$ regardless of whether the null hypothesis is true or not, but that the $MSR$ is here a measure of how much of the variation in the response can be attributed to differences at each of the factor levels. Indeed, if we use the first of our possible constraints, $\sum_{i=1}^{k} n_i\tau_i = 0$, to define the $\tau_i$'s, then:

$$E(MSR) = \sigma^2 + \frac{1}{k-1}\sum_{i=1}^{k} n_i\tau_i^2,$$

which will tend to be much larger than $\sigma^2$ when the null hypothesis is not true. Therefore, we will formally reject $H_0$ at significance level $\alpha$ if the test statistic

$$F = \frac{MSR}{MSE} = \frac{s_b^2}{s_w^2},$$

exceeds the $(1-\alpha)$-quantile of the $F_{k-1,n-k}$-distribution, which is the distribution of the $F$-ratio under the assumption that the null hypothesis is true.

Generally, of course, we would conduct this overall hypothesis test initially, and if we reject $H_0$ we would then begin to examine the individual $\mu_i$'s to discover how the individual level effects were related to one another. In addition, we should note the problems of simultaneous inference. As in the regression setting, if we want to construct confidence intervals for several parameters (or linear combinations of the parameters) we must take into account the fact that the overall simultaneous confidence level that all the intervals actually contain the values they are estimating will generally be smaller than the confidence levels for each of the individual intervals separately. However, it is a simple procedure to employ the Bonferroni (or some similar) method for expanding the individual intervals to ensure that the overall confidence is at least as high as desired. The Bonferroni method is based on the probability inequality:

$$Pr(A_1 \cap A_2 \cap \cdots \cap A_g) \geq 1 - \sum_{i=1}^{g} Pr(A_i^c).$$

Using this inequality, we can see that if each of the events $A_i$ have probability $1-\alpha/g$, then we have $Pr(A_i^c) = \alpha/g$ and

$$Pr(A_1 \cap A_2 \cap \cdots \cap A_g) \geq 1 - \sum_{i=1}^{g} \alpha/g = 1 - \alpha.$$

In other words, we can see that if we want the simultaneous confidence of $g$ confidence intervals to be at least $1-\alpha$, then we can guarantee this is the case if we use individual confidence intervals which each have confidence level $1-\alpha/g$.

*Example 1 - Corn Yield Data:* To investigate the effect of various fertilisers on corn yield, a farmer divided a field into 24 plots of the same size and shape and applied one of four fertiliser treatments to the plots, each treatment being used for six plots. The four treatments applied were: no fertiliser (i.e., a control); $K_2O + N$ (i.e., potassium oxide and nitrogen); $K_2O + P_2O_5$

(i.e., potassium oxide and phosphorus oxide); and, $N + P_2O_5$ (i.e., nitrogen and phosphorus oxide). The yield from each plot (in bushels per acre) was:

| Treatment | Yields | | | | | | "Within Factor" Averages |
|---|---|---|---|---|---|---|---|
| Control: | 99 | 40 | 61 | 72 | 76 | 84 | 72 |
| $K_2O + N$: | 96 | 84 | 82 | 104 | 99 | 105 | 95 |
| $K_2O + P_2O_5$: | 63 | 57 | 81 | 59 | 64 | 72 | 66 |
| $N + P_2O_5$: | 79 | 92 | 91 | 87 | 78 | 71 | 83 |
| "Overall" average: | 79 | | | | | | |

Using *S-Plus* we can calculate the ANOVA table:

```
> corn <- read.table("Corn.txt",header=T)
> attach(corn)
> names(corn)
[1] "yield" "fert"
> corn.aov <- aov(yield ~ fert, data=corn)
> summary(corn.aov)
            Df Sum of Sq Mean Sq F Value      Pr(F)
fert         3      2940   980.0 5.99022 0.00438693
Residuals 20      3272   163.6
```

So, there is clearly a significant difference in the level effects, since the $p$-value for the test of $H_0 : \mu_1 = \ldots = \mu_4$ is 0.0044. The task which then remains is to find specifically how the level effects differ from one another. Now, this is clearly a situation in which the second reparameterisation of our model is appropriate (i.e., using the constraint that $\tau_1 = 0$) since the first level of the factor is a control group. One quantity in which we might initially be interested would be the averaged effect of any fertiliser on the corn yield; in other words, we might want to find a confidence interval for the quantity:

$$\frac{\tau_2 + \tau_3 + \tau_4}{3} = \frac{(\mu_2 - \mu_1) + (\mu_3 - \mu_1) + (\mu_4 - \mu_1)}{3} = \frac{\mu_2 + \mu_3 + \mu_4}{3} - \mu_1.$$

We can construct a 95% confidence intervals for this quantity as:

```
> lvl.mns <- tapply(yield,fert,mean)
> ni <- tapply(yield,fert,length)
> lvl.mns
 Control K2O+N K2O+P2O5 N+P2O5
      72    95       66     83
> h <- c(-1,1/3,1/3,1/3)
```

[NOTE: Make sure that you put the $h_i$ values in the appropriate order by checking how *S-Plus* has ordered the level means.]

```
> est <- t(h)%*%lvl.mns
> MSE <- sum((yield-fitted(corn.aov))^2)/corn.aov$df.residual
> sd <- sqrt(MSE)*sqrt(sum((h^2)/ni))
> upper <- est+qt(0.975,corn.aov$df.residual)*sd
> lower <- est-qt(0.975,corn.aov$df.residual)*sd
> c(lower,est,upper)
[1] -3.244102   9.333333  21.910768
```

Interestingly, this confidence interval contains zero, indicating that on average the fertilisers have no significant effect. Of course, our initial result indicated that at least one of the fertilisers

must have a significant effect, so this new result indicates that some of the fertilisers have no effect or that the effect of some fertilisers is negative. To investigate this further, we construct a simultaneous 95% confidence region for the three level effects $\tau_2 = \mu_2 - \mu_1$, $\tau_3 = \mu_3 - \mu_1$ and $\tau_4 = \mu_4 - \mu_1$ using the Bonferroni method.

```
> h1 <- c(-1,1,0,0)
> h2 <- c(-1,0,1,0)
> h3 <- c(-1,0,0,1)
> sd1 <- sqrt(MSE)*sqrt(sum((h1^2)/ni))
> sd2 <- sqrt(MSE)*sqrt(sum((h2^2)/ni))
> sd3 <- sqrt(MSE)*sqrt(sum((h3^2)/ni))
> t2hat <- t(h1)%*%lvl.mns
> t3hat <- t(h2)%*%lvl.mns
> t4hat <- t(h3)%*%lvl.mns
> ests <- c(t2hat,t3hat,t4hat)
> sds <- c(sd1,sd2,sd3)
> uppers <- ests+qt(1-0.05/6,corn.aov$df.residual)*sds
> lowers <- ests-qt(1-0.05/6,corn.aov$df.residual)*sds
> ints <- cbind(uppers,ests,lowers)
> ints
      uppers ests     lowers
[1,] 42.29308   23   3.706922
[2,] 13.29308   -6 -25.293078
[3,] 30.29308   11  -8.293078
```

Thus, it appears that the first fertiliser is the only one which is significantly enhancing corn yields. However, the third fertiliser is not too far from significant, and we note that the first and third fertilisers are the ones which contained nitrogen. So, we might lastly wish to test whether fertilisers with nitrogen significantly increase yield over fertilisers without nitrogen, and we are thus interested in the linear combination of parameters:

$$\frac{\tau_2 + \tau_4}{2} - \tau_3 = \frac{(\mu_2 - \mu_1) + (\mu_4 - \mu_1)}{2} - (\mu_3 - \mu_1) = \frac{\mu_2 + \mu_4}{2} - \mu_3.$$

```
> h <- c(0,1/2,-1,1/2)
> est <- t(h)%*%lvl.mns
> sd <- sqrt(MSE)*sqrt(sum((h^2)/ni))
> upper <- est+qt(0.975,corn.aov$df.residual)*sd
> lower <- est-qt(0.975,corn.aov$df.residual)*sd
> c(lower,est,upper)
[1] 9.659615 23.000000 36.340385
```

So, it does appear that fertilisers with nitrogen will significantly increase corn yields.

