Chapter 18

Analysis of Covariance - ANCOVA

Contents

18.1 Introduction .......................................................... 1168
18.2 Assumptions .......................................................... 1171
18.3 Comparing individual regression lines ......................... 1172
18.4 Comparing means after covariate adjustments ............... 1175
18.5 Power and sample size .............................................. 1176
18.6 Example: Degradation of dioxin - multiple locations ....... 1176
18.6.1 Prologue ............................................................ 1186
18.7 Example: Change in yearly average temperature with regime shifts ............... 1187
18.8 Example - More refined analysis of stream-slope example .... 1194
18.9 Example: Comparing Fulton’s Condition Factor \(K\) among groups .... 1202
18.9.1 Fulton’s condition factor for all fish ...................... 1204
18.9.2 Comparing Fulton’s condition factor among groups .......... 1207
18.10 Final Notes .......................................................... 1212

The suggested citation for this chapter of notes is:

In Course Notes for Beginning and Intermediate Statistics.
Available at http://www.stat.sfu.ca/~cschwarz/CourseNotes

18.1 Introduction

In previous chapters, we looked at comparing group means from data collected from a single-factor completely randomized design and analyzed using ANOVA. We also looked at estimating the slope of a straight line between two variables. In both cases the response variable, \(Y\), was continuous (interval or ratio scale). In the case of ANOVA, the \(X\) variables was nominal or ordinal in scale and served to identify the treatment groups. In the regression setting, the \(X\) variable was also continuous.

The Analysis of Covariance (ANCOVA) is a combination of both analyses. Groups are identified by a nominal or ordinal scale variable and a continuous covariate is also measured.
There are two uses of ANCOVA which, on the surface, appear to be separate analyses. In fact, both analyses are identical.

The first use is to check if the regression line for the groups are parallel. If there is evidence that the individual regression lines are not parallel, then a separate regression line must be fit for each group for prediction purposes. If there is no evidence of non-parallelism, then the next task is to see if the lines are co-incident, i.e. have both the same intercept and the same slope. If there is evidence that the lines are not coincident, then a series of parallel lines are fit to the data. All of the data are used to estimate the common slope. If there is no evidence that the lines are not coincident, then all of the data can be simply pooled together and a single regression line fit for all of the data.

The three possibilities are shown below for the case of two groups - the extension to many groups is obvious:
Parallel lines for the two groups.
Second, ANCOVA has been used to test for differences in means among the groups when some of the variation in the responsible variable can be “explained” by a covariate. For example, the effectiveness of two different diets can be compared by randomizing people to the two diets and measuring the weight change during the experiment. However, some of the variation in weight change may be related to initial weight. Perhaps by “standardizing” everyone to some common weight, we can more easily detect differences among the groups.


### 18.2 Assumptions

As before, it is important before the analysis is started to verify the assumptions underlying the analysis. As ANCOVA is a combination of ANOVA and Regression, the assumptions are similar. Both goals of ANCOVA have similar assumptions:

- The response variable $Y$ is continuous (interval or ratio scaled)
• The data are collected under a completely randomized design\footnote{It is possible to relax this assumption - this is beyond the scope of this course.} This implies that the treatment must be randomized completely over the entire set of experimental units if an experimental study, or units must be selected at random from the relevant populations if an observational study.

• There must be no outliers. Plot $Y$ vs. $X$ for each group separately to see if there are any points that don’t appear to follow the straight line.

• The relationship between $Y$ and $X$ must be linear for each group\footnote{It is possible to relax this assumption as well, but is again, beyond the scope of this course.} Check this assumption by looking at the individual plots of $Y$ vs. $X$ for each group.

• The variance must be equal for both groups around their respective regression lines. Check that the spread of the points is equal around the range of $X$ and that the spread is comparable between the two groups. This can be formally checked by looking at the MSE from a separate regression line for each group as MSE estimates the variance of the data around the regression line.

• The residuals must be normally distributed around the regression line for each group. This assumption can be check by examining the residual plots from the fitted model for evidence of non-normality. For large samples, this is not too crucial; for small sample sizes, you will likely have inadequate power to detect anything but gross departures.

\section*{18.3 Comparing individual regression lines}

You saw in earlier chapters, that a statistical model is a powerful shorthand to describe what analysis is fit to a set of data. The model must describe the treatment structure, the experimental unit structure, and the randomization structure. Let $Y$ be the response variable; $X$ be the continuous X-variable, and $\text{Group}$ be the group factor.

In all cases that follow, we are assuming that a completely randomized design was used for the randomization structure. This implies that there are no explicit terms for the randomization structure in the model.

Similarly, there is a single size of experimental unit with no blocking or sub-sampling occurring. This also implies there will be no terms in the model for the experimental unit structure. In more advanced courses, the analyses in this chapter can be extended to more complex designs.

In earlier chapters, we saw that the model for a single-factor completely randomized design is

$$Y = \text{Group}$$

This is read as saying that variation in $Y$ can be partially explained by an overall grand mean (never specified) with differences in the mean caused by $\text{Groups}$ plus an implicit random noise (which is never specified).

Again, from an earlier chapter, we say that the model for a regression of $Y$ on $X$ is

$$Y = X$$

This is read as saying that the variation in $Y$ can be partially explained by an intercept (never specified) plus changes in the $X$ plus an implicit random noise (which is never specified).

As ANCOVA is a combination of the above two analyses, it will not be surprising that the models will have terms corresponding to both $\text{Group}$ and $X$. Again, there are three cases:
If the lines for each group are not parallel:

\[ Y_1 = \text{Group} \times \text{Group} \times X \]

The terms can be in any order. This is read as variation in \( Y \) can be explained a common intercept (never specified) followed by group effects (different intercepts), a common slope on \( X \), and an “interaction” between \( \text{Group} \) and \( X \) which is interpreted as different slopes for each group. This model is almost equivalent to fitting a separate regression line for each group. The only advantage to using this joint model for all groups is similar to that enjoyed by using ANOVA - all of the groups contribute to a better estimate of residual error. If the number of data points per group is small, this can lead to improvements in precision compared to fitting each group individually.

If the lines are parallel across groups, but not coincident:
the appropriate model is

\[ Y^2 = \text{Group} \times X \]

The terms can be in any order. The only difference between this and the previous model is that this simpler model lacks the Group*X “interaction” term. It would not be surprising then that a statistical test to see if this simpler model is tenable would correspond to examining the \( p \)-value of the test on the Group*X term from the complex model. This is exactly analogous to testing for interaction effects between factors in a two-factor ANOVA.

Lastly, if the lines are co-incident:
the appropriate model is

\[ Y^3 = X \]

Now the difference between this model and the previous model is the Group term that has been dropped. Again, it would not be surprising that this corresponds to the test of the Group effect in the formal statistical test. The test for co-incident lines should only be done if there is insufficient evidence against the hypothesis of parallelism.

While it is possible to test for a non-zero slope, this is rarely done.

### 18.4 Comparing means after covariate adjustments

This is straightforward and is illustrated in other example when the means of the categorical variables is compared.

To be added later
18.5 Power and sample size

to be added later

- use the MSE as the estimate of variance for testing MEANS and for testing the slope.

18.6 Example: Degradation of dioxin - multiple locations

An unfortunate byproduct of pulp-and-paper production used to be dioxins - a very hazardous material. This material was discharged into waterways with the pulp-and-paper effluent where it bioaccumulated in living organisms such as crabs. Newer processes have eliminated this byproduct, but the dioxins in the organisms takes a long time to degrade.

Government environmental protection agencies take samples of crabs from affected areas each year and measure the amount of dioxins in the tissue. The following example is based on a real study.

Each year, four crabs are captured from two monitoring stations which are situated quite a distance apart on the same inlet where the pulp mill was located. The liver is excised and the livers from all four crabs are composited together into a single sample\textsuperscript{3}\textsuperscript{3}. The dioxins levels in this composite sample is measured. As there are many different forms of dioxins with different toxicities, a summary measure, called the Total Equivalent Dose (TEQ) is computed from the sample.

