Chapter 7

Analysis of Covariance: Comparing Regression Lines

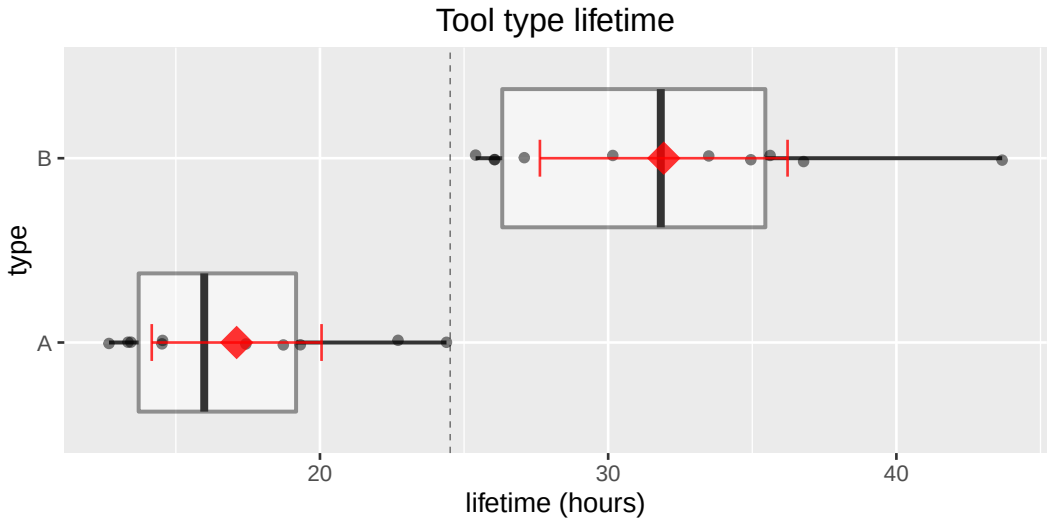
Suppose that you are interested in comparing the typical lifetime (hours) of two tool types (A and B). A simple analysis of the data given below would consist of making side-by-side boxplots followed by a two-sample test of equal means (or medians). The standard two-sample test using the pooled variance estimator is a special case of the one-way ANOVA with two groups. The summaries suggest that the distribution of lifetimes for the tool types are different. In the output below, μ_i is population mean lifetime for tool type i ($i = A, B$).

```
#### Example: Tool lifetime
tools <- read.table("http://statacumen.com/teach/ADA2/ADA2_notes_Ch07_tools.dat"
                    , header = TRUE)
str(tools)
```

```
## 'data.frame': 20 obs. of  3 variables:
##  $ lifetime: num  18.7 14.5 17.4 14.5 13.4 ...
##  $ rpm     : int  610 950 720 840 980 530 680 540 890 730 ...
##  $ type    : Factor w/ 2 levels "A","B": 1 1 1 1 1 1 1 1 1 1 ...
```

	lifetime	rpm	type		lifetime	rpm	type
1	18.7300	610	A	11	30.1600	670	B
2	14.5200	950	A	12	27.0900	770	B
3	17.4300	720	A	13	25.4000	880	B
4	14.5400	840	A	14	26.0500	1000	B
5	13.4400	980	A	15	33.4900	760	B
6	24.3900	530	A	16	35.6200	590	B
7	13.3400	680	A	17	26.0700	910	B
8	22.7100	540	A	18	36.7800	650	B
9	12.6800	890	A	19	34.9500	810	B
10	19.3200	730	A	20	43.6700	500	B

```
library(ggplot2)
p <- ggplot(tools, aes(x = type, y = lifetime))
# plot a reference line for the global mean (assuming no groups)
p <- p + geom_hline(aes(yintercept = mean(lifetime)),
                    colour = "black", linetype = "dashed", size = 0.3, alpha = 0.5)
# boxplot, size=.75 to stand out behind CI
p <- p + geom_boxplot(size = 0.75, alpha = 0.5)
# points for observed data
p <- p + geom_point(position = position_jitter(w = 0.05, h = 0), alpha = 0.5)
# diamond at mean for each group
p <- p + stat_summary(fun.y = mean, geom = "point", shape = 18, size = 6,
                      colour="red", alpha = 0.8)
# confidence limits based on normal distribution
p <- p + stat_summary(fun.data = "mean_cl_normal", geom = "errorbar",
                      width = .2, colour="red", alpha = 0.8)
p <- p + labs(title = "Tool type lifetime") + ylab("lifetime (hours)")
p <- p + coord_flip()
print(p)
```



A two sample t-test comparing mean lifetimes of tool types indicates a difference between means.

```
t.summary <- t.test(lifetime ~ type, data = tools)
t.summary
```

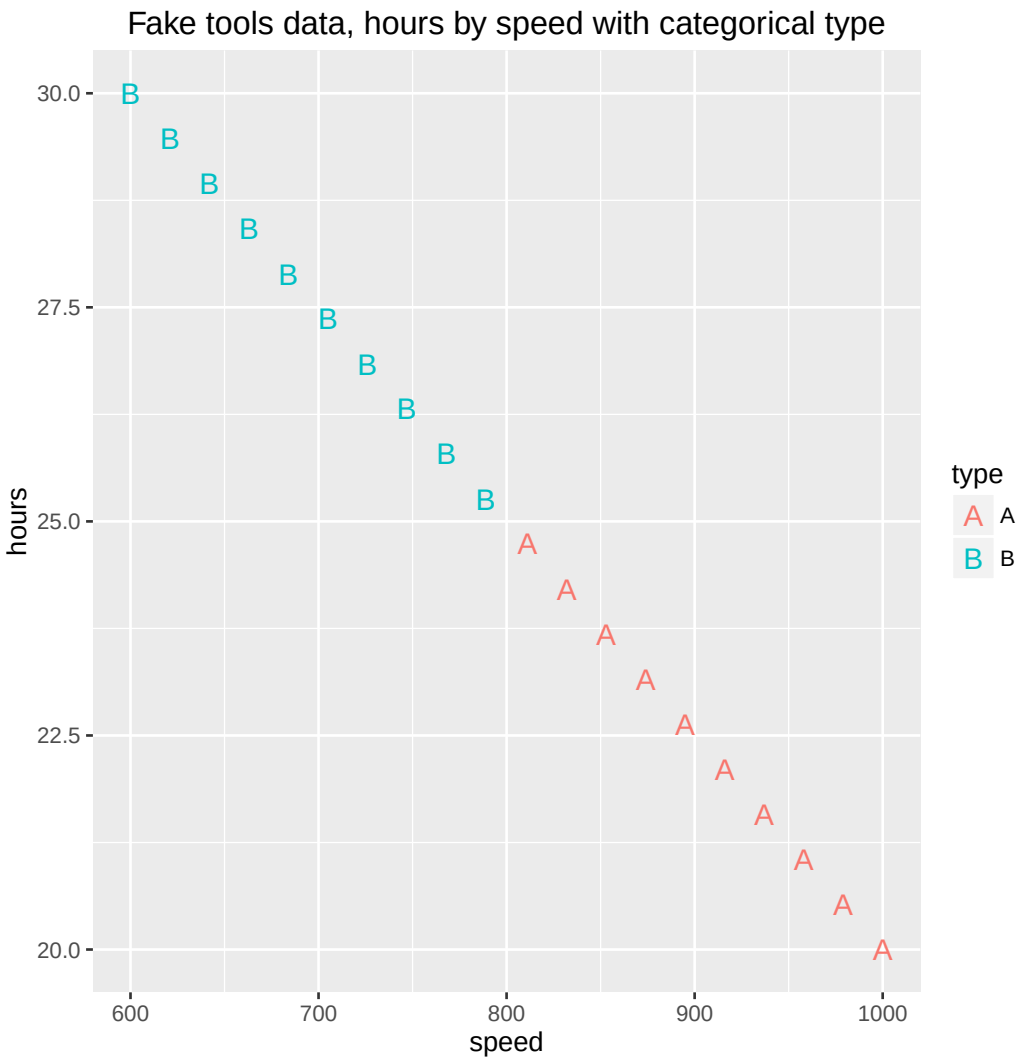
```
##
##  Welch Two Sample t-test
##
```

```
## data:  lifetime by type
## t = -6.435, df = 15.93, p-value = 8.422e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -19.70128  -9.93472
## sample estimates:
## mean in group A mean in group B
##          17.110          31.928
```