iv. Indicator Variables and Regression

Suppose that the categorical predictor in a one-way ANOVA model has only two levels (which would indicate that we were simply in a situation where a two-sample $t$-test was appropriate). In this case, there are only two parameters (excluding $\sigma^2$), $\mu_1$ and $\mu_2$. If we employ the reparameterisation based on our second constraint (i.e., that $\tau_1 = 0$), then for data values at the first level of the factor, the model shows:

$$Y_{1j} = \mu + \epsilon_{1j} = \beta_0 + \epsilon_{1j}$$

while for data values at the second level of the factor, the model shows:

$$Y_{2j} = \mu + \tau_2 + \epsilon_{2j} = \beta_0 + \beta_1 + \epsilon_{2j},$$

where we have simply used a re-naming of the parameters ($\mu$ renamed to $\beta_0$ and $\tau_2$ renamed to $\beta_1$) to help demonstrate the connections to multiple regression. Indeed, the above descriptions show that we can write the overall model as:

$$Y_{ij} = \beta_0 + \beta_1 z_{ij} + \epsilon_{ij},$$

where $z$ is now a *numerical* (i.e., continuous) predictor which has a value of zero for data points at the first level of the factor and a value of one for data points at the second factor level, i.e.,

$$z_{ij} = \begin{cases} 0 & \text{if } i = 1 \\ 1 & \text{if } i = 2 \end{cases}.$$

So, we see how numerical codes can be interpreted as continuous predictors in order to show the connection between regression and ANOVA models with factors having only two levels. Suppose now that the factor has three levels. Then, again using the second constraint for our reparameterisation, the breakdown of the model can be seen as:

$$\begin{aligned} Y_{1j} &= \mu + \epsilon_{1j} = \beta_0 + \epsilon_{1j} \\ Y_{2j} &= \mu + \tau_2 + \epsilon_{2j} = \beta_0 + \beta_1 + \epsilon_{2j} \\ Y_{3j} &= \mu + \tau_3 + \epsilon_{3j} = \beta_0 + \beta_2 + \epsilon_{3j}. \end{aligned}$$

Now, if we try to employ the same connection between ANOVA and regression that we used in the two-level case, we might attempt to write our model as:

$$Y_{ij} = \beta_0 + \beta_1 z_{ij} + \epsilon_{ij},$$

where the predictor $z$ is now defined as, say:

$$z_{ij} = \begin{cases} 0 & \text{if } i = 1 \\ 1 & \text{if } i = 2 \\ 2 & \text{if } i = 3 \end{cases}.$$

not a good indicator assignment

Unfortunately, this does not work, since it would imply that for responses observed at the third factor level, the expression:

$$Y_{3j} = \beta_0 + \beta_1 z_{3j} + \epsilon_{3j} = \beta_0 + 2\beta_1 + \epsilon_{3j},$$

must be equivalent to $Y_{3j} = \beta_0 + \beta_2 + \epsilon_{3j}$, which clearly cannot be true except in the special case that $\beta_2 = 2\beta_1$, and this certainly cannot be guaranteed for all populations. So, we cannot connect a three-level factor to a regression with one predictor. However, a closer inspection reveals that we can write the model as:

$$Y_{ij} = \beta_0 + \beta_1 z_{1,ij} + \beta_2 z_{2,ij} + \epsilon_{ij},$$

add a new term

where $z_1$ and $z_2$ are defined as:

$$z_{1,ij} = \begin{cases} 1 & \text{if } i = 2 \\ 0 & \text{if } i \neq 2 \end{cases} \quad \text{and} \quad z_{2,ij} = \begin{cases} 1 & \text{if } i = 3 \\ 0 & \text{if } i \neq 3 \end{cases}.$$

In fact, for the general $k$-level factor, we can write the ANOVA model (under the second reparameterisation constraint that $\tau_1 = 0$) as a multiple regression model with $k-1$ such predictors, generally call *indicator* or *dummy* variables.

In general then, a one-way ANOVA on a factor having $k$ levels is equivalent to a multiple regression of the form:

$$Y = \beta_0 + \beta_1 z_1 + \ldots + \beta_{k-1} z_{k-1} + \epsilon,$$

where $z_l$ $(l = 1, \ldots, k-1)$ is an indicator variable which takes the value zero for all data points except those at the $(l+1)^{\text{st}}$ level of the factor, where it takes the value one; i.e.,

$$z_{l,ij} = \begin{cases} 1 & \text{if } i = l+1 \\ 0 & \text{if } i \neq l+1 \end{cases}.$$

Finally, note that we can write this model in matrix notation as $Y = X\beta + \epsilon$ where the rows of the design matrix $X$ are of the form $(1, 0, \ldots, 0, 1, 0, \ldots, 0)$, with the second one residing in the column corresponding to the indicator variable which takes on the value one for the appropriate factor level of the data point to which the row refers. [NOTE: It is possible to use variants of the indicator variables to connect multiple regression with the ANOVA models under different constraints on the $\tau_i$'s. Indeed, *S-Plus* sometimes uses a different default set of constraints, so that the regression parameter estimates do not have the interpretation we have employed here. Of course, the overall tests of regression significance will be the same regardless of the parameterisation.]

*Example 1 (cont'd) - Corn Yield Data:* All of the preceding calculations done in *S-Plus* using the `aov()` function can also be performed using the `lm()` function:

```
> fert1 <- ifelse(fert=="K2O+N",1,0)
> fert2 <- ifelse(fert=="K2O+P2O5",1,0)
> fert3 <- ifelse(fert=="N+P2O5",1,0)
> ferts <- cbind(fert1,fert2,fert3)
> corn.lm <- lm(yield ~ ferts)
> anova(corn.lm)
Analysis of Variance Table

Response:  yield

Terms added sequentially (first to last)

          Df Sum of Sq Mean Sq F Value      Pr(F)
ferts      3      2940   980.0 5.99022 0.00438693
Residuals 20      3272   163.6
> corn.lm

Call:
lm(formula = yield ~ ferts)

Coefficients:
 (Intercept) fertsfert1 fertsfert2 fertsfert3
          72         23         -6         11

Degrees of freedom:  24 total; 20 residual
Residual standard error:  12.79062
```

In addition, the `factor()` function in *S-Plus* will alleviate the need to create the indicator variables ourselves:

```
> corn.lm <- lm(yield ~ factor(fert))
```

```
> anova(corn.lm)
Analysis of Variance Table

Response:  yield

Terms added sequentially (first to last)
             Df Sum of Sq Mean Sq F Value      Pr(F)
factor(fert)  3      2940   980.0 5.99022 0.00438693
Residuals    20      3272   163.6
> corn.lm

Call:
lm(formula = yield ~ factor(fert))

Coefficients:
 (Intercept) factor(fert)1 factor(fert)2 factor(fert)3
          79          11.5     -5.833333      1.333333

Degrees of freedom:  24 total; 20 residual
Residual standard error:  12.79062
```

Note that the `factor()` function does not use either of our two constraints, but instead uses a third, more complicated constraint. Thus, while the overall ANOVA table will still be what we expect, we cannot interpret the parameter estimates in the same way.

## III. Analysis of Covariance Models

Once we have seen how indicator variables can be used to incorporate categorical predictors into the standard regression framework, it is not a difficult step to expand one-way ANOVA models by including other continuous predictors into the model. In this section, we will still focus on the case where we have only a single categorical predictor, however, the case of several categorical predictors is not substantially different and will be dealt with in the final section of this chapter.

### i. Model and Assumptions

Suppose that we have data on a response variable, $Y_{ij}$, and a categorical predictor, $f_i$, as well as on a continuous predictor $x_{ij}$. The simplest linear model structure for such a setting is the so-called *ANalysis of COVAriance* (or *ANCOVA*) model:

$$Y_{ij} = \mu_i + \beta x_{ij} + \epsilon_{ij},$$

where the $\mu_i$'s are the effects on the response corresponding to the factor levels and $\beta$ is the slope of the relationship between the response and the continuous predictor $x$. In other words, $\mu_i$ is the expected response value for individuals at the $i^{\text{th}}$ level of the factor when $x = 0$. Similarly, $\beta$ represents the expected change in the response for a unit change in the predictor $x$ within any of the factor levels. Since the slope of the relationship between the response and the continuous predictor is assumed to be the same within each level of the factor, ANCOVA models are sometimes referred to as *parallel regression* models, the level effects for the factor simply corresponding to the different intercept terms for each of the parallel regression lines fit within each level of the factor. We also note that the same reparameterisation concerns which were discussed for ANOVA models are equally relevant for ANCOVA models, so that we will generally re-write our model as:

$$Y_{ij} = \mu + \tau_i + \beta x_{ij} + \epsilon_{ij},$$

where we must employ one of our constraints on the $\tau_i$'s so as not to overparameterise the model.