As seen earlier, the appropriate response variable is $\log(TEQ)$.

Is the rate of decline the same for both sites? What is the estimated difference or ratio in concentrations between the two sites?

Here are the raw data:

\textsuperscript{3} Compositing is a common analytical tool. There is little loss of useful information induced by the compositing process - the only loss of information is the among individual-sample variability which can be used to determine the optimal allocation between samples within years and the number of years to monitor.
### Table 1

<table>
<thead>
<tr>
<th>Site</th>
<th>Year</th>
<th>TEQ</th>
<th>log(TEQ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1990</td>
<td>179.05</td>
<td>5.19</td>
</tr>
<tr>
<td>a</td>
<td>1991</td>
<td>82.39</td>
<td>4.41</td>
</tr>
<tr>
<td>a</td>
<td>1992</td>
<td>130.18</td>
<td>4.87</td>
</tr>
<tr>
<td>a</td>
<td>1993</td>
<td>97.06</td>
<td>4.58</td>
</tr>
<tr>
<td>a</td>
<td>1994</td>
<td>49.34</td>
<td>3.90</td>
</tr>
<tr>
<td>a</td>
<td>1995</td>
<td>57.05</td>
<td>4.04</td>
</tr>
<tr>
<td>a</td>
<td>1996</td>
<td>57.41</td>
<td>4.05</td>
</tr>
<tr>
<td>a</td>
<td>1997</td>
<td>29.94</td>
<td>3.40</td>
</tr>
<tr>
<td>a</td>
<td>1998</td>
<td>48.48</td>
<td>3.88</td>
</tr>
<tr>
<td>a</td>
<td>1999</td>
<td>49.67</td>
<td>3.91</td>
</tr>
<tr>
<td>a</td>
<td>2000</td>
<td>34.25</td>
<td>3.53</td>
</tr>
<tr>
<td>a</td>
<td>2001</td>
<td>59.28</td>
<td>4.08</td>
</tr>
<tr>
<td>a</td>
<td>2002</td>
<td>34.92</td>
<td>3.55</td>
</tr>
<tr>
<td>a</td>
<td>2003</td>
<td>28.16</td>
<td>3.34</td>
</tr>
<tr>
<td>b</td>
<td>1990</td>
<td>93.07</td>
<td>4.53</td>
</tr>
<tr>
<td>b</td>
<td>1991</td>
<td>105.23</td>
<td>4.66</td>
</tr>
<tr>
<td>b</td>
<td>1992</td>
<td>188.13</td>
<td>5.24</td>
</tr>
<tr>
<td>b</td>
<td>1993</td>
<td>133.81</td>
<td>4.90</td>
</tr>
<tr>
<td>b</td>
<td>1994</td>
<td>69.17</td>
<td>4.24</td>
</tr>
<tr>
<td>b</td>
<td>1995</td>
<td>150.52</td>
<td>5.01</td>
</tr>
<tr>
<td>b</td>
<td>1996</td>
<td>95.47</td>
<td>4.56</td>
</tr>
<tr>
<td>b</td>
<td>1997</td>
<td>146.80</td>
<td>4.99</td>
</tr>
<tr>
<td>b</td>
<td>1998</td>
<td>85.83</td>
<td>4.45</td>
</tr>
<tr>
<td>b</td>
<td>1999</td>
<td>67.72</td>
<td>4.22</td>
</tr>
<tr>
<td>b</td>
<td>2000</td>
<td>42.44</td>
<td>3.75</td>
</tr>
<tr>
<td>b</td>
<td>2001</td>
<td>53.88</td>
<td>3.99</td>
</tr>
<tr>
<td>b</td>
<td>2002</td>
<td>81.11</td>
<td>4.40</td>
</tr>
<tr>
<td>b</td>
<td>2003</td>
<td>70.88</td>
<td>4.26</td>
</tr>
</tbody>
</table>


The data are imported into *R* in the usual fashion:

```r
crabs <- read.csv("dioxin2.csv", header=TRUE, 
as.is=TRUE, strip.white=TRUE, 
na.string=".")
crabs$Site <- factor(crabs$Site) 
crabs$logTEQ <- NULL # drop this and recompute later 
head(crabs) 
str(crabs)
```

Note that both *Year* and *WHO.TEQ* are numeric (*R* doesn’t have the concept of scale of variables).
However, we must declare the Site variable to be a FACTOR, i.e. a categorical variable. In general, it is recommended that alphanumeric codes be used for categorical variables, i.e. don’t code the sites as 1 and 2 because then there is the possibility that R will treat the sites as a continuous variable if you forget to declare the variable as a factor. With alphanumeric codes, R will either figure it out, or issue an error message if you forget to declare the variable as a factor.

Part of the raw data and the structure of the data frame are shown below:

<table>
<thead>
<tr>
<th>Site</th>
<th>Year</th>
<th>WHO.TEQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a</td>
<td>1990</td>
</tr>
<tr>
<td>2</td>
<td>a</td>
<td>1991</td>
</tr>
<tr>
<td>3</td>
<td>a</td>
<td>1992</td>
</tr>
<tr>
<td>4</td>
<td>a</td>
<td>1993</td>
</tr>
<tr>
<td>5</td>
<td>a</td>
<td>1994</td>
</tr>
<tr>
<td>6</td>
<td>a</td>
<td>1995</td>
</tr>
</tbody>
</table>

I recommend recomputing derived variables (e.g. the log() of the TEQ) on the fly, rather than reading them in. This way I avoid any errors where the derived variables are not in sync with the rest of the data. The ordering of the rows is NOT important; however, it is often easier to find individual data points if the data is sorted by the X value. This is particularly true if you want to do a Durbin-Watson test where most packages assume that the data has been temporally ordered. It is common practice in many statistical packages to add extra rows at the end of data set for future predictions.

As usual, start with an initial plot of the data. We already know that we will be plotting on the log-scale so we will skip the first plot on the anti-log scale. In cases with multiple groups, it is often helpful to use a different plotting symbol for each group. This is done in the usual way using the ggplot2 package. Notice how the aes() function can specify the different plotting symbols and colors should be used for the different sites and how the ggplot() function creates the legend.

```r
plotprelimlog <- ggplot(data=crabs, aes(x=Year, y=logTEQ, shape=Site, color=Site)) +
  ggtitle("log(Dioxin) levels over time") +
  xlab("Year") + ylab("log(Dioxin) levels (WHO.TEQ)") +
  geom_point(size=4)
plotprelimlog
```
Before fitting the various models, begin with an exploratory examination of the data looking for outliers and checking the assumptions.

The initial scatter plot doesn’t show any obvious outliers.

Each year’s data is independent of other year’s data as a different set of crabs was selected. Similarly, the data from one site are independent from the other site. This is an observational study, so the question arises of how exactly were the crabs were selected? In this study, crab pots were placed on the floor of the sea to capture the available crabs in the area.

When ever multiple sets of data are collected over time, there is always the worry about common year effects (also known as process error). For example, if the response variable was body mass of small fish, then poor growing conditions in a single year could depress the growth of fish in all locations. This would then violate the assumption of independence as the residual in one site in a year would be related to the residual in another site in the same year. You would then tend to see the residuals “paired” with negative residuals from the fitted line at one site matched (by year) with negative residuals at the other site. In this example, this is unlikely to have occurred. Degradation of dioxin is relatively independent of external environmental factors and the variation that we see about the two regression lines is related solely to sampling error based on the particular set of crabs that that were sampled. It seems unlikely that the residuals are related.

Start by fitting a simple regression to EACH site to see if make sense to pool the data into a single model.