This comparison is potentially misleading because the samples are not comparable. A one-way ANOVA is most appropriate for designed experiments where all the factors influencing the response, other than the treatment (tool type), are fixed by the experimenter. The tools were operated at different speeds. If speed influences lifetime, then the observed differences in lifetimes could be due to differences in speeds at which the two tool types were operated.

Fake example For example, suppose speed is inversely related to lifetime of the tool. Then, the differences seen in the boxplots above could be due to tool type B being operated at lower speeds than tool type A. To see how this is possible, consider the data plot given below, where the relationship between lifetime and speed is identical in each sample. A simple linear regression model relating hours to speed, ignoring tool type, fits the data exactly, yet the lifetime distributions for the tool types, ignoring speed, differ dramatically. (The data were generated to fall exactly on a straight line). The regression model indicates that you would expect identical mean lifetimes for tool types A and B, if they were, or could be, operated at identical speeds. This is not exactly what happens in the actual data. However, I hope the point is clear.

```
#### Example: Tools, fake
toolsfake <- read.table("http://statacumen.com/teach/ADA2/ADA2_notes_Ch07_toolsfake.dat"
                        , header = TRUE)
library(ggplot2)
p <- ggplot(toolsfake, aes(x = speed, y = hours, colour = type, shape = type))
p <- p + geom_point(size=4)
  library(R.oo) # for ascii code lookup
  p <- p + scale_shape_manual(values=charToInt(sort(unique(toolsfake$type))))
p <- p + labs(title="Fake tools data, hours by speed with categorical type")
print(p)
```



As noted in the Chapter 6 SAT example, you should be wary of group comparisons where important factors that influence the response have not been accounted for or controlled. In the SAT example, the differences in scores were affected by a change in the ethnic composition over time. A two-way ANOVA with two factors, time and ethnicity, gave the most sensible analysis.

For the tool lifetime problem, you should compare groups (tools) after adjusting the lifetimes to account for the influence of a measurement variable, speed. The appropriate statistical technique for handling this problem is called **analysis of covariance** (ANCOVA).

7.1 ANCOVA

A natural way to account for the effect of speed is through a multiple regression model with lifetime as the response and two predictors, speed and tool type. A binary **categorical variable**, here tool type, is included in the model as a **dummy variable** or **indicator variable** (a $\{0, 1\}$ variable).

Consider the model

$$\text{Tool lifetime} = \beta_0 + \beta_1 \text{ typeB} + \beta_2 \text{ rpm} + e,$$

where typeB is 0 for type A tools, and 1 for type B tools. For type A tools, the model simplifies to:

$$\begin{aligned} \text{Tool lifetime} &= \beta_0 + \beta_1(0) + \beta_2 \text{ rpm} + e \\ &= \beta_0 + \beta_2 \text{ rpm} + e. \end{aligned}$$

For type B tools, the model simplifies to:

$$\begin{aligned} \text{Tool lifetime} &= \beta_0 + \beta_1(1) + \beta_2 \text{ rpm} + e \\ &= (\beta_0 + \beta_1) + \beta_2 \text{ rpm} + e. \end{aligned}$$

This ANCOVA model fits two regression lines, one for each tool type, but restricts the slopes of the regression lines to be identical. To see this, let us focus on the interpretation of the regression coefficients. For the ANCOVA model,

β_2 = slope of population regression lines for tool types A and B.

and

β_0 = intercept of population regression line for tool A (called the reference group).

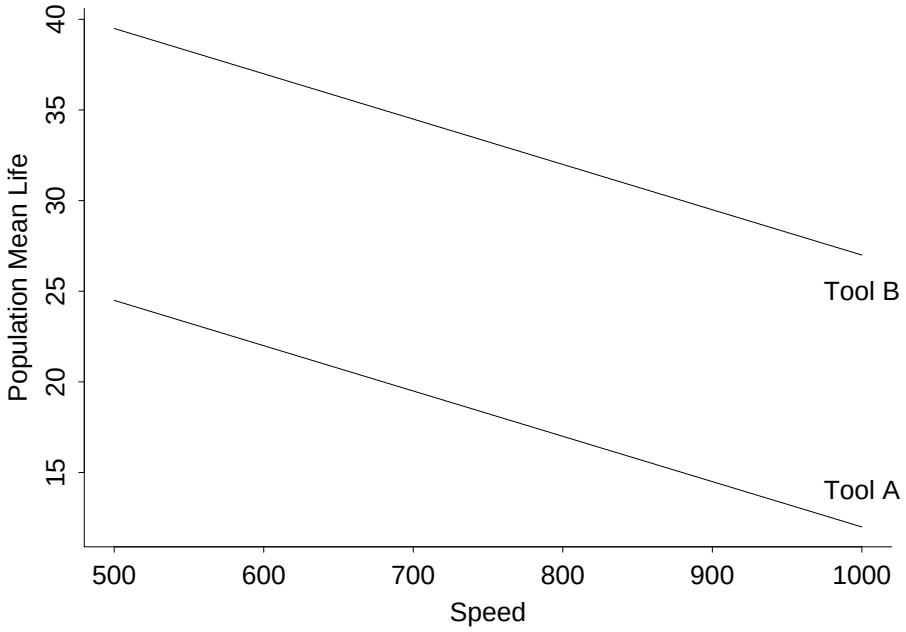
Given that

$\beta_0 + \beta_1$ = intercept of population regression line for tool B,

it follows that

β_1 = difference between tool B and tool A intercepts.