Now, we have already seen how indicator variables can be used in a one-way ANOVA model, and the same treatment can be employed for ANCOVA models. Thus, if the factor $f_i$ has $k$ levels, then we can employ the constraint $\tau_1 = 0$ and use $k-1$ indicator variables to re-write the ANCOVA model as:

$$Y_{ij} = \mu + \tau_i + \beta x_{ij} + \epsilon_{ij} = \beta_0 + \beta_1 z_{1,ij} + \ldots + \beta_{k-1} z_{k-1,ij} + \beta_k x_{ij} + \epsilon_{ij},$$

where $z_{l,ij}$ is just the appropriate indicator variable for the $(l+1)^{\text{st}}$ factor level, $\beta_0 = \mu$, $\beta_k = \beta$ and $\beta_l$ is just a renaming of $\tau_{l+1}$, $l = 1, \ldots, k-1$.

Of course, as with all our linear models, we will assume that the error terms, $\epsilon_{ij}$, are independent, homoscedastic and standard normally distributed with zero mean and common variance $\sigma^2$.

ii. Parameter Estimation

Using our indicator variables, we can write our ANCOVA model in regression form as $Y = X\beta + \epsilon$, where $X$ is the design matrix, with rows of the form $(1, 0, \ldots, 0, 1, 0, \ldots, 0, x_{ij})$, the internal "1" residing in the column appropriate to the indicator variable associated with the factor level at which the response value was observed. Once we have written the model in this form, it is clear that the least-squares (as well as the maximum likelihood) estimator of the $\beta$ vector is just the usual

$$b = (X^T X)^{-1} X^T Y.$$

In fact, once we have made the connection between the ANCOVA model and a multiple regression model, we can simply employ all of our standard multiple regression techniques for conducting predictions and hypothesis tests. For this reason, we here only briefly highlight how some of the standard regression formulae reduce in the case of a simple ANCOVA model. Specifically, we note that the general least-squares estimator can be shown to yield:

$$\begin{aligned} b_0 &= \overline{Y}_1 - b_k \overline{x}_1 \\ b_1 = (\overline{Y}_2 - \overline{Y}_1) - b_k(\overline{x}_2 - \overline{x}_1), \ldots, b_{k-1} &= (\overline{Y}_k - \overline{Y}_1) - b_k(\overline{x}_k - \overline{x}_1) \\ b_k &= \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_i)(Y_{ij} - \overline{Y}_i)}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_i)^2}, \end{aligned}$$

where $\overline{x}_i$ is the average of the continuous predictor values for data points within the $i^{\text{th}}$ level of the factor. Note that since $\mu_i = \tau_i + \mu = \beta_{i-1} + \beta_0$, this shows that:

$$\hat{\mu}_i = b_{i-1} + b_0 = (\overline{Y}_i - \overline{Y}_1) - b_k(\overline{x}_i - \overline{x}_1) + \overline{Y}_1 - b_k \overline{x}_1 = \overline{Y}_i - b_k \overline{x}_i \qquad (i = 2, \ldots, k).$$

These estimates are, of course, unbiased and the variance-covariance matrix of $b$ is given by $\sigma^2 (X^T X)^{-1}$. In particular, this yields:

$$Var(b_k) = \sigma^2 \left\{ \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_i)^2 \right\}^{-1},$$

from which the variances of the other parameter estimates can be derived using the fact that $b_k$ can be shown to be independent of the $\overline{Y}_i$'s, so that:

$$Var(b_0) = \sigma^2\left\{\frac{1}{n_1} + \frac{\overline{x}_1^2}{\sum_{i=1}^k\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2}\right\}$$
$$Var(b_l) = \sigma^2\left\{\frac{1}{n_{l+1}} + \frac{1}{n_1} + \frac{(\overline{x}_{l+1}-\overline{x}_1)^2}{\sum_{i=1}^k\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2}\right\} \qquad (l = 1,\ldots,k-1)$$
$$Var(\hat{\mu}_l) = \sigma^2\left\{\frac{1}{n_l} + \frac{\overline{x}_l^2}{\sum_{i=1}^k\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2}\right\} \qquad (l = 2,\ldots,k).$$

Using these basic results, we can construct all of our usual hypothesis tests, confidence intervals and predictions. In particular, we can again construct confidence intervals and hypothesis tests for the contrasts $\sum_{i=1}^k h_i\mu_i$ based on the estimator

$$\sum_{i=1}^k h_i\hat{\mu}_i = \sum_{i=1}^k h_i(\overline{Y}_i - b_k\overline{x}_i) = \sum_{i=1}^k h_i\overline{Y}_i - b_k\sum_{i=1}^k h_i\overline{x}_i,$$

which has variance:

$$Var\left(\sum_{i=1}^k h_i\hat{\mu}_i\right) = \sigma^2\sum_{i=1}^k\frac{h_i^2}{n_i} + \left(\sum_{i=1}^k h_i\overline{x}_i\right)^2 Var(b_k) = \sigma^2\left\{\sum_{i=1}^k\frac{h_i^2}{n_i} + \frac{\left(\sum_{i=1}^k h_i\overline{x}_i\right)^2}{\sum_{i=1}^k\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2}\right\},$$

and we can estimate this variance by replacing $\sigma^2$ by the observed value of the $MSE$.
Of course, if we write our linear combination in terms of the $\beta$ vector, then we can always use the multiple regression fact that

$$Var(c^Tb) = \sigma^2c^T(X^TX)^{-1}c.$$

Finally, we can easily generalise our simple ANCOVA model to include several continuous predictors and then even employ our model selection techniques. Note, however, that the specific formulae for the least-squares estimators and their variances become much more complicated, and we must rely on the general forms derived from our multiple regression discussions.

*Example 2 - Teacher Effectiveness Data:* A study of 23 student teachers was designed to investigate what factors are important in teaching effectiveness. Twelve male and eleven female student teachers were evaluated and given an overall effectiveness score. In addition, each participating student teacher was given four standardised tests, and their scores were recorded for use as the predictor variables. The model that we wish to fit is:

$$Y_i = \beta_0 + \beta_1z_i + \beta_2x_{1i} + \beta_3x_{2i} + \beta_4x_{3i} + \beta_5x_{4i} + \epsilon_i,$$

where $z_i$ is the indicator variable associated with gender. Notice that we drop the double subscripting (i.e., $Y_{ij}$) when we write our models in multiple regression format. Suppose we are interested in whether gender has an effect on the teacher effectiveness ratings (which might indicate some gender bias in the evaluation procedure). To fit this model in *S-Plus*:

```
> tchreff <- read.table("teachEff.txt",header=T)
> attach(tchreff)
> names(tchreff)
[1] "teff"   "gender" "x1"     "x2"     "x3"     "x4"
```

```
> tchreff.lm <- lm(teff ~ x1 + x2 + x3 + x4 + factor(gender))
> anova(tchreff.lm)
Analysis of Variance Table

Response:  teff

Terms added sequentially (first to last)
               Df Sum of Sq  Mean Sq  F Value     Pr(F)
x1              1  31463.31 31463.31 10.40721 0.0049617
x2              1  14064.48 14064.48  4.65215 0.0456228
x3              1  22488.27 22488.27  7.43851 0.0143316
x4              1  42355.56 42355.56 14.01007 0.0016193
factor(gender)  1    881.34   881.34  0.29152 0.5962485
Residuals      17  51394.78  3023.22
```