Start by fitting a line to each group separately, i.e. two separate fits, one for each Site. We use the `lm()` function to fit the regression model to each site. This is accomplished within a `dply()` function which is the modern way in R to do by-group processing.

\[^4\] If you actually try and fit a process error term to this model, you find that the estimated process error is zero.
CHAPTER 18. ANALYSIS OF COVARIANCE - ANCOVA

# Fit a separate line for each year
plyr::d_ply(crabs, "Site", function(x){
  # fit a separate line for each site
  cat("\n\n***Separate fit for site :", as.character(x$Site[1]),"\n")
  fit <- lm( logTEQ ~ Year, data=x)
  print(summary(fit))
  print(confint(fit)) # confidence interval on slope
})

The formula in the *lm()* function is what tells R that the response variable is *logTEQ* because it appears to the left of the tilde sign, and that the predictor variable is *Year* because it appears to the right of the tilde sign.

The *summary()* function produces the table that contains the estimates of the regression coefficients and their standard errors and various other statistics.

***Separate fit for site : a

Call:
*lm(formula = logTEQ ~ Year, data = x)*

Residuals:
  Min 1Q Median 3Q Max
-0.59906 -0.16260 -0.01206 0.14054 0.51449

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept) 218.91364 42.79187  5.116  0.000255 ***
Year        -0.10762  0.02143  -5.021  0.000299 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3233 on 12 degrees of freedom
Multiple R-squared: 0.6775,   Adjusted R-squared: 0.6506
F-statistic: 25.21 on 1 and 12 DF,  p-value: 0.0002986

(Intercept) 125.6781579 312.14911470
Year        -0.1543185 -0.06091975

***Separate fit for site : b

Call:
*lm(formula = logTEQ ~ Year, data = x)*

Residuals:
  Min 1Q Median 3Q Max
-0.5567 -0.2399  0.0224 0.2013 0.5059

(Intercept) 125.6781579 312.14911470
Year        -0.1543185 -0.06091975
The estimated slope for the $a$ site is $-0.107$ (se 0.02) while the estimated slope for the $b$ site is $-0.06$ (se 0.02). The 95% confidence intervals overlap considerably so the population slopes could be the same for the two groups.

The MSE from site $a$ is 0.10 and the MSE from site $b$ is 0.12. This corresponds to standard deviations (RMSE) about the regression line of $\sqrt{0.10} = 0.32$ and $\sqrt{0.12} = 0.35$ which are very similar so that assumption of equal standard deviations about the regression line for the two sites seems reasonable.

The residual plots (not shown) also look reasonable.

The assumptions appear to be satisfied, so let us now fit the various models.

First, fit the model allowing for separate lines for each group, i.e. the non-parallel slope model. We use the `lm()` function to fit the regression model with non-parallel slopes:

```r
# Fit the regression line with non-parallel slopes and look at the ANOVA table.
# Because lm() produces type I (increment tests), you should specify the
# contrast in the fit and use the Anova() function from the car package
crabs.fit.np <- lm(logTEQ ~ Site + Year + Year:Site, data=crabs,
    contrast=list(Site='contr.sum'))
car::Anova(crabs.fit.np, type="III")
```

Note that because $R$ computes Type I (incremental) tests of hypotheses, you must specify the interaction term last in the model. Or, as a better solution, specify that the contrast matrix for the factors is the sum-to-zero form and use the `Anova()` function from the `car` package to obtain the Type III (marginal) tests.

The `Anova()` function produces the table that contains the test for the hypothesis of parallel slopes.

```r
Anova Table (Type III tests)
Response: logTEQ
    Sum Sq Df F value Pr(>F)
(Intercept) 3.3408 1 29.372 1.442e-05 ***
Site 0.2612 1 2.297 0.1427
Year 3.1756 1 27.921 2.028e-05 ***
```

© 2019 Carl James Schwarz 1181
The \( p \)-value is 0.14 indicating very little evidence against the hypothesis of parallel slopes.

You can also estimate the individual slopes. The estimates from the non-parallel slope model should be very similar to those seen earlier when the data were split by \textit{Site}. We use the \texttt{lstrends()} function from the \texttt{emmeans} package to fit the regression model with non-parallel slopes:

```r
# estimate the individual slopes and compare them
crabs.fit.np.emmo <- emmeans::emtrends(crabs.fit.np, ~Site, var="Year")
summary(crabs.fit.np.emmo, infer=TRUE)
summary(pairs(crabs.fit.np.emmo), infer=TRUE)
```

which gives:

<table>
<thead>
<tr>
<th>Site</th>
<th>Year.trend</th>
<th>SE</th>
<th>df</th>
<th>lower.CL</th>
<th>upper.CL</th>
<th>t.ratio</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>-0.1076</td>
<td>0.0224</td>
<td>24</td>
<td>-0.154</td>
<td>-0.0615</td>
<td>-4.813</td>
<td>0.0001</td>
</tr>
<tr>
<td>b</td>
<td>-0.0595</td>
<td>0.0224</td>
<td>24</td>
<td>-0.106</td>
<td>-0.0133</td>
<td>-2.660</td>
<td>0.0137</td>
</tr>
</tbody>
</table>

Confidence level used: 0.95

contrast estimate  SE  df   lower.CL upper.CL   t.ratio  p.value  
 a - b  -0.0482  0.0316   24    -0.113  0.0171  -1.523  0.1409   

Confidence level used: 0.95

We see that the both sites have negative trends and that the difference in the slopes is very small, .048 (se .03), and there is no evidence of a difference in the slopes between the two sites. The \( p \)-value from this comparison matches the \( p \)-value from the test for no interaction, as it must. DO NOT USE the values from the \texttt{summary()} function as they depend on the underlying contrast matrix used.

Consequently, we can refit the model, dropping the interaction term:

Again, because \textit{R} does incremental tests, specify the \textit{Site} term last in the model:

```r
# Fit the regression line with parallel slopes.
# Because \texttt{lm()} produces type I (increment tests), you should specify the contrast in the fit and use the \texttt{Anova()} function from the \texttt{car} package

# Fit the regression line with parallel slopes.
# Because \texttt{lm()} produces type I (increment tests), you should specify the contrast in the fit and use the \texttt{Anova()} function from the \texttt{car} package
crabs.fit.p <- lm( logTEQ ~ Year + Site, data=crabs, contrast=list(Site='contr.sum'))
car::Anova(crabs.fit.p, type='III')
```

The \texttt{anova()} table output is:

\begin{verbatim}
Anova Table (Type III tests)
\end{verbatim}
We now have a small \( p \)-value (0.0017) for the Site effect indicating that there is evidence that two lines are not coincident, i.e. they are parallel with different intercepts. This would mean that the rate of decay of the dioxin appears to be equal in both sites, but the initial concentration appears to be different.

It is possible to extract all of the individual pieces using the standard methods (specialized functions to be applied to the results of a model fitting):

```r
# Extract the individual parts of the fit using the standard methods. Note that because Site is a factor
# DO NOT USE THE ESTIMATES from the summary() to estimate the site effect because these estimates depend on the
# internal parameterization used. Use the emmeans() function instead
summary(crabs.fit.p)
coef(crabs.fit.p)
sqrt(diag(vcov(crabs.fit.p))) # gives the SE
confint(crabs.fit.p)
names(summary(crabs.fit.p))
summary(crabs.fit.p)$r.squared
summary(crabs.fit.p)$sigma
```

These results are suitable for any continuous variable (e.g. Year), but be VERY CAUTIOUS about interpreting the estimates for the categorical variable Site as these values depend on the internal parameterization used by \( R \).
CHAPTER 18. ANALYSIS OF COVARIANCE - ANCOVA

F-statistic: 19.47 on 2 and 25 DF,  p-value: 7.985e-06

(Intercept)    Year   Site1
171.07518130  -0.08354254  -0.23043104
(Intercept)    Year   Site1
32.38777458   0.01622224   0.06539395

The common slope has a value of \(-0.083\) (se 0.016). Because the analysis was done on the log-scale, this implies that the dioxin levels changed by a factor of \(\exp(-0.083) = 0.92\) from year to year, i.e. about a 8% decline each year. The 95% confidence interval for the slope on the log-scale is from \((-0.12 \rightarrow -0.05)\) which corresponds to a potential factor between \(\exp(-0.12) = 0.88\) to \(\exp(-0.05) = 0.95\) per year, i.e. between a 12% and 5% decline per year.