A picture of the population regression lines for one version of the model is given below.

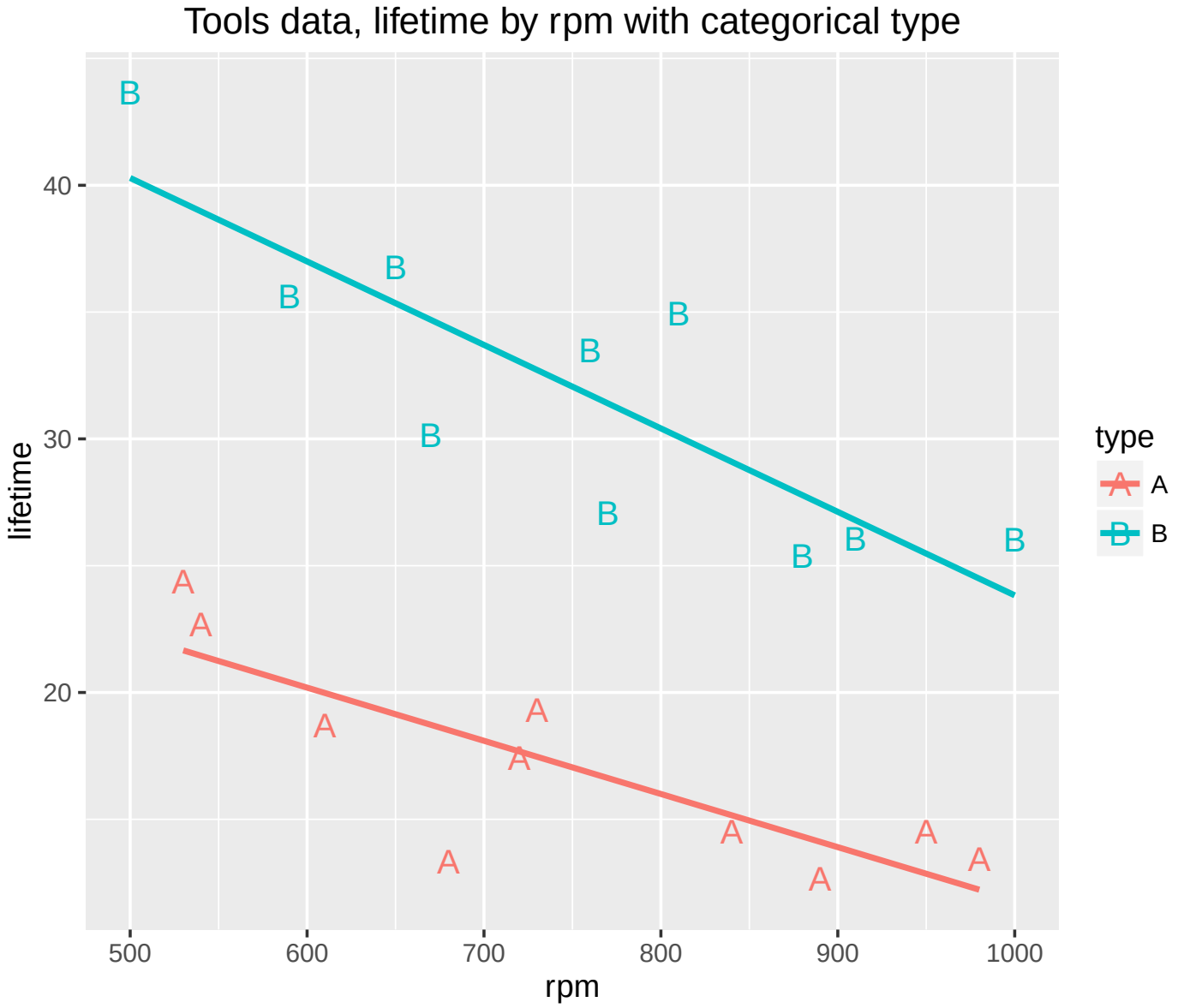


An important feature of the ANCOVA model is that β_1 measures the difference in mean response for the tool types, regardless of the speed. A test of $H_0 : \beta_1 = 0$ is the primary interest, and is interpreted as a **comparison of the tool types, after adjusting or allowing for the speeds at which the tools were operated**.

The ANCOVA model is plausible. The relationship between lifetime and speed is roughly linear within tool types, with similar slopes but unequal intercepts across groups. The plot of the studentized residuals against the fitted values shows no gross abnormalities, but suggests that the variability about the regression line for tool type A is somewhat smaller than the variability for tool type B. The model assumes that the variability of the responses is the same for each group. The QQ-plot does not show any gross deviations from a straight line.

```
#### Example: Tool lifetime
library(ggplot2)
p <- ggplot(tools, aes(x = rpm, y = lifetime, colour = type, shape = type))
```

```
p <- p + geom_point(size=4)
  library(R.oo) # for ascii code lookup
  p <- p + scale_shape_manual(values=charToInt(sort(unique(tools$type))))
p <- p + geom_smooth(method = lm, se = FALSE)
p <- p + labs(title="Tools data, lifetime by rpm with categorical type")
print(p)
```




```
lm.l.r.t <- lm(lifetime ~ rpm + type, data = tools)
#library(car)
#Anova(aov(lm.l.r.t), type=3)
summary(lm.l.r.t)
```