So, it appears that gender is not a significant factor in explaining the variation in teaching efficiency. However, this may be due to a poorly chosen model. So, using the model selection procedures developed for general regression models, we can construct the following table:

| Model | $p$ | $s_p^2$ | $R_a^2$ | $PRESS_p$ | $C_p$ |
|---|---|---|---|---|---|
| $gndr$ | 2 | 7537 | -0.020 | 190595 | 33.356 |
| $gndr + x_1$ | 3 | 6528 | 0.117 | 165737 | 26.188 |
| $gndr + x_2$ | 3 | 6718 | 0.091 | 171696 | 27.445 |
| $gndr + x_3$ | 3 | 7274 | 0.016 | 192406 | 31.120 |
| $gndr + x_4$ | 3 | 5330 | 0.279 | 140434 | 18.261 |
| $gndr + x_1 + x_2$ | 4 | 6124 | 0.172 | 156444 | 23.486 |
| $gndr + x_1 + x_3$ | 4 | 5468 | 0.260 | 143321 | 19.365 |
| $gndr + x_1 + x_4$ | 4 | 4954 | 0.330 | 138254 | 16.133 |
| $gndr + x_2 + x_3$ | 4 | 6503 | 0.120 | 172834 | 25.870 |
| $gndr + x_2 + x_4$ | 4 | 3336 | 0.549 | 88544 | 5.962 |
| $gndr + x_3 + x_4$ | 4 | 5179 | 0.300 | 139945 | 17.546 |
| $gndr + x_1 + x_2 + x_3$ | 5 | 5230 | 0.293 | 140236 | 18.141 |
| $gndr + x_1 + x_2 + x_4$ | 5 | 3377 | 0.543 | 95535 | 7.108 |
| $gndr + x_1 + x_3 + x_4$ | 5 | 4280 | 0.421 | 122882 | 12.482 |
| $gndr + x_2 + x_3 + x_4$ | 5 | 3223 | 0.564 | 86473 | 6.188 |
| $full$ | 6 | 3023 | 0.591 | 85057 | 6.000 |

This table indicates that a good model to look at might be:

$$Y = \beta_0 + \beta_1 z + \beta_3 x_2 + \beta_5 x_4 + \epsilon,$$

where $z$ is the indicator variable for gender, since this model has one of the largest $R_a^2$ values, one of the lowest prediction sum-of-squares, a Mallows' $C_p$ value close to $p$ and is parsimonious in that it has only 4 parameters. We might also try a forward stepwise regression:

```
> unique(gender)
[1] M F
> gndr <- ifelse(gender=="M",0,1)
> tchreff.all <- cbind(gndr,x1,x2,x3,x4,teff)
> forstp <- stepwise(tchreff.all[,1:5],tchreff.all[,6],f.crit=c(3.5,3))
> forstp$which
      gndr x1 x2 x3 x4
1(+5)    F  F  F  F  T
2(+3)    F  F  T  F  T
```

which arrives at the suggested model. So, fitting this model we find:

```
> anova(lm(teff ~ x2 + x4 + factor(gender)))
Analysis of Variance Table

Response:  teff

Terms added sequentially (first to last)
                Df Sum of Sq  Mean Sq  F Value     Pr(F)
x2               1  26605.58 26605.58  7.97634 0.0108354
x4               1  68084.02 68084.02 20.41155 0.0002354
factor(gender)   1   4582.44  4582.44  1.37381 0.2556456
Residuals       19  63375.70  3335.56
```

which still indicates that gender is not a significant factor, though its $p$-value is now smaller.

iii. Interaction and Non-parallel Regressions

IS IT REALLY PARALLEL?

In the preceding discussion, we have assumed that the slope of the relationship between the response and the continuous predictors is the same within each level of the factor. However, as with all assumptions, we would like to be able to test the plausibility of this model structure. To accomplish this, we need to expand our *parallel regression* ANCOVA model to include the possibility of different slopes within each factor level. Once we have included extra terms in the model to allow for the possibility of non-parallel slopes, we can then use our standard sum of squares approach to test whether these extra terms are significantly adding to the explanation of the response variation.

Suppose that we have a factor with only two levels, and we want to incorporate the possibility of *non-parallel* regression lines within the two factor levels. The model is then:

$$Y_{1j} = \mu_1 + \gamma_1 x_{1j} + \epsilon_{1j} = \beta_0 + \beta_2 x_{1j} + \epsilon_{1j},$$

for data points at the first factor level, and

$$Y_{2j} = \mu_2 + \gamma_2 x_{2j} + \epsilon_{2j} = \mu_1 + \tau_2 + (\gamma_1 + \delta_2) x_{2j} + \epsilon_{2j} = \beta_0 + \beta_1 + \beta_2 x_{2j} + \beta_3 x_{2j} + \epsilon_{2j},$$

for data at the second factor level. Such a model structure is often referred to as a *non-parallel regressions* ANCOVA. Note that, as with a one-way ANOVA, we have reparameterised the model so that within the second factor level $\tau_2 = \beta_1$ is the difference in the intercepts and $\delta_2 = \gamma_2 - \gamma_1 = \beta_3$ is the difference in the slopes within the two factor levels. Clearly, we can use an indicator variable for the second factor level, $z_{ij}$, to represent the different intercepts, but we need to do the same for the slopes. A little inspection shows that we can make use of the same indicator and write the model as:

$$Y_{ij} = \beta_0 + \beta_1 z_{ij} + \beta_2 x_{ij} + \beta_3 (z_{ij} x_{ij}) + \epsilon_{ij}.$$

Of course, a single indicator variable will not suffice for the case of a factor with $k > 2$ levels. However, just as for the ANOVA, we can write our ANCOVA model in such cases using $k - 1$ indicator variables:

$$Y_i = \beta_0 + \beta_1 z_{1i} + \ldots + \beta_{k-1} z_{(k-1)i} + \beta_k x_i + \beta_{k+1} z_{1i} x_i + \ldots + \beta_{2k-1} z_{(k-1)i} x_i + \epsilon_i,$$

where we have again moved to the single subscripting notation for convenience, and the predictor $z_l$ is the indicator variable for the $(l+1)^{\text{st}}$ factor level. In matrix notation, we can write the model as:

$$Y = X\beta + \epsilon = \underline{1}_n \beta_0 + Z\beta_{(1)} + x\beta_k + Z_x \beta_{(2)} + \epsilon,$$

where $x = (x_1, \ldots, x_n)^T$ is the column vector containing the values of the continuous predictor, $Z$ is the appropriate matrix of indicator values (i.e., the design matrix for the ANOVA model, ignoring the continuous predictor and the intercept columns), $Z_x$ is an $n \times (k-1)$ matrix with $l^{\text{th}}$ column of the form $(z_{l1}x_1, \ldots, z_{ln}x_n)^T$ (which can be constructed using the *S-Plus* multiplication `Zx <- Z*x`), $\underline{1}_n$ is a column vector of length $n$ in which all the elements are 1, $\beta_{(1)} = (\beta_1, \ldots, \beta_{k-1})^T$ and $\beta_{(2)} = (\beta_{k+1}, \ldots, \beta_{2k-1})^T$. In other words, we have partitioned the model into four distinct pieces, the first containing the baseline intercept, the second containing the "intercept" effects within each of the factor levels (i.e., the difference in intercepts from the baseline level), the third containing the slope within the first factor level and the fourth containing the differences in slopes among the remaining factor levels. Once we have written the model this way, of course, we can again simply employ all of our standard multiple regression techniques for inference, analysis and diagnostics.