While it is possible to estimate the difference between the parallel lines from the information produced by the `summary()` function, this is VERY DANGEROUS as these numbers could change depending on the internal parameterization adopted by \(R\). In the case of categorical variables, the preferred method is to use the `emmeans()` function in the `emmeans` package.

```
# Estimate the size of the site effect. Do not use
# the output from summary() directly as this depends on the
# internal parameterization used by R. We use the emmeans() package
# crabs.fit.p.emmo <- emmeans::emmeans(crabs.fit.p, ~Site)
# sitediff <- pairs(crabs.fit.p.emmo)
# summary(sitediff, infer=TRUE)
```

giving

```
                  contrast  estimate     SE df lower.CL upper.CL t.ratio p.value
a - b             -0.461 0.131 25   -0.73 -0.191  -3.524 0.0017
Confidence level used: 0.95
```

The estimated difference between the lines (on the log-scale) is estimated to be 0.46 (se 0.13). Because the analysis was done on the log-scale, this corresponds to a ratio of \(\exp(0.46) = 1.58\) in median dioxin levels between the two sites, i.e. site \(b\) has 1.58× the dioxin level as site \(a\), on average. Because the slopes are parallel and declining, the dioxin levels are falling in both sites, but the 1.58 times ratio remains consistent.

Predictions and confidence intervals for the mean response and prediction intervals for individual responses can be computed in much the same way as for simple regression. Now we need predictions

---

Caution. There is also a `emmeans()` function in the `lmerTest` package which has different functionality.
for EACH site at each year.

The shaded region is the confidence interval for the mean response, and the bounds are the prediction intervals for single responses.

Computation of confidence intervals for individual responses must be interpreted carefully because the individual samples are composites and not the reading in individual crabs.

It always important to examine the model diagnostics. The $R$ diagnostic plot:
fails to show any evidence of a problem in the fit.

As usual you should examine the residual and other diagnostic plots to ensure that the model is appropriate. Because this is time-series data, you should also check for autocorrelation over time. Check the computer code for details.

18.6.1 Prologue

It turns out that the above analysis committed the cardinal sin of PSEUDO-REPLICATION (Hurlbert, 1984). Pseudo-replication is quite common in studies over time because of the failure to recognize the potential for year-specific effects, i.e. in certain years, both site readings tend to be above their corresponding regression line and in other years, both readings tend to be below their corresponding trend lines.

The proper analysis of this data set requires the use of the Mixed Linear Model where a random effect for the effect of year is added to the model:

\[
\log(TEQ) = Year + YearC(R) + Site + Site : Year
\]
CHAPTER 18. ANALYSIS OF COVARIANCE - ANCOVA

Here $Year$ is the term corresponding to the trendline, while $YearC(R)$ is the random effect of the categorical levels of year.

Please see me for details and consult the $R$ code for the models used.

18.7 Example: Change in yearly average temperature with regime shifts

The ANCOVA technique can also be used for trends when there are KNOWN regime shifts in the series. The case when the timing of the shift is unknown is more difficult and not covered in this course.

For example, consider a time series of annual average temperatures measured at Tuscaloosa, Alabama from 1901 to 2001. It is well known that shifts in temperature can occur whenever the instrument or location or observer or other characteristics of the station change. The data is available in the tuscaloosa.csv file in the Sample Program Library at http://www.stat.sfu.ca/~cschwarz/Stat-Ecology-Datasets.

The data are imported into $R$ in the usual fashion:

```r
tusctemp <- read.csv("tuscaloosa.csv", header=TRUE, as.is=TRUE, strip.white=TRUE, na.string="") # here missing values are blanks or null cells
tusctemp <- tusctemp[complete.cases(tusctemp[,c("Year","Epoch","Avg.Temp..C."))],] # drop the missing values
tusctemp$Epoch <- factor(tusctemp$Epoch)
head(tusctemp)
str(tusctemp)
```

The $Epoch$ variable is declared as a factor and years where the average temperature is a mixture of reading a two different sites are removed. Part of the raw data and the structure of the data frame are shown below:

<table>
<thead>
<tr>
<th>Year</th>
<th>Avg.Temp..C.</th>
<th>Epoch</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1901</td>
<td>16.00</td>
<td>e1</td>
</tr>
<tr>
<td>2</td>
<td>1902</td>
<td>17.56</td>
<td>e1</td>
</tr>
<tr>
<td>3</td>
<td>1903</td>
<td>16.64</td>
<td>e1</td>
</tr>
<tr>
<td>4</td>
<td>1904</td>
<td>17.21</td>
<td>e1</td>
</tr>
<tr>
<td>5</td>
<td>1905</td>
<td>16.96</td>
<td>e1</td>
</tr>
<tr>
<td>6</td>
<td>1906</td>
<td>17.45</td>
<td>e1</td>
</tr>
</tbody>
</table>

A time series plot of the data is constructed in the usual way using the ggplot2 package.

```r
plotprelim <- ggplot(data=tusctemp, aes(x=Year, y=Avg.Temp..C., shape=Epoch, color=Epoch)) +
```

©2019 Carl James Schwarz
The plot clearly shows a shift in the readings in 1939 (thermometer changed), 1957 (station moved), and possibly in 1987 (location and thermometer changed). There is an obvious outlier around 1940 – this reading needs to be investigated further and the analysis should be repeated with this point removed to see if the results are dramatically different.

It turns out that cases where the number of epochs tends to increase with the number of data points has some serious technical issues with the properties of the estimators. See


for details. Basically, if the number of parameters tends to increase with sample size, this violates one of the assumptions for maximum likelihood estimation. This would lead to estimates which may not even be consistent! For example, suppose that the recording changed every two years. Then the two data points should still be able to estimate the common slope, but this corresponds to the well known problem with case-control studies where the number of pairs increases with total sample size. Fortunately, Lu and Lund (2007) showed that this violation is not serious.

The analysis proceeds as in the dioxin example with two sites, except that now the series is broken into different epochs corresponding to the sets of years when conditions remained stable at the recording site. In this case, this corresponds to the years 1901-1938 (inclusive); 1940-1956 (inclusive); 1958-1986 (inclusive), and 1989-2000 (inclusive). Note that the years 1939, 1957, and 1987 are NOT used because...
the average temperature in these two years is an amalgam of two different recording conditions.

For example, the data file (around the first regime change) may look like:

<table>
<thead>
<tr>
<th>Year</th>
<th>Avg.Temp..C.</th>
<th>Epoch</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>36</td>
<td>18.15</td>
<td>e1</td>
<td>&lt;NA&gt;</td>
</tr>
<tr>
<td>37</td>
<td>17.87</td>
<td>e1</td>
<td>&lt;NA&gt;</td>
</tr>
<tr>
<td>38</td>
<td>18.95</td>
<td>e1</td>
<td>&lt;NA&gt;</td>
</tr>
<tr>
<td>40</td>
<td>16.09</td>
<td>e2</td>
<td>&lt;NA&gt;</td>
</tr>
<tr>
<td>41</td>
<td>17.81</td>
<td>e2</td>
<td>&lt;NA&gt;</td>
</tr>
<tr>
<td>42</td>
<td>17.54</td>
<td>e2</td>
<td>&lt;NA&gt;</td>
</tr>
<tr>
<td>43</td>
<td>17.77</td>
<td>e2</td>
<td>&lt;NA&gt;</td>
</tr>
<tr>
<td>44</td>
<td>18.05</td>
<td>e2</td>
<td>&lt;NA&gt;</td>
</tr>
<tr>
<td>45</td>
<td>17.84</td>
<td>e2</td>
<td>&lt;NA&gt;</td>
</tr>
</tbody>
</table>

Notice that we have deleted these odd years data.