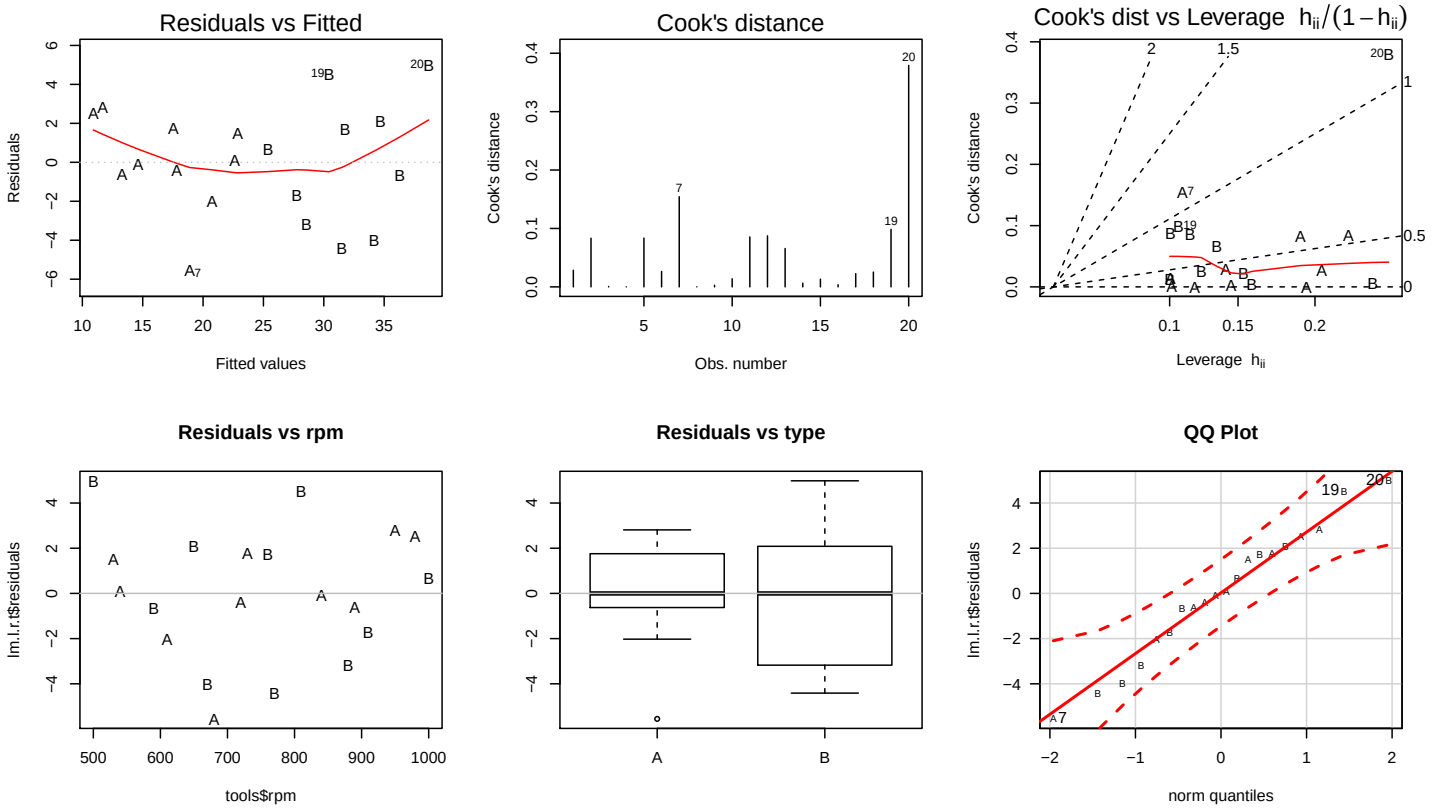
```
##
## Call:
## lm(formula = lifetime ~ rpm + type, data = tools)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5527 -1.7868 -0.0016  1.8395  4.9838
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.98560    3.51038  10.536 7.16e-09 ***
## rpm         -0.02661    0.00452  -5.887 1.79e-05 ***
## typeB       15.00425    1.35967  11.035 3.59e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.039 on 17 degrees of freedom
## Multiple R-squared:  0.9003,Adjusted R-squared:  0.8886
## F-statistic: 76.75 on 2 and 17 DF,  p-value: 3.086e-09
```

```
# plot diagnostics
par(mfrow=c(2,3))
plot(lm.l.r.t, which = c(1,4,6), pch=as.character(tools$type))

plot(tools$rpm, lm.l.r.t$residuals, main="Residuals vs rpm", pch=as.character(tools$type))
  # horizontal line at zero
  abline(h = 0, col = "gray75")

plot(tools$type, lm.l.r.t$residuals, main="Residuals vs type")
  # horizontal line at zero
  abline(h = 0, col = "gray75")

# Normality of Residuals
library(car)
qqPlot(lm.l.r.t$residuals, las = 1, id.n = 3, main="QQ Plot", pch=as.character(tools$type))
```

```
##  7 20 19
##  1 20 19
```

```
## residuals vs order of data
#plot(lm.l.r.t£residuals, main="Residuals vs Order of data")
#  # horizontal line at zero
#  abline(h = 0, col = "gray75")
```



The fitted relationship for the combined data set is

$$\text{Predicted Lifetime} = 36.99 + 15.00\ \text{typeB} - 0.0266\ \text{rpm}.$$

Assigning the LS estimates to the appropriate parameters, the fitted relationships for the two tool types must be, for tool type B:

$$\begin{aligned}\text{Predicted Lifetime} &= (36.99 + 15.00) - 0.0266\ \text{rpm} \\ &= 51.99 - 0.0266\ \text{rpm},\end{aligned}$$

and for tool type A:

$$\text{Predicted Lifetime} = 36.99 - 0.0266\ \text{rpm}.$$

The t-test of $H_0 : \beta_1 = 0$ checks whether the intercepts for the population regression lines are equal, assuming equal slopes. The t-test p-value < 0.0001 suggests that the population regression lines for tools A and B have unequal intercepts. The LS lines indicate that the average lifetime of either type tool decreases by 0.0266 hours for each increase in 1 RPM. Regardless of the lathe

speed, the model predicts that type B tools will last 15 hours longer (i.e., the regression coefficient for the typeB predictor) than type A tools. Summarizing this result another way, the t-test suggests that there is a significant difference between the lifetimes of the two tool types, after adjusting for the effect of the speeds at which the tools were operated. The estimated difference in average lifetime is 15 hours, regardless of the lathe speed.

7.2 Generalizing the ANCOVA Model to Allow Unequal Slopes

I will present a flexible approach for checking equal slopes and equal intercepts in ANCOVA-type models. The algorithm also provides a way to build regression models in studies where the primary interest is comparing the regression lines across groups rather than comparing groups after adjusting for a regression effect. The approach can be applied to an arbitrary number of groups and predictors. For simplicity, I will consider a problem with three groups and a single regression effect.

The data[1] below are the IQ scores of identical twins, one raised in a foster home (IQF) and the other raised by natural parents (IQN). The 27 pairs are divided into three groups by social status of the natural parents (H=high, M=medium, L=low). I will examine the regression of IQF on IQN for each of the three social classes.