However, there is one particular hypothesis test which is clearly of central interest. This is the so-called test for *additivity* or for parallel regressions:

$$H_0 : \beta_{(2)} = 0 \qquad \text{versus} \qquad H_A : \beta_{(2)} \neq 0.$$

The focus here is to see whether the *interaction* or *multiplicative* terms (which are so named because they allow the effects of the factor levels to modify, that is interact with, the effect of the continuous predictor) can be removed from the model, implying that the slope of the relationship between the response and the continuous predictor is the same at each level of the factor (i.e., the regression lines within each factor level are parallel). The appropriate $F$-statistic for this test is simply:

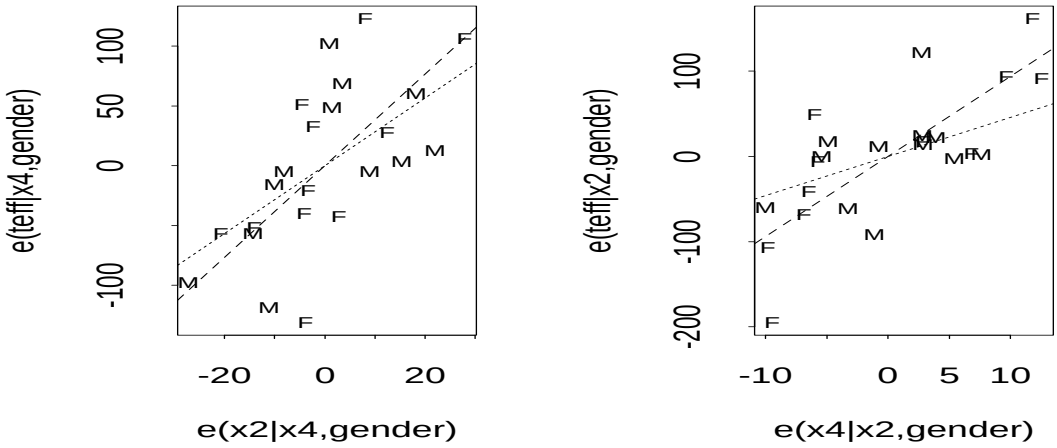
$$F = \frac{SSR(\beta_{(2)}|\beta_{(1)}, \beta_k)/(k-1)}{MSE_{full}},$$

which has an $F$-distribution with $k-1$ numerator and $n-\{k+1+(k-1)\} = n-2k$ denominator degrees of freedom as our standard regression approach would indicate.

Extension of this model to the case of several continuous predictor variables is quite straightforward, simply adding the appropriate multiplicative terms between the indicator variables and any continuous predictors for which non-parallel slopes are desired. Finally, before moving on to an example of this procedure, we note that an ANCOVA model with a $k$-level factor and $c$ continuous predictors has $k + c$ parameters, and if we add in the possibility of non-parallel slopes for $c'$ of the continuous predictors, the model now has $k + c + c'(k-1)$ parameters (i.e., an additional $k-1$ parameters for each predictor with non-parallel slopes), which can become quite large, and thus we cannot realistically expect reliable results unless we have a reasonably large dataset (indeed, recall that we require the sample size to be greater than the total number of parameters in our model in order to have enough information to estimate all the model parameters in any multiple regression).

*Example 2 (cont'd) - Teaching Efficiency Data:* In the parallel regression ANCOVA analysis of the teaching efficiency data, we saw that gender did not appear to play a significant role. However, it is possible that the assumption of equal slopes for males and females may have covered a possible relationship between efficiency scores and gender. One simple way to investigate whether a non-parallel regressions ANCOVA model should be employed is to examine added variable plots for each continuous predictor, employing a different plotting symbol for data points within each of the different factor levels. If it appears that there are different slopes among the different groups of symbols, then it may be worth fitting the larger model

and testing whether or not there is a significant effect due to non-parallel slopes (assuming of course that there is sufficient data to fit the extra parameters into the model). For the teaching efficiency data, we have seen that the second and fourth continuous predictors were the most important ones, and so we could create added variable plots as:




```
> rsd1 <- residuals(lm(x2 ~ factor(gender) + x4))
> rsd2 <- residuals(lm(teff ~ factor(gender) + x4))
> plot(rsd1,rsd2,type="n",xlab="e(x2|x4,gender)", ylab="e(teff|x4,gender)")
> points(rsd1[gender=="M"],rsd2[gender=="M"],pch="M",cex=0.7)
> points(rsd1[gender=="F"],rsd2[gender=="F"],pch="F",cex=0.7)
> abline(lsfit(rsd1[gender=="M"],rsd2[gender=="M"])$coef,lty=2)
> abline(lsfit(rsd1[gender=="F"],rsd2[gender=="F"])$coef,lty=3)
> rsd1 <- residuals(lm(x4 ~ factor(gender) + x2))
> rsd2 <- residuals(lm(teff ~ factor(gender) + x2))
> plot(rsd1,rsd2,type="n",xlab="e(x4|x2,gender)", ylab="e(teff|x2,gender)")
> points(rsd1[gender=="M"],rsd2[gender=="M"],pch="M",cex=0.7)
> points(rsd1[gender=="F"],rsd2[gender=="F"],pch="F",cex=0.7)
> abline(lsfit(rsd1[gender=="M"],rsd2[gender=="M"])$coef,lty=2)
> abline(lsfit(rsd1[gender=="F"],rsd2[gender=="F"])$coef,lty=3)
```

These plots do not seem to indicate any need for the use of non-parallel regressions, as the fitted lines associated with males and females appear to have quite similar slopes (though perhaps the right-hand plot has some small cause for concern). However, we could confirm this with a formal statistical test as follows:

```
> Zx2 <- gndr*x2
> Zx4 <- gndr*x4
> Zx <- cbind(Zx2,Zx4)
> eff.lm <- lm(teff ~ x2 + x4 + gndr + Zx)
> anova(eff.lm)
Analysis of Variance Table

Response:  teff

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value     Pr(F)
x2         1  26605.58 26605.58  7.78668 0.0125548
x4         1  68084.02 68084.02 19.92620 0.0003411
gndr       1   4582.44  4582.44  1.34115 0.2628433
```

```
Zx          2   5289.96  2644.98  0.77411 0.4767027
Residuals 17  58085.74  3416.81
```

As was suggested by the plots, we can now confirm that the `Zx` term is non-significant and we would accept the null hypothesis that all the $\beta$'s associated with these predictors are zero (i.e., that all the slopes within each of the factor levels are the same). If it had turned out that the `Zx` term was indeed significant, we would then have had to construct a more detailed analysis of variance table to examine the individual effects of `Zx2` and `Zx4` separately.

Finally, notice that this analysis of variance table shows that the sums of squares and mean squares for the first three rows (i.e., the rows associated with the model terms `x2`, `x4` and `gndr`) are the same as they were for our parallel regression ANCOVA analysis. Moreover, the sum of the two remaining entries (i.e., those for `Zx` and `Residuals`) is equal to the residual sum of squares from the parallel regression ANCOVA analysis. This should come as no surprise, since this is precisely how sequential sums-of-squares operate in a multiple regression setting. Indeed, had we fit the `gndr` predictor first in our `lm()` command, the sum of squares in the first line of the ANOVA table would have corresponded to the factor sum of squares arrived at from a one-way ANOVA model for this data.

## IV. Two-way Analysis of Variance

As our final topic in this section, we shall briefly examine the two-way ANOVA model. Such models are appropriate for datasets in which we have a continuous numerical response variable and two categorical predictors. As with one-way ANOVA and ANCOVA models, we can write these models in regression form and thus much of the theory of these new models carries over directly from our original multiple regression discussions.

i. Model and Assumptions

Suppose that we have a response variable, $Y_{ijk}$, measured at particular levels of two categorical variables. The triple subscripting is designed to indicate that the observed response value $Y_{ijk}$ is the $k^{\text{th}}$ measurement ($k = 1, \ldots, n$) observed at the $i^{\text{th}}$ level of the first factor ($i = 1, \ldots, I$) and the $j^{\text{th}}$ level of the second factor ($j = 1, \ldots, J$). Notice that we have assumed that the same number of observations has been made within each possible combination of the factor levels; in other words, we will assume that we have a *balanced* design, so that the total number of observations is $N = nIJ$. This restriction is not overly critical, but it allows us to present simpler formulae initially. Typically, the simplest model for our data will be the *additive* model

$$Y_{ijk} = \mu_i + \nu_j + \epsilon_{ijk} = \mu + \tau_i + \alpha_j + \epsilon_{ijk},$$

where $\mu_i + \nu_j$ is the expected response within the $(i, j)^{\text{th}}$ level combination of the two factors, $\mu_i$ representing the effect on the expected response of the $i^{\text{th}}$ level of the first factor and $\nu_j$ the effect of the $j^{\text{th}}$ level of the second factor. This is referred to as the additive model since its structure assumes that the effects of the two factors are additive, that is, that the effect of the either factor is not changed depending on the level of the other factor at which the observations are being made.