Note that Year and Avg Temp and both set to have continuous scale; but Epoch should have a nominal or ordinal scale (in JMP parlance), a FACTOR (in R parlance), or declared as a class (categorical) variable in SAS parlance.

Model filling proceeds as before by first the model:

\[
\text{AvgTemp} = \text{Year} \times \text{Epoch} \times \text{Year} \times \text{Epoch}
\]

to see if the change in AvgTemp per year is consistent among Epochs and then fitting the model:

\[
\text{AvgTemp} = \text{Year} \times \text{Epoch}
\]

to estimate the common trend (after adjusting for shifts among the Epochs).

We first run a separate regression line for each epoch (not shown) to check for outliers, to check that the slopes are similar; and to check that the MSE are comparable among epochs. Then we start with the non-parallel slope model to check for evidence against parallelism.

The non-parallel slope model is fit:

```r
# Fit the regression line with non-parallel slopes and look at the ANOVA table
# Because lm() produces type I (increment tests), you need to specify the contrast type
# use the Anova() function from the car package
# Be sure that Epoch has been declared as a factor.
tusctemp.fit.np <- lm( Avg.Temp..C. ~ Epoch + Year + Year:Epoch, data=tusctemp, contrasts=list(Epoch="contr.sum"))
Anova(tusctemp.fit.np,type="III")
```

giving

\[
\text{Anova Table (Type III tests)}
\]

6If the exact day of the change were known, it is possible to weight the two epochs in these years and include the data points.
There is no strong evidence that the slopes are different among the epochs ($p = .10$) despite the plot showing a potentially differential slope in the 3rd epoch:

You can also estimate the individual slopes. The estimates from the non-parallel slope model should be very similar to those seen earlier when the data were split by Epoch. We use the lstrends() function from the (emmeans package to fit the regression model with non-parallel slopes:

```r
tusctemp.fit.np.emmo <- emmeans::emtrends(tusctemp.fit.np, ~Epoch, var="Year")
summary(tusctemp.fit.np.emmo, infer=TRUE)
emmeans::CLD(tusctemp.fit.np.emmo)
```

which gives:

<table>
<thead>
<tr>
<th>Epoch</th>
<th>Year trend</th>
<th>SE</th>
<th>df</th>
<th>lower.CL</th>
<th>upper.CL</th>
<th>t.ratio</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>e1</td>
<td>0.0403</td>
<td>0.00777</td>
<td>89</td>
<td>0.02484</td>
<td>0.0557</td>
<td>5.184</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>e2</td>
<td>0.0481</td>
<td>0.02600</td>
<td>89</td>
<td>-0.00360</td>
<td>0.0997</td>
<td>1.848</td>
<td>0.0679</td>
</tr>
<tr>
<td>e3</td>
<td>0.0108</td>
<td>0.01166</td>
<td>89</td>
<td>-0.01239</td>
<td>0.0339</td>
<td>0.924</td>
<td>0.3581</td>
</tr>
</tbody>
</table>

©2019 Carl James Schwarz
We see that the all epochs have positive trends, but for some epoch’s, there was no evidence that the slope was different from zero. The compact letter display does not find evidence of a difference among the slopes – this is not surprising given that the \( p \)-value for the test of no interaction was large. The confidence interval for the slope in the third epoch is so large that the slope during this epoch could match the other slopes.

DO NOT USE the values from the `summary()` function as they depend on the underlying contrast matrix used.

The simpler model with common slopes is then fit.

```r
# Fit the regression line with parallel slopes.  # Because lm() produces type I (increment tests), you need to specify the contrast type  # use the Anova() function from the car package  # Be sure that Epoch has been declared as a factor.
tusctemp.fit.p <- lm( Avg.Temp..C. ~ Year + Epoch, data=tusctemp, contrasts=list(Epoch="contr.sum"))
Anova(tusctemp.fit.p, type='III')
```

```r
Anova Table (Type III tests)

Response: Avg.Temp..C.
    Sum Sq Df F value   Pr(>F)
(Intercept) 4.2774 1 14.963 0.0002045 ***
Year 8.0230 1 28.065 7.971e-07 ***
Epoch 23.5990 3 27.517 8.621e-13 ***
Residuals 26.3000 92
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

with fitted (common slope) lines:
The estimated change in average temperature is \(0.033 \pm 0.006\) per year. The 95\% confidence interval does not cover 0.

No further model simplification is possible and there is good evidence that the common slope is different from zero:

The residual/diagnostic plots (against predicted and the order in which the data were collected):
Whenever time series data are used, autocorrelation should be investigated. The Durbin-Watson test is applied to the residuals. The Durbin-Watson test is available in the `lmtest` and `car` package.

```r
# check for autocorrelation using Durbin-Watson test.
# You can use the durbinWatsontest in the car package or the
dwtest in the lmtest package
# For small sample sizes both are fine; for larger sample sizes use the lmtest package
# Note the difference in the default direction of the alternative hypothesis

durbinWatsonTest(tusctemp.fit.p) # from the car package
dwtest(tusctemp.fit.p) # from the lmtest package
```

Note that the default action of the two functions uses a different alternate hypothesis for computing the \( p \)-values (one function returns the one-sided \( p \)-value while the other function returns the two-sided
**p-value** and use different approximations to compute the **p-values**. Hence the results may look slightly different:

<table>
<thead>
<tr>
<th>Lag</th>
<th>Autocorrelation</th>
<th>D-W Statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.1808638</td>
<td>2.299493</td>
<td>0.284</td>
</tr>
</tbody>
</table>

Alternative hypothesis: \( \rho \neq 0 \)

Durbin-Watson test

data: tusctemp.fit.p

\( DW = 2.2995, \text{ p-value} = 0.861 \)

alternative hypothesis: true autocorrelation is greater than 0

with no obvious problems detected.

The leverage plots (against *year*) and other diagnostic plots

also reveal nothing amiss.

### 18.8 Example - More refined analysis of stream-slope example

In the chapter on paired comparisons, the example of the effect of stream slope was examined based on:

In that paper, stream slope was (roughly) categorized into high or low slope classes and a paired-analysis was performed. In this section, we will use the actual stream slopes to examine the relationship between fish density and stream slope.

Recall that a stream reach is a portion of a stream from 10 to several hundred metres in length that exhibits consistent slope. The slope influences the general speed of the water which exerts a dominant influence on the structure of physical habitat in streams. If fish populations are influenced by the structure of physical habitat, then the abundance of fish populations may be related to the slope of the stream.

Reach-scale stream slope and the structure of associated physical habitats are thought to affect trout populations, yet previous studies confound the effect of stream slope with other factors that influence trout populations.

Past studies addressing this issue have used sampling designs wherein data were collected either using repeated samples along a single stream or measuring many streams distributed across space and time.

Reaches on the same stream will likely have correlated measurements making the use of simple statistical tools problematical. [Indeed, if only a single stream is measured on multiple locations, then this is an example of pseudo-replication and inference is limited to that particular stream.]

Inference from streams spread over time and space is made more difficult by the inter-stream differences and temporal variation in trout populations if samples are collected over extended periods of time. This extra variation reduces the power of any survey to detect effects.

For this reason, a paired approach was taken. A total of twenty-three streams were sampled from a large watershed. Within each stream, two reaches were identified and the actual slope gradient was measured.

In each reach, fish abundance was determined using electro-fishing methods and the numbers converted to a density per 100 m$^2$ of stream surface.

Table 19.1 presents the (fictitious but based on the above paper) raw data.
Notice that the density varies considerably among stream but appears to be fairly consistent within each stream.

The raw data is available in the datafile called paired-stream.csv in the Sample Programs Library at http://www.stat.sfu.ca/~cschwarz/Stat-Ecology-Datasets

The data are read in the usual way:
The `stream` variable must be declared as a factor.

As noted earlier, this is an example of an Analytical Survey. The treatments (low or high slope) cannot be randomized within stream – the randomization occurs by selecting streams at random from some larger population of potential streams. As noted in the early chapter on Observational Studies, causal inference is limited whenever a randomization of experimental units to treatments cannot be performed.

Unlike the example presented in other chapters where the slope is divided (arbitrarily) into two class (low and high slope), we will now use the actual slope. A simple regression CANNOT be used because of the non-independence introduced by measuring two reaches on the same stream. However, an ANOCOVA will prove to be useful here.

First, it seem sensible that the response to stream slope will will be multiplicative rather than additive, i.e. an increase in the stream slope will change the fish density by a common fraction, rather than simply changing the density by a fixed amount. For example, it may turn out that a 1 unit change in the slope, reduces density by 10% - if the density before the change was 100 fish/m$^2$, then after the change, the new density will be 90 fish/m$^2$. Similarly, if the original density was only 10 fish/m$^2$, then the final density will be 9 fish/m$^2$. In both cases, the reduction is a fixed fraction, and NOT the same fixed amount (a change of 10 vs. 1).

Create the `log(density)` column in the usual fashion (not illustrated here). In cases like this, the natural logarithm is preferred because the resulting estimates have a very nice simple interpretation.

An appropriate model will be one where each stream has a separate intercept (corresponding to the different productivities of each stream - acting like a block), with a common slope for all streams. The simplified model syntax would look like

$$\log(density) = stream \ times \ slope$$

where the term `stream` represents a nominal scaled variable and gives the different intercepts and the term `slope` is the effect of the common slope on the `log(density)`.

This model is fit using the `lm()` function:

```r
# Analyze using Ancova lm()
# Fit the linear model and get the ANOVA table and test for effects
# Because lm() gives Type I (incremental) tests by default, we need
# to set the contrast matrix.
fish.fit <- lm(logdensity ~ StreamF + slope, data=fish,
               contrasts=list(StreamF="contr.sum"))
```

The `Stream` variable must be declared as a factor and because the `lm()` default is to do incremental (Type I) hypothesis tests, we need to specify the `contrast` option in the function call.

First is a plot of the common slope fit to each stream:
CHAPTER 18. ANALYSIS OF COVARIANCE - ANCOVA

This shows a gradual increase as slope increases. This plot is hard to interpret, but a plot of observed vs. predicted values is clearer:

Generally, the observed are close to the predicted values except for two potential outliers. By clicking on these points, it is shown that both points belong to stream 2 where it appears that the slope increases causes a large decrease in density contrary to the general pattern seen in the other streams.

The effect tests:

Anova Table (Type III tests)

<table>
<thead>
<tr>
<th>Response: logdensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sum Sq</td>
</tr>
<tr>
<td>(Intercept)</td>
</tr>
<tr>
<td>StreamF</td>
</tr>
<tr>
<td>slope</td>
</tr>
</tbody>
</table>

©2019 Carl James Schwarz
fail to detect any influence of slope. Indeed the estimated coefficient associated with a change in slope is found as:

| Estimate Std. Error  t value Pr(>|t|) |
|-------------------------------|---------------------------------|
| slope 0.02465254 0.02998367 0.8221988 0.4197861 |

It is estimated to be .025 (se .0299) and there is no evidence that the change in slope is different from zero. Because the natural log transform was used and the data on the log scale was used, “smallish” slope coefficients have an approximate interpretation. In this example, a slope of .025 on the (natural) log scale implies that the estimated fish density INCREASES by 2.5% every time the slope increases by one percentage point.

Residual plots also show the odd behavior of stream 2:

If this rogue stream is “eliminated” from the analysis, the resulting plots do not show any problems (try it), but now there is evidence of an effect ($p = .035$):

```
After removing stream 2
Anova Table (Type III tests)
Response: logdensity
```
The estimated change in log-density per percentage point change in the slope is found to be:

\[
\text{Estimate Std. Error } t \text{ value } Pr(>|t|) \\
\text{slope } 0.05073439 \quad 0.02251963 \quad 2.252896 \quad 0.03508547
\]

i.e. the slope is .05 (se .02) which is interpreted that a percentage point increase in stream slope increases fish density by 5%. This easy interpretation occurs because the natural log transform was used. If the common (base 10) log transform was used, there is no longer such a simple interpretation.

The remaining residual plot and leverage plots show no problems.

Yet another alternate analysis!

Because the treatment only has two levels, the same answers can also be obtained by estimating the difference of the change in the log(density) to the change in slope. If the slope-class had three or more levels, this analysis could not be done, and the previous analysis would be the preferred route. We first need to put both the slope and log(density) on the same line. The \texttt{melt()} and \texttt{dcast()} functions do the trick:

```r
# Rehape the data from long to wide format
fish.melt <- melt(fish, id.vars=c("Stream","slope.class"), measure.vars=c("slope","logdensity"))
cat('Data frame after melting \n')
head(fish.melt)
fish.wide <- dcast(fish.melt, Stream ~variable+slope.class, value.var=c("value"))
cat('Data frame now in wide format after casting\n')
head(fish.wide)
```

This creates a data table with separate columns for the log(density) and the stream slope for both the high and low slope categories:

<table>
<thead>
<tr>
<th>Stream</th>
<th>slope.class</th>
<th>variable</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>low</td>
<td>slope</td>
<td>0.7</td>
</tr>
<tr>
<td>2</td>
<td>high</td>
<td>slope</td>
<td>4.0</td>
</tr>
<tr>
<td>3</td>
<td>low</td>
<td>slope</td>
<td>2.4</td>
</tr>
<tr>
<td>4</td>
<td>high</td>
<td>slope</td>
<td>6.0</td>
</tr>
<tr>
<td>5</td>
<td>low</td>
<td>slope</td>
<td>0.7</td>
</tr>
<tr>
<td>6</td>
<td>high</td>
<td>slope</td>
<td>2.6</td>
</tr>
</tbody>
</table>

Now create two new variables (create new columns and write a formula for each column) representing the differences in the log(density) and slope between the high and low slope classes:
Data frame now in wide format after casting

<table>
<thead>
<tr>
<th>Stream</th>
<th>slope_high</th>
<th>slope_low</th>
<th>logdensity_high</th>
<th>logdensity_low</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>4.0</td>
<td>0.7</td>
<td>3.044522</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>6.0</td>
<td>2.4</td>
<td>1.131402</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>2.6</td>
<td>0.7</td>
<td>1.856298</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4.0</td>
<td>1.3</td>
<td>2.867899</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>4.4</td>
<td>0.6</td>
<td>1.945910</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>3.2</td>
<td>1.3</td>
<td>3.218876</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stream</th>
<th>slope_high</th>
<th>slope_low</th>
<th>logdensity_high</th>
<th>logdensity_low</th>
<th>diff_slope</th>
<th>diff_logdensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>4.0</td>
<td>0.7</td>
<td>3.044522</td>
<td>3.3</td>
<td>0.33647224</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>6.0</td>
<td>2.4</td>
<td>1.131402</td>
<td>3.6</td>
<td>-1.26649316</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>2.6</td>
<td>0.7</td>
<td>1.856298</td>
<td>1.9</td>
<td>0.08134564</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4.0</td>
<td>1.3</td>
<td>2.867899</td>
<td>2.7</td>
<td>0.36646295</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>4.4</td>
<td>0.6</td>
<td>1.945910</td>
<td>3.8</td>
<td>0.12136086</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>3.2</td>
<td>1.3</td>
<td>3.218876</td>
<td>1.9</td>
<td>-0.46499109</td>
</tr>
</tbody>
</table>

Finally, we wish to fit a line through the origin through these data points. Notice how the formula for the model uses \(+0\) to indicate no intercept. An alternate way to specify no intercept is \(-1\) in place of the \(+0\).