There is no **a priori** reason to assume that the regression lines for the three groups have equal slopes or equal interepts. These are, however, reasonable hypotheses to examine. The easiest way to check these hypotheses is to fit a multiple regression model to the combined data set, and check whether certain carefully defined regression effects are zero. The most general model has six parameters, and corresponds to fitting a simple linear regression model to the three groups separately ($3 \times 2 = 6$).

[1]The data were originally analyzed by Sir Cyril Burt.

Two indicator variables are needed to uniquely identify each observation by social class. For example, let $I_1 = 1$ for H status families and $I_1 = 0$ otherwise, and let $I_2 = 1$ for M status families and $I_2 = 0$ otherwise. The indicators I_1 and I_2 jointly assume 3 values:

Status	I_1	I_2
L	0	0
M	0	1
H	1	0

Given the indicators I_1 and I_2 and the predictor IQN, define two **interaction** or **product effects**: $I_1 \times$ IQN and $I_2 \times$ IQN.

7.2.1 Unequal slopes ANCOVA model

The most general model allows separate slopes and intercepts for each group:

$$\text{IQF} = \beta_0 + \beta_1 I_1 + \beta_2 I_2 + \beta_3 \text{ IQN} + \beta_4 I_1 \text{ IQN} + \beta_5 I_2 \text{ IQN} + e. \quad (7.1)$$

This model is best understood by considering the three status classes separately. If status = L, then $I_1 = I_2 = 0$. For these families

$$\text{IQF} = \beta_0 + \beta_3 \text{ IQN} + e.$$

If status = M, then $I_1 = 0$ and $I_2 = 1$. For these families

$$\begin{aligned} \text{IQF} &= \beta_0 + \beta_2(1) + \beta_3 \text{ IQN} + \beta_5 \text{ IQN} + e \\ &= (\beta_0 + \beta_2) + (\beta_3 + \beta_5) \text{ IQN} + e. \end{aligned}$$

Finally, if status = H, then $I_1 = 1$ and $I_2 = 0$. For these families

$$\begin{aligned} \text{IQF} &= \beta_0 + \beta_1(1) + \beta_3 \text{ IQN} + \beta_4 \text{ IQN} + e \\ &= (\beta_0 + \beta_1) + (\beta_3 + \beta_4) \text{ IQN} + e. \end{aligned}$$

The regression coefficients β_0 and β_3 are the intercept and slope for the L status population regression line. The other parameters measure differences in intercepts and slopes across the three groups, using L status families as a **baseline** or **reference group**. In particular:

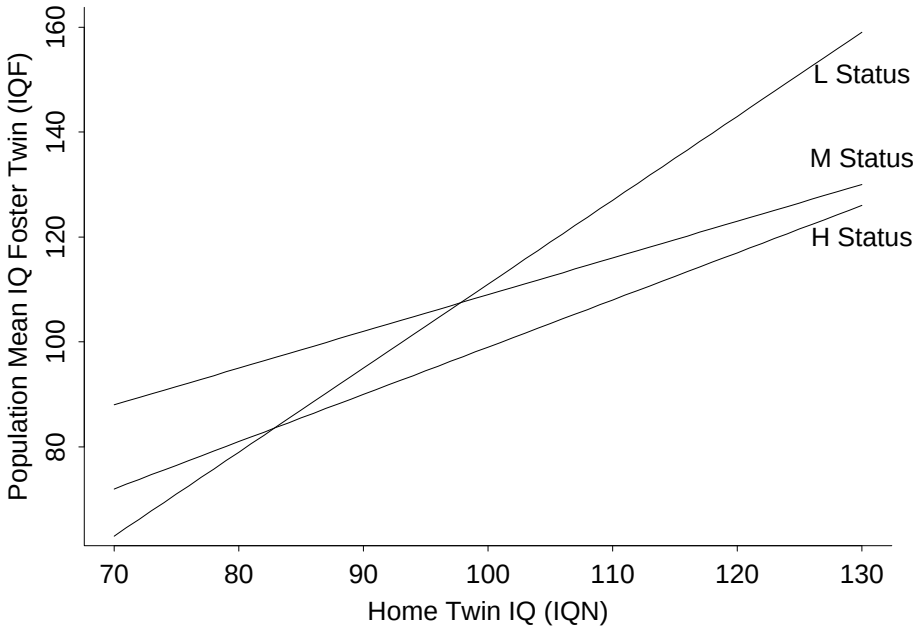
β_1 = difference between the intercepts of the H and L population regression lines.

β_2 = difference between the intercepts of the M and L population regression lines.

β_4 = difference between the slopes of the H and L population regression lines.

β_5 = difference between the slopes of the M and L population regression lines.

The plot gives a possible picture of the population regression lines corresponding to the general model (7.1).