As with one-way ANOVA, we have reparameterised the model, so that $\mu$ is an "overall" or "grand" mean, and the parameters $\tau_i$ and $\alpha_j$ measure the discrepancies of the effects of each factor level from the the overall expected response, $\mu$. Of course, again in analogy to the one-way ANOVA, our reparameterisation, while intuitively useful, has too many parameters (note that there are $I + J$ parameters in our original model structure and $I + J + 1$ in our reparameterised model.) It turns out that we need two constraints to remedy the situation

(NOTE: The reason that we need two constraints instead of just one is somewhat technical, but arises from the fact that even the original model has a so-called *identifiability* problem, since adding any constant $c$ to each of the $\mu_i$'s and subtracting the same constant $c$ from each of the $\nu_j$'s results in the same model even though the parameter values are different, implying that this model is itself overparameterised and requiring of a constraint, which in the case of a balanced design is generally taken to be of the form $\sum_{j=1}^{J} \nu_j = 0$.) Typically, we shall assume one of the following sets of constraints:

- the "baseline" or "control group" structure: $\tau_1 = \alpha_1 = 0$; or,
- the "grand mean" constraints: $\sum_{i=1}^{I} \tau_i = \sum_{j=1}^{J} \alpha_j = 0$.

Under the first constraint, the parameter $\mu = \mu_1 + \nu_1$ becomes the expected value for responses at the first level of both factors (which will often be a control or a baseline group), while $\tau_i = \mu_i - \mu_1$ and $\alpha_j = \nu_j - \nu_1$. Under the second constraint, the parameter $\mu$ becomes a true overall mean in the sense that

$$\begin{aligned} E(\overline{Y}) &= \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n} E(Y_{ijk}) = \frac{1}{nIJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n}(\mu + \tau_i + \alpha_j) \\ &= \frac{1}{nIJ}\left(nIJ\mu + nJ\sum_{i=1}^{I}\tau_i + nI\sum_{j=1}^{J}\alpha_j\right) \\ &= \mu. \end{aligned}$$

An important consequence of this calculation is that

$$\begin{aligned} \mu = E(\overline{Y}) &= \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n} E(Y_{ijk}) \\ &= \frac{1}{nIJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n}(\mu_i + \nu_j) \\ &= \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}(\mu_i + \nu_j) \\ &= \frac{1}{I}\sum_{i=1}^{I}\mu_i + \frac{1}{J}\sum_{j=1}^{J}\nu_j. \end{aligned}$$

And therefore, under this constraint,

$$\tau_i = \mu_i - \frac{1}{I}\sum_{l=1}^{I}\mu_l \qquad \text{and} \qquad \alpha_j = \nu_j - \frac{1}{J}\sum_{l=1}^{J}\nu_l.$$

In either case, we will also assume our usual independence, homoscedasticity and normality for the error vector $\epsilon$. In other words, we will assume that

$$\epsilon \sim N(0, \sigma^2 I_n),$$

where $I_n$ represents the $n \times n$ identity matrix.

ii. Parameter Estimation and Hypothesis Testing

We note that the interpretation of the $\mu_i$'s implies that the expected effect due to the first factor is the same for all levels of the second factor, and thus a reasonable estimator is clearly:

$$\hat{\mu}_i = \frac{1}{nJ}\sum_{j=1}^{J}\sum_{k=1}^{n} Y_{ijk} = \overline{Y}_{i\bullet},$$

where the subscripting notation "$i\bullet$" indicates that the averaging has taken place over all the levels of the second factor. Note that $\overline{Y}_{i\bullet}$ is just the average of all response values at the $i^{\text{th}}$ level of the first factor. The estimated second factor effect (adjusted for the first factor by employing the constraint that $\sum_{j=1}^{J}\nu_j = 0$) is:

$$\hat{\nu}_j = \frac{1}{nI}\sum_{i=1}^{I}\sum_{k=1}^{n}(Y_{ijk} - \hat{\mu}_i) = \overline{Y}_{\bullet j} - \frac{1}{I}\sum_{i=1}^{I}\overline{Y}_{i\bullet} = \overline{Y}_{\bullet j} - \overline{Y},$$

where the subscripting notation "$\bullet j$" now indicates that the averaging has taken place over all the levels of the first factor. In addition, it is easily deduced (again using the constraint $\sum_{i=1}^{J}\nu_j = 0$) that:

$$E(\overline{Y}_{i\bullet}) = \mu_i \qquad \text{and} \qquad Var(\overline{Y}_{i\bullet}) = \frac{\sigma^2}{nJ}$$
$$E(\overline{Y}_{\bullet j}) = \nu_j + \frac{1}{I}\sum_{i=1}^{I}\mu_i \qquad \text{and} \qquad Var(\overline{Y}_{\bullet j}) = \frac{\sigma^2}{nI}.$$

Furthermore, once we have these estimates, we can easily estimate the parameters $\mu$, $\tau_i$ and $\alpha_j$. Of course, their interpretations and consequently their estimators will depend on the chosen set of constraints. Under the first set of constraints, we have:

$$\hat{\mu} = \hat{\mu}_1 + \hat{\nu}_1 = \overline{Y}_{1\bullet} + \overline{Y}_{\bullet 1} - \overline{Y}$$
$$\hat{\tau}_i = \hat{\mu}_i - \hat{\mu}_1 = \overline{Y}_{i\bullet} - \overline{Y}_{1\bullet}$$
$$\hat{\alpha}_j = \hat{\nu}_j - \hat{\nu}_1 = \overline{Y}_{\bullet j} - \overline{Y}_{\bullet 1}$$

Under the second set of constraints we note that:

$$\frac{1}{I}\sum_{i=1}^{I}\overline{Y}_{i\bullet} = \frac{1}{I}\sum_{i=1}^{I}\frac{1}{nJ}\sum_{j=1}^{J}\sum_{k=1}^{n}Y_{ijk} = \overline{Y} = \frac{1}{J}\sum_{j=1}^{J}\frac{1}{nI}\sum_{i=1}^{I}\sum_{k=1}^{n}Y_{ijk} = \frac{1}{J}\sum_{j=1}^{J}\overline{Y}_{\bullet j},$$

so that we have:

$$\hat{\mu} = \overline{Y}$$
$$\hat{\tau}_i = \hat{\mu}_i - \frac{1}{I}\sum_{l=1}^{I}\hat{\mu}_l = \overline{Y}_{i\bullet} - \frac{1}{I}\sum_{l=1}^{I}\overline{Y}_{l\bullet} = \overline{Y}_{i\bullet} - \overline{Y}$$
$$\hat{\alpha}_j = \hat{\nu}_j - \frac{1}{J}\sum_{l=1}^{J}\hat{\nu}_l = \overline{Y}_{\bullet j} - \frac{1}{J}\sum_{l=1}^{J}\overline{Y}_{\bullet l} = \overline{Y}_{\bullet j} - \overline{Y}$$

In either case, the variances of these estimators can typically be calculated using the fact that the $\overline{Y}_{i\bullet}$'s are independent of each other as are the $\overline{Y}_{\bullet j}$'s. Indeed, as with one-way ANOVA, we can calculate the variance (and thus create confidence intervals and conduct hypothesis tests) for any linear combination of the $\mu_i$'s or of the $\nu_j$'s this way. However, we note that the $\overline{Y}_{i\bullet}$'s are generally *not* independent of the $\overline{Y}_{\bullet j}$'s and thus for more general contrasts involving both the $\mu_i$'s and the $\nu_j$'s we will have to be more careful (e.g., for the variance of $\hat{\mu} = \hat{\mu}_1 + \hat{\nu}_1 = \overline{Y}_{1\bullet} + \overline{Y}_{\bullet 1}$ under the constraint $\tau_1 = \alpha_1 = 0$). Typically, we will have to write our estimates in terms of the quantities $\overline{Y}_{ij} = (1/n)\sum_{k=1}^{n}Y_{ijk}$ which are independent of each other. Fortunately, there is really no need for writing out all of the necessary formulae, since shortly we will demonstrate that a two-way ANOVA can be written in the form of a multiple regression and then all of the standard regression methods can be used to construct

estimates, confidence intervals and hypothesis tests for any linear combination of the regression parameters, $c^T\beta$ for any vector of constants, $c$.
Moreover, we are generally much more interested in whether either of the factors have any effect on the response at all. In other words, we will want to test the null hypotheses:

$$H_0 : \tau_1 = \ldots = \tau_I = 0 \qquad \text{and} \qquad H_0 : \alpha_1 = \ldots = \alpha_J = 0.$$