```
# fit a line through the difference that goes through the origin
fish.wide.fit <- lm( diff.logdensity ~ 0+ diff.slope, data=fish.wide)
```

This give the following output:

|             | Estimate | Std. Error | t value | Pr(>|t|) |
|-------------|----------|------------|---------|---------|
| diff.slope  | 0.02465254 | 0.02998367 | 0.8221988 | 0.4197861 |
 CHAPTER 18. ANALYSIS OF COVARIANCE - ANCOVA

We obtain the same estimated effect and \( se \). The outlier from Stream 2 is readily evident. When this outlier is excluded and the analysis is repeated, there is again evidence of an effect that matches the previous analysis after removing Stream 2.

18.9 Example: Comparing Fulton’s Condition Factor \( K \) among groups

Not all fish within a lake are identical. How can a single summary measure be developed to represent the condition of fish within a lake?

In general, the relationship between fish weight and length follows a power law:

\[
W = aL^b
\]

where \( W \) is the observed weight; \( L \) is the observed length, and \( a \) and \( b \) are coefficients relating length to weight. The usual assumption is that heavier fish of a given length are in better condition than lighter fish. Condition indices are a popular summary measure of the condition of the population.

There are at least eight different measures of condition which can be found by a simple literature search. Conne (1989) raises some important questions about the use of a single index to represent the two-dimensional weight-length relationship.

One common measure is Fulton’s \( K \):

\[
K = \frac{Weight}{(\text{Length}/100)^3}
\]


©2019 Carl James Schwarz

1202
This index makes an implicit assumption of isometric growth, i.e. as the fish grows, its body proportions and specific gravity do not change.

How can $K$ be computed from a sample of fish, and how can $K$ be compared among different subset of fish from the same lake or across lakes?

The B.C. Ministry of Environment takes regular samples of rainbow trout using a floating and a sinking net. For each fish captured, the weight (g), length (mm), sex, and maturity of the fish was recorded.


The data are imported into R in the usual fashion:

```r
fish <- read.csv("rainbow-condition.csv", header=TRUE, as.is=TRUE, strip.white=TRUE)
fish$K <- fish$Weight..g./(fish$Len..mm./100)**3
fish$SpeciesF <- factor(fish$Species)
fish$MaturityF <- factor(fish$Maturity)
fish$SexF <- factor(fish$Sex)
fish[1:5,]
```

Part of the raw data are shown below:

<table>
<thead>
<tr>
<th>Net.Type</th>
<th>Fish</th>
<th>Len..mm.</th>
<th>Weight..g.</th>
<th>Species</th>
<th>Sex</th>
<th>Maturity</th>
<th>K</th>
<th>SpeciesF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sinking</td>
<td>1</td>
<td>360</td>
<td>686</td>
<td>RB</td>
<td>F</td>
<td>MATURING</td>
<td>14.70336</td>
<td>RB</td>
</tr>
<tr>
<td>Sinking</td>
<td>2</td>
<td>385</td>
<td>758</td>
<td>RB</td>
<td>F</td>
<td>MATURING</td>
<td>13.28272</td>
<td>RB</td>
</tr>
<tr>
<td>Sinking</td>
<td>3</td>
<td>295</td>
<td>284</td>
<td>RB</td>
<td>M</td>
<td>MATURING</td>
<td>11.06247</td>
<td>RB</td>
</tr>
<tr>
<td>Sinking</td>
<td>4</td>
<td>285</td>
<td>292</td>
<td>RB</td>
<td>F</td>
<td>MATURING</td>
<td>12.61387</td>
<td>RB</td>
</tr>
<tr>
<td>Sinking</td>
<td>5</td>
<td>380</td>
<td>756</td>
<td>RB</td>
<td>F</td>
<td>MATURING</td>
<td>13.77752</td>
<td>RB</td>
</tr>
</tbody>
</table>

$K$ was computed for each individual fish, and the resulting histogram is:
There is a range of condition numbers among the individual fish with an average (among the fish caught) $K$ of about 13.6.

18.9.1 Fulton’s condition factor for all fish

Deriving a single summary measure to represent the entire population of fish in the lake depends heavily on the sampling design used to capture fish.

Some case must be taken to ensure that the fish collected are a simple random sample from the fish in the population. If a net of a single mesh size are used, then this has a selectivity curve and the nets are typically more selective for fish of a certain size. In this experiment, several different mesh sizes were used to try and ensure that all fish of all sizes have an equal chance of being selected.

As well, if regression methods have an advantage in that a simple random sample from the population is no longer required to estimate the regression coefficients. As an analogy, suppose you are interested in the relationship between yield of plants and soil fertility. Such a study could be conducted by finding a random sample of soil plots, but this may lead to many plots with similar fertility and only a few plots with fertility at the tails of the relationship. An alternate scheme is to deliberately seek out soil plots with a range of fertilities or to purposely modify the fertility of soil plots by adding fertilizer, and then fit a regression curve to these selected data points.

Fulton’s index is often re-expressed for regression purposes as:

$$ W = K \left( \frac{L}{100} \right)^3 $$

This looks like a simple regression between $W$ and $\left( \frac{L}{100} \right)^3$ but with no intercept.

A plot of these two variables:
shows a tight relationship among fish but with possible increasing variance with length.

There is some debate about the proper way to estimate the regression coefficient $K$. Classical regression methods (least squares) implicitly implies that all of the “error” in the regression is in the vertical direction, i.e. conditions on the observed lengths. However, the structural relationship between weight and length likely is violated in both variables. This would lead to the error-in-variables problem in regression, which has a long history. Fortunately, the relationship between the two variables is often sufficiently tight that it really doesn’t matter which method is used to find the estimates.

We wish to fit a regression between the two variable but constraining the intercept to be zero. The `lm()` function can be used and you specify $-1$ in the model to remove the intercept and force the line through the point (0,0). Alternatively, you can use $+0$ to do the same thing.

```
k.fit <- lm(W ~ -1 +L, data=fish)
summary(k.fit)
```

This gives rise to the fitted line and statistics about the fit:
Note that $R^2$ really doesn’t make sense in cases where the regression is forced through the origin because the null model to which it is being compared is the line $Y = 0$ which is silly. For this reason, many packages will not report a value of $R^2$ if the intercept is contained to be zero.

The estimated value of $K$ is 13.72 (SE 0.099).

The residual plot and other diagnostic plots

---

8 Consult any of the standard references on regression such as Draper and Smith for more details.
shows clear evidence of increasing variation with the length variable. This usually implies that a weighted regression is needed with weights proportional to the $1/\text{length}^2$ variable. In this case, such a regression gives essentially the same estimate of the condition factor ($\hat{K} = 13.67, SE = .11$).

### 18.9.2 Comparing Fulton’s condition factor among groups

This dataset has a number of sub-groups – do all of the subgroups have the same condition factor? For example, suppose we wish to compare the $K$ value for immature and mature fish. As noted by Garcia-Berthou (2001), this is best done through a technique called Analysis of Covariance (ANCOVA).

We start with a model that has a separate $K$ for each maturity class. The simplified syntax for this model is:

---

\[ W = (\text{Len}/100)^3 \cdot (\text{Len}/100)^3 \cdot \text{Maturity} \]

Note that unlike traditional ANOCOVA models, the model is lacking the simple effect of maturity. The reason for this is that unlike traditional ANCOVA models, the intermediate model with parallel slopes really doesn’t make sense when the regression lines are forced through the origin. This syntax specifies that variation in length are attributable to variations in length and an interaction between the two variables. This latter term represents the differential \( K \) between the maturity classes.

Here is where some care must be taken with some computer packages. For example, JMP “centers” (i.e. subtracts the mean) continuous \( X \) variables when they participate in an interaction or similar term.

Hence, if you just try and implement this above model directly in JMP, you will actually fit the model:

\[ W = (\text{Len}/100)^3 \cdot ((\text{Len}/100)^3 - (\text{Len}/100)^3) \cdot \text{Maturity} \]

which, when expanded, actually adds an intercept term to the model. Ordinarily, in regression models with intercepts, this would NOT be a problem – it is because the model is being forced through the intercept that this causes a problem.

You will need to check your computer package carefully to see if centwring is automatically done, and if so, it must be turned off. Fortunately \( R \) does not centre covariates, so this is not a concern. Because of the way that \( R \) converts the formula to the internal design matrices, the summary table shows a row of estimates that are missing (all NA), and a warning about a singularity is presented. However, the increment test for the interaction term (the default method) rather than forcing \( R \) to get the Type III (or marginal tests) gives the correct results!