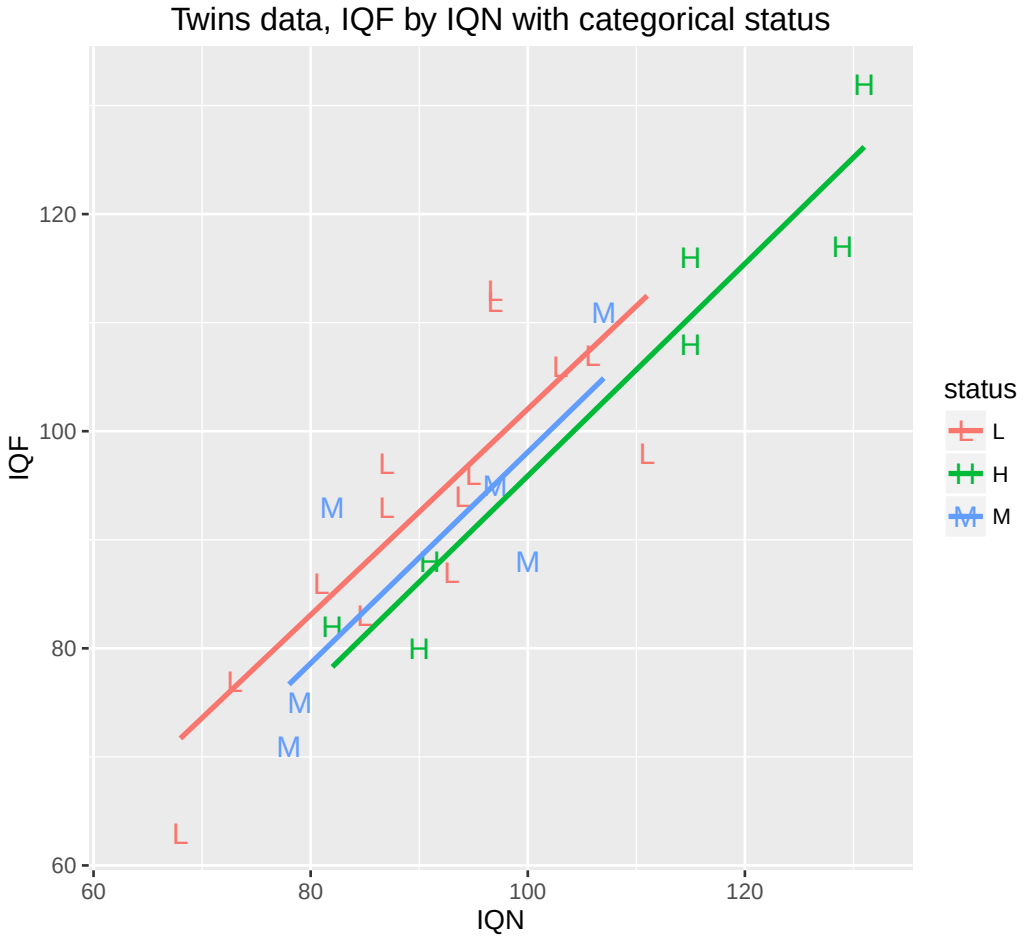
We fit the general model to the twins data.

```
#### Example: Twins
twins <- read.table("http://statacumen.com/teach/ADA2/ADA2_notes_Ch07_twins.dat"
                    , header = TRUE)
# set "L" as baseline level
twins$status <- relevel(twins$status, "L")
str(twins)
```

```
## 'data.frame': 27 obs. of  3 variables:
##  $ IQF   : int  82 80 88 108 116 117 132 71 75 93 ...
##  $ IQN   : int  82 90 91 115 115 129 131 78 79 82 ...
```

```
##  $ status: Factor w/ 3 levels "L","H","M": 2 2 2 2 2 2 2 3 3 3 ...
```

```
library(ggplot2)
p <- ggplot(twins, aes(x = IQN, y = IQF, colour = status, shape = status))
p <- p + geom_point(size=4)
  library(R.oo) # for ascii code lookup
  p <- p + scale_shape_manual(values=charToInt(sort(unique(twins$status))))
p <- p + geom_smooth(method = lm, se = FALSE)
p <- p + labs(title="Twins data, IQF by IQN with categorical status")
  # equal axes since x- and y-variables are same quantity
  dat.range <- range(twins[,c("IQF","IQN")])
  p <- p + xlim(dat.range) + ylim(dat.range) + coord_equal(ratio=1)
print(p)
```




```
lm.f.n.s.ns <- lm(IQF ~ IQN*status, data = twins)
library(car)
Anova(aov(lm.f.n.s.ns), type=3)
```

```
## Anova Table (Type III tests)
##
## Response: IQF
##              Sum Sq Df F value   Pr(>F)
## (Intercept)   11.61  1  0.1850   0.6715
## IQN         1700.39  1 27.1035 3.69e-05 ***
## status         8.99  2  0.0716   0.9311
```

```
## IQN:status     0.93  2  0.0074   0.9926
## Residuals   1317.47 21
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

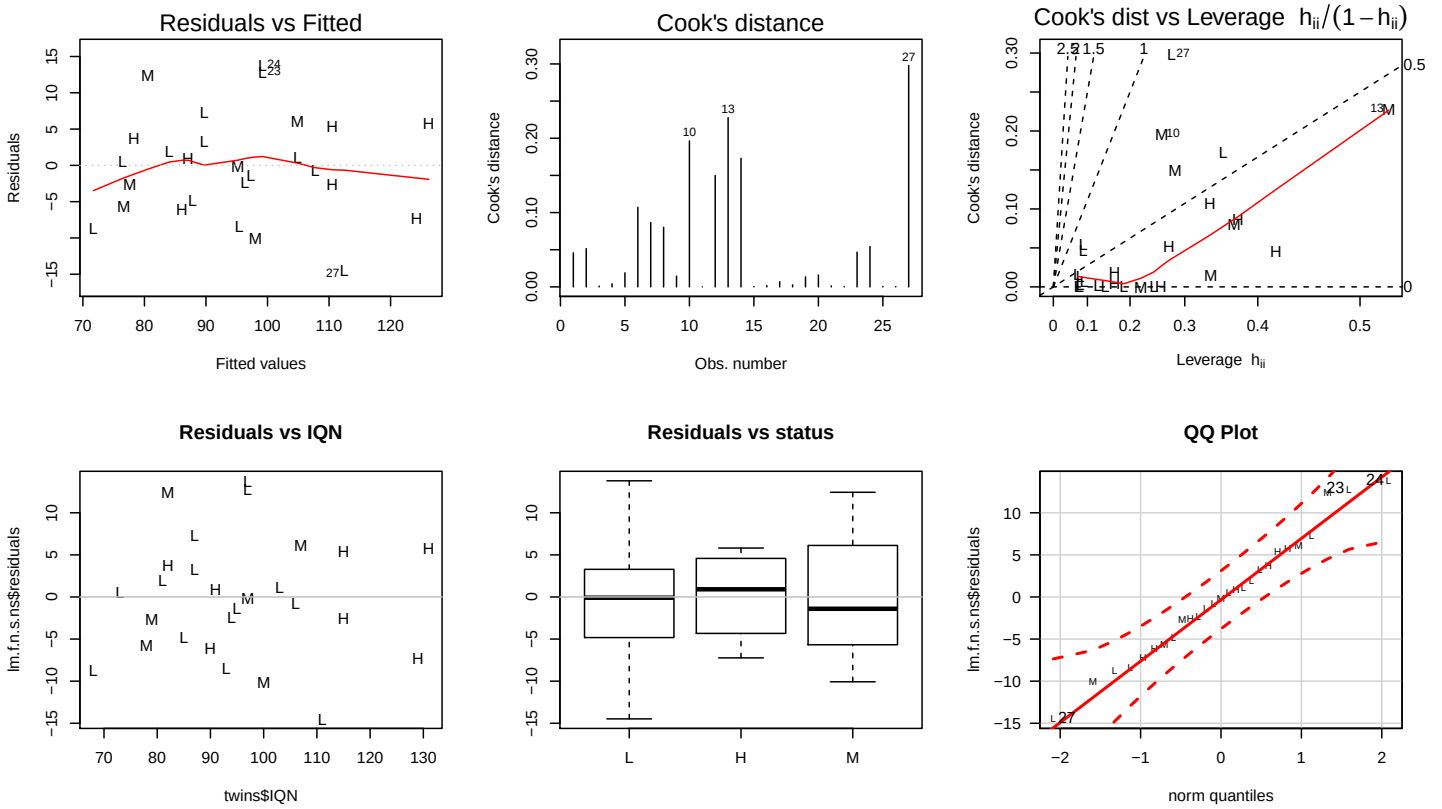
```
summary(lm.f.n.s.ns)
```