Rather than develop a separate theory for the two-way ANOVA, we will now demonstrate how the additive model can be written as a multiple regression, which then allows us to employ all of our previous results for such models. Under the constraint $\tau_1 = \alpha_1 = 0$, we can develop a regression model in exact analogy with the one-way ANOVA situation, so that the model can be written:

$$Y_{ijk} = \beta_0 + \beta_1 z_{1,ijk} + \ldots + \beta_{I-1} z_{(I-1),ijk} + \beta_I w_{1,ijk} + \ldots + \beta_{I+J-2} w_{(J-1),ijk} + \epsilon_{ijk},$$

where $\beta_0 = \mu_1 + \nu_1$ (which equals $\mu$ under the chosen constraints),

$$\beta_l = \begin{cases} \tau_{l+1} & \text{for } l = 1, \ldots, I-1 \\ \alpha_{l+2-I} & \text{for } l = I, \ldots, I+J-2 \end{cases}$$

and the $z_l$'s are the indicators associated with the first factor, while the $w_l$'s are the indicators associated with the second factor, so that:

$$z_{l,ijk} = \begin{cases} 1 & \text{if } i = l+1 \\ 0 & \text{if } i \neq l+1 \end{cases} \qquad \text{and} \qquad w_{l,ijk} = \begin{cases} 1 & \text{if } j = l+1 \\ 0 & \text{if } j \neq l+1 \end{cases}.$$

Using matrix notation, we can write this model more succinctly as:

$$Y = \underline{1}_n \beta_0 + Z\beta_{(1)} + W\beta_{(2)} + \epsilon,$$

where the matrix $Z$ has columns corresponding to the indicators for the first factor (i.e., $Z$ is the design matrix for a one-way ANOVA model of the response on the first factor alone, excluding the initial column associated with the baseline intercept term $\beta_0$), $W$ is the design matrix containing the indicators associated with the second factor (i.e., $W$ is the design matrix for a one-way ANOVA model of the response on the second factor alone, again excluding the initial column associated with the baseline intercept), $\underline{1}_n$ is again a column vector of length $n$ having all elements equal to unity, $\beta_{(1)} = (\beta_1, \ldots, \beta_{I-1})^T$ is the vector of parameters associated with the first factor, and $\beta_{(2)} = (\beta_I, \ldots, \beta_{I+J-2})^T$ is the vector of parameters associated with the second factor.
It is now a simple matter to test whether either of the factors is explaining a significant portion of the variation in the response variable, by using standard ANOVA table breakdowns with the appropriate sequential sums of squares. For instance, to test whether the second factor is significant, we would test the null hypothesis $H_0 : \beta_{(2)} = 0$, using the sequential $F$-statistic

$$F = \frac{SSR(\beta_{(2)}|\beta_{(1)}, \beta_0)/(J-1)}{MSE_{full}},$$

which has an $F$-distribution with $J-1$ numerator and $nIJ - (I+J-1) = N - I - J + 1$ denominator degrees of freedom when the null hypothesis is true.

*Example 3 - Used Car Trade-In Data:* An experiment was conducted to see whether gender and age had any effect on how much trade-in value an individual was offered for their used

car. Eighteen men and eighteen women, six members of both genders in each of three age categories were given a similar mid-sized six-year-old car and asked to see how much they were offered in trade for the car. The data (in hundreds of dollars) is listed below:

| Young | | Middle-Aged | | Elderly | |
|---|---|---|---|---|---|
| Male | Female | Male | Female | Male | Female |
| 21 | 21 | 30 | 26 | 25 | 23 |
| 23 | 22 | 29 | 29 | 22 | 19 |
| 19 | 20 | 26 | 27 | 23 | 20 |
| 22 | 21 | 28 | 28 | 21 | 21 |
| 22 | 19 | 27 | 27 | 22 | 20 |
| 23 | 25 | 27 | 29 | 21 | 20 |

We want to determine if the offered trade-in value for used car depends on age and gender. To do so, we will fit a basic two-way ANOVA model:

```
> tradein <- read.table("Tradein.txt",header=T)
> names(tradein)
[1] "value" "gender" "age"
> attach(tradein)
> tradein.aov <- aov(value ~ age + gender)
> summary(tradein.aov)
            Df Sum of Sq  Mean Sq  F Value     Pr(F)
age          2  316.7222 158.3611 66.05069 0.0000000
gender       1    5.4444   5.4444  2.27082 0.1416397
Residuals 32   76.7222   2.3976
```

So, gender does not appear to be significant after age has already been included. To examine the reverse relationship, that is, whether age is significant after gender has been included in the model, we need only fit the model terms in the reverse order:

```
> tradein.aov <- aov(value ~ gender + age)
> summary(tradein.aov)
            Df Sum of Sq  Mean Sq  F Value     Pr(F)
gender       1    5.4444   5.4444  2.27082 0.1416397
age          2  316.7222 158.3611 66.05069 0.0000000
Residuals 32   76.7222   2.3976
```

Notice that the sums of squares associated with each factor have not changed. This is due to the balanced design in this problem (i.e., the fact that 6 data points were observed in each possible combination of the factor levels), which is an example of a more general category of predictor variable structure called an *orthogonal design*, but this concept is beyond the scope of this course. In general, of course, the sequential sums-of-squares will change when the order of the predictors changes. Indeed, in an unbalanced two-way ANOVA, fitting the predictors in a reversed order will typically lead to different sums of squares in the corresponding ANOVA tables. So, we can now conclude that age does appear to be a significant factor in the offered trade-in value of a used car. To see how age effects the value, we might construct confidence intervals for the "gender-baseline" mean levels $\mu + \tau_i$ for $i = 1, 2, 3$ (where the $\tau_i$'s refer to the levels of the age factor). The simplest way to do this is to fall back on our usual regression approach:

```
> gender1 <- ifelse(gender=="M",1,0)
```

```
> age1 <- ifelse(age=="Middle",1,0)
> age2 <- ifelse(age=="Elderly",1,0)
> tradein.lm <- lm(value ~ age1 + age2 + gender1)
> coefficients(tradein.lm)
 (Intercept) age1        age2   gender1
    21.11111 6.25 -0.08333333 0.7777778
> Xmat <- cbind(1,age1,age2,gender1)
> XtXi <- solve(t(Xmat)%*%Xmat)
> MSE <- sum(residuals(tradein.lm)^2)/32
> cc1 <- c(1,0,0,0)
> cc2 <- c(1,1,0,0)
> cc3 <- c(1,0,1,0)
> est1 <- t(cc1)%*%coefficients(tradein.lm)
> est2 <- t(cc2)%*%coefficients(tradein.lm)
> est3 <- t(cc3)%*%coefficients(tradein.lm)
> ests <- c(est1,est2,est3)
> sd1 <- sqrt(MSE)*sqrt(t(cc1)%*%XtXi%*%cc1)
> sd2 <- sqrt(MSE)*sqrt(t(cc2)%*%XtXi%*%cc2)
> sd3 <- sqrt(MSE)*sqrt(t(cc3)%*%XtXi%*%cc3)
> sds <- c(sd1,sd2,sd3)
> upper <- ests + (qt(0.975,32)*sds)
> lower <- ests - (qt(0.975,32)*sds)
> cbind(lower,ests,upper)
        lower     ests    upper
[1,] 20.05978 21.11111 22.16245
[2,] 26.30978 27.36111 28.41245
[3,] 19.97644 21.02778 22.07911
```

We have used here the fact that, under the chosen constraint that $\tau_1 = \alpha_1 = 0$, we have $\mu + \tau_1 = \mu = \beta_0$, $\mu + \tau_2 = \beta_0 + \beta_1$ and $\mu + \tau_3 = \beta_0 + \beta_2$. In addition, recall that $Var(c^T b) = \sigma^2 c^T (X^T X)^{-1} c$, which we have used to calculate the confidence intervals for our parameters of interest. So, it appears that middle-aged persons (regardless of gender) are offered nearly \$600 more for their trade-ins than either young or elderly people.

iii. Interaction and Testing Additivity

ie. test two factors are not dependent

As with the parallel regression ANCOVA, where we desired to test our assumption of equal slopes, we similarly would like to test our assumption of additivity in the two-way ANOVA model. In other words, we want to test the assumption that the effect on the response of the $i^{\text{th}}$ level of the first factor does not depend on the level of the second factor at which we have observed the data point. To do this, we need to expand our model to incorporate the possibility of differing effects at each of the factor levels depending on the level of the other factor. We do this with the so-called *cell-means* model:

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk},$$

where $\mu_{ij}$ is the expected response at the $(i,j)^{\text{th}}$ factor level combination or *cell*. Notice that this model now contains $IJ$ parameters which will generally be much greater than the $I + J - 1$ parameters of the additive model, and thus we must be careful to ensure that we have enough

datapoints to guarantee that we can estimate all of the parameters. In order to write this model in terms of a multiple regression, we note that we can reparameterise the model as:

$$Y_{ijk} = \mu + \tau_i + \alpha_j + \gamma_{ij} + \epsilon_{ijk},$$

where $\mu$, $\tau_i$ and $\alpha_j$ retain their original interpretations, and the $\gamma_{ij} = \mu_{ij} - \mu - \tau_i - \alpha_j$ are referred to as the factor level *interactions*, which take account of the individual effects of the specific factor level combinations over and above the general effects associated with the parameters $\mu$, $\tau_i$ and $\alpha_j$. Of course, under this new reparameterisation, there are now $IJ + I + J + 1$ parameters, and thus we require $I + J + 1$ constraints to reduce the parameterisation back down to the appropriate number of effective parameters; namely, $IJ$. Again, there are two common sets of constraints:

- the "baseline" or "control group" structure: $\tau_1 = \alpha_1 = \gamma_{i1} = \gamma_{1j} = 0$ for all $i$ and $j$; or,
- the "grand mean" constraints: $\sum_{l=1}^{I} \tau_l = \sum_{l=1}^{J} \alpha_l = \sum_{l=1}^{J} \gamma_{il} = \sum_{l=1}^{I} \gamma_{lj} = 0$.

Under either constraint, the parameters $\mu$, $\tau_i$ and $\alpha_j$ are estimated using the same procedure as before (note that the estimates for the "baseline" structure change slightly to $\hat{\mu} = \overline{Y}_{11}$, $\hat{\tau}_i = \overline{Y}_{i1} - \overline{Y}_{11}$ and $\hat{\alpha}_j = \overline{Y}_{1j} - \overline{Y}_{11}$), and it remains only to estimate the $\gamma_{ij}$'s. Under the "baseline" constraints, we have:

$$\hat{\gamma}_{ij} = \hat{\mu}_{ij} - \hat{\mu} - \hat{\tau}_i - \hat{\alpha}_j = \overline{Y}_{ij} - \overline{Y}_{i1} - \overline{Y}_{1j} + \overline{Y}_{11} \qquad (\text{for } i = 2, \ldots I;\ j = 2, \ldots, J);$$

while under the "grand mean" constraints:

$$\hat{\gamma}_{ij} = \hat{\mu}_{ij} - \hat{\mu} - \hat{\tau}_i - \hat{\alpha}_j = \overline{Y}_{ij} - \overline{Y}_{i\bullet} - \overline{Y}_{\bullet j} + \overline{Y}_{\bullet\bullet}.$$

Again, we can use the independence of the $\overline{Y}_{ij}$'s to calculate the variances of these estimators, though it is unnecessary once we write our model in a multiple regression format, since we can as usual take advantage of all previous regression calculations. To write the model in this form, we need only recognise (as we did for the non-parallel regressions ANCOVA) that the appropriate new predictors to include in the model are the products of all the indicators. So, under the "baseline" constraints, the new model in matrix terms is:

$$Y = \underline{1}_n \beta_0 + Z\beta_{(1)} + W\beta_{(2)} + (Z \otimes W)\beta_{(3)} + \epsilon,$$

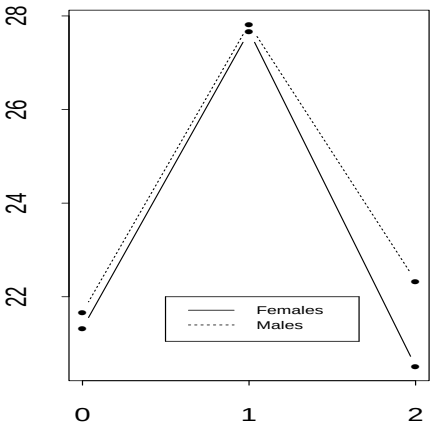
where $\underline{1}_n$, $Z$, $W$, $\beta_{(1)}$ and $\beta_{(2)}$ are all defined as they were for the additive model and $\beta_{(3)} = (\beta_{I+J-1}, \ldots, \beta_{IJ-1})$, the parameters associated with the interaction terms (i.e., the $\beta$'s which correspond to the $\gamma_{ij}$'s) while the matrix $(Z \otimes W)$ has $(I-1)(J-1)$ columns, corresponding to the products of each of the possible indicator pairs $(z_i, w_j)$.

Now, it is the $\gamma_{ij}$'s that we will need to investigate in order to test the plausibility of the additivity assumption (i.e., we must test the null hypothesis $H_0 : \gamma_{ij} = 0$ for all $i = 1, \ldots, I$ and $j = 1, \ldots, J$). Such a hypothesis can be written using our multiple regression format as $H_0 : \beta_{(3)} = 0$, and thus can again be tested using a standard sequential $F$-test.

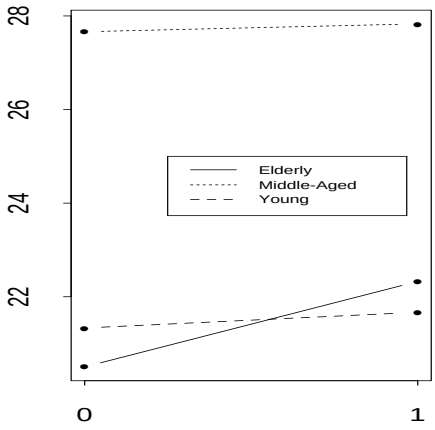
*Example 3 (cont'd) - Used Car Trade-In Data:* We have seen that gender does not appear significant in our additive model analysis. Perhaps this is due to the fact that the additive model is not appropriate for this data and we need to include an interaction term. The most useful pictorial tool for investigating such a possibility is the so-called *cell-means plot*, in which the observed averages within each of the possible factor level combinations are plotted versus the levels of one of the factors, and then the data points corresponding to the same level of the other factor are connected by a line. So, for the data at hand, we would have:

```
> cell.mns <- tapply(value,list(age,gender),mean)
> cell.mns
                F        M
Elderly 20.50000 22.33333
 Middle 27.66667 27.83333
  Young 21.33333 21.66667
```

use cell plots to see if additive model is needed.




Age Group
(0=Young; 1=Middle-Aged; 2=Elderly)




Gender
(0=Female; 1=Male)

If the connected points in such a plot appear to be parallel, then this is an indication that additivity is a justifiable model for the data. On the other hand, if the lines of the connected points criss-cross one another or are generally dissimilar in shape and structure, this indicates that some interaction is likely to be necessary in the model. For the data in this example, the connected points in the first of the plots appear to be parallel. However, the second plot shows some crossing. Nonetheless, the two crossed lines are still rather parallel, and thus an additive model seems appropriate. In addition, these plots also show that gender does not appear to be an important factor in the determination of offered trade in value, since in the first plot the sets of connected points for each gender are extremely close, while in the second plot the lines for each age group are nearly flat. Moreover, we can see that it is the middle-aged individuals who are being offered more than either young or elderly individuals. We can formally test the appropriateness of additivity using the full interactive model as follows:

```
> intrctn <- cbind(age1*gender1,age2*gender1)
> ageind <- cbind(age1,age2)
> tradein.lm <- lm(value ~ ageind + gender1 + intrctn)
> anova(tradein.lm)
Analysis of Variance Table

Response:  value

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value    Pr(F)
ageind     2  316.7222 158.3611 66.29070 0.000000
gender1    1    5.4444   5.4444  2.27907 0.141592
intrctn    2    5.0556   2.5278  1.05814 0.359700
Residuals 30   71.6667   2.3889
```

Thus, the interaction term is not significant, confirming the graphical evidence that the additive model is appropriate for this data. Finally, notice that again the sums of squares associated

with the original indicators are the same as they were in the ANOVA table for the additive model analysis, following our fundamental principle of how sequential sums of squares behave.