```r
mat.fit <- lm(W ~ -1 + L + L:MaturityF, data=fish)
summary(mat.fit)
```

giving:

```
Call:
lm(formula = W ~ -1 + L + L:MaturityF, data = fish)

Residuals:
     Min      1Q  Median      3Q     Max
-107.910 -10.679   2.267  15.407  94.711

Coefficients: (1 not defined because of singularities)
                     Estimate Std. Error   t value Pr(>|t|)
(Intercept)            NA         NA        NA      NA
L                     13.8055   0.1029 134.143   <2e-16 ***
L:MaturityFIM       -0.7024   0.3129  -2.245  0.0266  *
L:MaturityFMATURE     NA         NA         NA        NA
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30.79 on 121 degrees of freedom
Multiple R-squared: 0.994,  Adjusted R-squared: 0.9939
F-statistic: 9981 on 2 and 121 DF,  p-value: < 2.2e-16
```

© 2019 Carl James Schwarz
CHAPTER 18. ANALYSIS OF COVARIANCE - ANCOVA

Note the use of the *No Intercept* option in many packages to force the line through the origin. Some computer packages will `complain` about the odd form of the model because it is missing the simple *maturity class* effect, but just ignore the complaints. This gives the summary output for the effect test of:

```
Analysis of Variance Table
Response: W

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
<td>1</td>
<td>18918849</td>
<td>18918849</td>
<td>19956.3765</td>
<td>&lt; 2e-16 ***</td>
</tr>
<tr>
<td>L:MaturityF</td>
<td>1</td>
<td>4779</td>
<td>4779</td>
<td>5.0411</td>
<td>0.02657 *</td>
</tr>
<tr>
<td>Residuals</td>
<td>121</td>
<td>114709</td>
<td>948</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The *p*-value for the last term in the table (the interaction term) of 0.027 indicates that there is strong evidence of a different $K$ between the two maturity classes.

The estimate of the slopes for the separate maturity classes are obtained using the *lstrend()* function from the *emmeans* package in the usual way:

```
# Estimate the individual slopes
mat.fit.emmo <- emmeans::emtrends(mat.fit, ~MaturityF, var="L")
summary(mat.fit.emmo, infer=TRUE)
summary(pairs(mat.fit.emmo), infer=TRUE)
```

which gives the estimated $K$ for each maturity class.

```
MaturityF  L.trend   SE   df  lower.CL  upper.CL  t.ratio p.value
IM          13.1  0.295  121    12.5     13.7   44.351 <.0001
MATURING    13.8  0.103  121    13.6     14.0   134.143 <.0001

Confidence level used: 0.95
contrast  estimate   SE   df  lower.CL  upper.CL  t.ratio  p.value
IM - MATURING -0.702  0.313  121   -1.32   -0.0831   -2.245   0.0266

Confidence level used: 0.95
```

If you fit a separate regression for the two maturity classes, you will get the two same estimates. The respective standard errors will be slightly different because the single model is able to pool over all of the data to estimate the standard errors, but separate estimates cannot do any pooling.

The separate fitted lines are shown below:
Similarly, a comparison of $K$ can be made among the three sex classes (M, F, and U) where immature fish cannot be sexed and are given the code U, but mature fish are further subdivided into M and F classes.

Analysis of Variance Table

Response: W

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
<td>1</td>
<td>18918849</td>
<td>18918849</td>
<td>20202.577&lt;2e-16 ***</td>
</tr>
<tr>
<td>L:SexF</td>
<td>2</td>
<td>7113</td>
<td>3557</td>
<td>3.798</td>
</tr>
<tr>
<td>Residuals</td>
<td>120</td>
<td>112375</td>
<td>936</td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

This shows evidence ($p = .025$) of a differential $K$ among the three sex classes (this is not unexpected). A comparison of the slopes is required:

<table>
<thead>
<tr>
<th>SexF L.trend</th>
<th>SE</th>
<th>df</th>
<th>lower.CL</th>
<th>upper.CL</th>
<th>t.ratio</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>13.6</td>
<td>0.115</td>
<td>120</td>
<td>13.4</td>
<td>13.8</td>
<td>117.798 &lt;.0001</td>
</tr>
<tr>
<td>M</td>
<td>14.2</td>
<td>0.195</td>
<td>120</td>
<td>13.8</td>
<td>14.6</td>
<td>72.657 &lt;.0001</td>
</tr>
<tr>
<td>U</td>
<td>13.6</td>
<td>0.417</td>
<td>120</td>
<td>12.7</td>
<td>14.4</td>
<td>32.476 &lt;.0001</td>
</tr>
</tbody>
</table>

Confidence level used: 0.95

<table>
<thead>
<tr>
<th>SexF L.trend</th>
<th>SE</th>
<th>df</th>
<th>lower.CL</th>
<th>upper.CL</th>
<th>.group</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>13.6</td>
<td>0.417</td>
<td>120</td>
<td>12.7</td>
<td>14.4</td>
</tr>
<tr>
<td>F</td>
<td>13.6</td>
<td>0.115</td>
<td>120</td>
<td>13.4</td>
<td>13.8</td>
</tr>
<tr>
<td>M</td>
<td>14.2</td>
<td>0.195</td>
<td>120</td>
<td>13.8</td>
<td>14.6</td>
</tr>
</tbody>
</table>
As the \( p \)-value is .0074 (unadjusted for multiple comparisons as given the \textit{JMP}) or .0202 (after a Tukey adjustment in \textit{SAS} or \textit{R}), there is also strong evidence of a differential \( K \) between the males and females as well.

A final plot of the three lines is:

![Graph](image)

Finally, because you have replicate fish at the same body length, it is possible to a formal lack-of-fit test. The idea behind this test is to compare the variation in data points at the same replicate lengths (pure error) with the deviations around the line from the model (model error). If the model fits well, the ratio of these two estimates of residual variance should be comparable:

The lack-of-fit test in \textit{R} is found using a direct model comparison between the model as fitted (using \textit{sex} and \textit{length}) vs a model where every unique combination of the two variable is treated as a level of a factor and so has an exact fit to the mean: \textit{emmeans} package in the usual way:
CHAPTER 18. ANALYSIS OF COVARIANCE - ANCOVA

```
pure.error.fit <- lm( W ~ interaction(fish$L, fish$Sex), data=fish)
anova(sex.fit, pure.error.fit)
```
giving

<table>
<thead>
<tr>
<th>Model 1: W ~ -1 + L + L:SexF</th>
<th>Model 2: W ~ interaction(fish$L, fish$Sex)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Res.Df RSS Df Sum of Sq F Pr(&gt;F)</td>
<td>Res.Df RSS Df Sum of Sq F Pr(&gt;F)</td>
</tr>
<tr>
<td>1 120 112375</td>
<td>2 46 42399 74 69975 1.0259 0.47</td>
</tr>
</tbody>
</table>

The $p$-value for the lack-of-fit test is quite large ($p = 0.47$) indicating no evidence of a lack of fit.

This same ANCOVA method can be used to compare the $K$ values across lakes or across time within the same lake. If you have a large number of lakes each measured multiple times, some very interesting models can be fit that are beyond the scope of these notes – please contact me. Similarly, interest may lie in modeling the $K$ as functions of other lake-specific covariates such as lake size, productivity, etc. Again, please contact me as this is beyond the scope of these notes.

**Statistical significance is not the same as biological significance!** While there was evidence of differential $K$ in this data set, this statistical significance does not imply biological importance. I have no idea of the observed differences in $K$ among these three groups has any meaning biologically.

### 18.10 Final Notes

Some sections need to be added here on the following topics:

- danger of ANCOVA is there is no overlap in the covariate
- choice between paired t-test, multi-variate test, or ANCOVA in the case of two time points