```
##
## Call:
## lm(formula = IQF ~ IQN * status, data = twins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.479  -5.248  -0.155   4.582  13.798
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.20461   16.75126   0.430    0.672
## IQN          0.94842    0.18218   5.206 3.69e-05 ***
## statusH     -9.07665   24.44870  -0.371    0.714
## statusM     -6.38859   31.02087  -0.206    0.839
## IQN:statusH  0.02914    0.24458   0.119    0.906
## IQN:statusM  0.02414    0.33933   0.071    0.944
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.921 on 21 degrees of freedom
## Multiple R-squared:  0.8041,Adjusted R-squared:  0.7574
## F-statistic: 17.24 on 5 and 21 DF,  p-value: 8.31e-07
```

```
# plot diagnostics
par(mfrow=c(2,3))
plot(lm.f.n.s.ns, which = c(1,4,6), pch=as.character(twins$status))

plot(twins$IQN, lm.f.n.s.ns$residuals, main="Residuals vs IQN", pch=as.character(twins$status))
  # horizontal line at zero
  abline(h = 0, col = "gray75")

plot(twins$status, lm.f.n.s.ns$residuals, main="Residuals vs status")
  # horizontal line at zero
  abline(h = 0, col = "gray75")

# Normality of Residuals
library(car)
qqPlot(lm.f.n.s.ns$residuals, las = 1, id.n = 3, main="QQ Plot", pch=as.character(twins$status))
```

```
## 27 24 23
##  1 27 26
```

```
## residuals vs order of data
#plot(lm.f.n.s.ns$residuals, main="Residuals vs Order of data")
#  # horizontal line at zero
#  abline(h = 0, col = "gray75")
```



The natural way to express the fitted model is to give separate prediction equations for the three status groups. Here is an easy way to get the separate fits. For the general model (7.1), the predicted IQF satisfies

$$\begin{aligned}\text{Predicted IQF} &= (\text{Intercept} + \text{Coeff for Status Indicator}) \\ &\quad + (\text{Coeff for Status Product Effect} + \text{Coeff for IQN}) \times \text{IQN}.\end{aligned}$$

For the baseline group, use 0 as the coefficients for the status indicator and product effect.

Thus, for the baseline group with status = L,

$$\begin{aligned}\text{Predicted IQF} &= 7.20 + 0 + (0.948 + 0)\ \text{IQN} \\ &= 7.20 + 0.948\ \text{IQN}.\end{aligned}$$

For the M status group with indicator I_2 and product effect $I_2 \times$ IQN:

$$\begin{aligned}\text{Predicted IQF} &= 7.20 - 6.39 + (0.948 + 0.024)\ \text{IQN} \\ &= 0.81 + 0.972\ \text{IQN}.\end{aligned}$$

For the H status group with indicator I_1 and product effect $I_1 \times$ IQN:

$$\begin{aligned} \text{Predicted IQF} &= 7.20 - 9.08 + (0.948 + 0.029)\ \text{IQN} \\ &= -1.88 + 0.977\ \text{IQN}. \end{aligned}$$

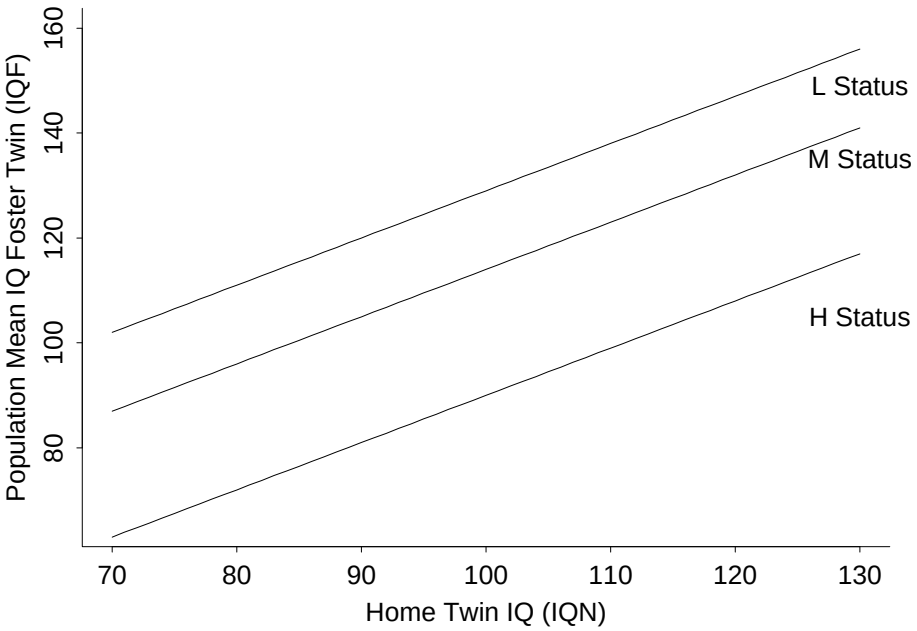
The LS lines are identical to separately fitting simple linear regressions to the three groups.

7.2.2 Equal slopes ANCOVA model

There are three other models of potential interest besides the general model. The **equal slopes** ANCOVA model

$$\text{IQF} = \beta_0 + \beta_1 I_1 + \beta_2 I_2 + \beta_3\ \text{IQN} + e$$

is a special case of (7.1) with $\beta_4 = \beta_5 = 0$ (no interaction). In the ANCOVA model, β_3 is the slope for all three regression lines. The other parameters have the same interpretation as in the general model (7.1), see the plot above. Output from the ANCOVA model is given below.



```
lm.f.n.s <- lm(IQF ~ IQN + status, data = twins)
library(car)
Anova(aov(lm.f.n.s), type=3)
```

```
## Anova Table (Type III tests)
##
## Response: IQF
##             Sum Sq Df F value    Pr(>F)
## (Intercept)   18.2  1  0.3181    0.5782
## IQN         4674.7  1 81.5521 5.047e-09 ***
## status       175.1  2  1.5276    0.2383
## Residuals   1318.4 23
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(lm.f.n.s)
```

```
##
## Call:
## lm(formula = IQF ~ IQN + status, data = twins)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.8235  -5.2366  -0.1111   4.4755  13.6978
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.6188     9.9628   0.564    0.578
## IQN           0.9658     0.1069   9.031 5.05e-09 ***
## statusH      -6.2264     3.9171  -1.590    0.126
## statusM      -4.1911     3.6951  -1.134    0.268
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.571 on 23 degrees of freedom
## Multiple R-squared:  0.8039,Adjusted R-squared:  0.7784
## F-statistic: 31.44 on 3 and 23 DF,  p-value: 2.604e-08
```

For the ANCOVA model, the predicted IQF for the three groups satisfies

$$\begin{aligned}\text{Predicted IQF} \;=\; &(\text{Intercept} + \text{Coeff for Status Indicator}) \\ &+(\text{Coeff for IQN}) \times \text{IQN}.\end{aligned}$$

As with the general model, use 0 as the coefficients for the status indicator and product effect for the baseline group.

For L status families:

$$\text{Predicted IQF} \;=\; 5.62 + 0.966\ \text{IQN},$$

for M status:

$$\begin{aligned} \text{Predicted IQF} &= 5.62 - 4.19 + 0.966 \text{ IQN} \\ &= 1.43 + 0.966 \text{ IQN}, \end{aligned}$$

and for H status:

$$\begin{aligned} \text{Predicted IQF} &= 5.62 - 6.23 + 0.966 \text{ IQN} \\ &= -0.61 + 0.966 \text{ IQN}. \end{aligned}$$

7.2.3 Equal slopes and equal intercepts ANCOVA model

The model with **equal slopes** and **equal intercepts**

$$\text{IQF} = \beta_0 + \beta_3 \text{ IQN} + e$$

is a special case of the ANCOVA model with $\beta_1 = \beta_2 = 0$. This model does not distinguish among social classes. The common intercept and slope for the social classes are β_0 and β_3, respectively.

The predicted IQF for this model is

$$\text{IQF} = 9.21 + 0.901 \text{ IQN}$$

for each social class.

```
lm.f.n <- lm(IQF ~ IQN, data = twins)
#library(car)
#Anova(aov(lm.f.n), type=3)
summary(lm.f.n)
```

```
##
## Call:
## lm(formula = IQF ~ IQN, data = twins)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.3512  -5.7311   0.0574   4.3244  16.3531
##
```

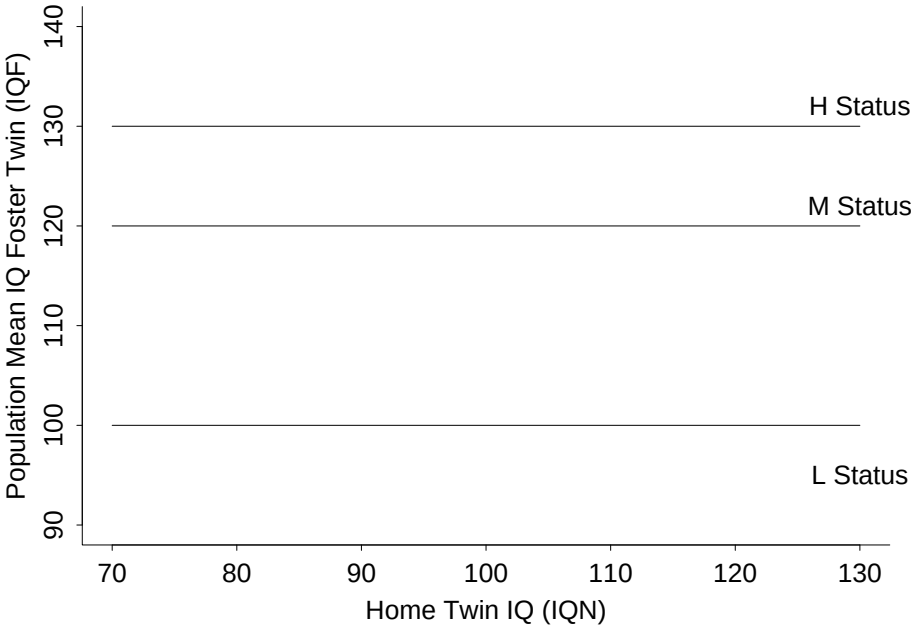
```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.20760    9.29990   0.990    0.332
## IQN          0.90144    0.09633   9.358  1.2e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.729 on 25 degrees of freedom
## Multiple R-squared:  0.7779,Adjusted R-squared:  0.769
## F-statistic: 87.56 on 1 and 25 DF,  p-value: 1.204e-09
```

7.2.4 No slopes, but intercepts ANCOVA model

The model with **no predictor** (IQN) effects

$$\text{IQF} \;=\; \beta_0 + \beta_1 I_1 + \beta_2 I_2 + e$$

is a special case of the ANCOVA model with $\beta_3 = 0$. In this model, social status has an effect on IQF but IQN does not. This model of **parallel regression lines** with **zero slopes** is identical to a one-way ANOVA model for the three social classes, where the intercepts play the role of the population means, see the plot below.



For the ANOVA model, the predicted IQF for the three groups satisfies

$$\text{Predicted IQF} \;=\; \text{Intercept} + \text{Coeff for Status Indicator}$$

Again, use 0 as the coefficients for the baseline status indicator.

For L status families:

$$\text{Predicted IQF} \;=\; 93.71,$$

for M status:

$$\begin{aligned}\text{Predicted IQF} &= 93.71 - 4.88 \\ &= 88.83,\end{aligned}$$

and for H status:

$$\begin{aligned}\text{Predicted IQF} &= 93.71 + 9.57 \\ &= 103.28.\end{aligned}$$

The predicted IQFs are the mean IQFs for the three groups.

```
lm.f.s <- lm(IQF ~ status, data = twins)
library(car)
Anova(aov(lm.f.s), type=3)
```

```
## Anova Table (Type III tests)
##
## Response: IQF
##             Sum Sq Df  F value Pr(>F)
## (Intercept) 122953  1 492.3772 <2e-16 ***
## status         732  2   1.4648 0.2511
## Residuals     5993 24
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(lm.f.s)
```

```
##
## Call:
## lm(formula = IQF ~ status, data = twins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.714 -12.274   2.286  12.500  28.714
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   93.714      4.223  22.190   <2e-16 ***
## statusH        9.571      7.315   1.308    0.203
## statusM       -4.881      7.711  -0.633    0.533
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.8 on 24 degrees of freedom
## Multiple R-squared:  0.1088,Adjusted R-squared:  0.03452
## F-statistic: 1.465 on 2 and 24 DF,  p-value: 0.2511
```

7.3 Relating Models to Two-Factor ANOVA

Recall the multiple regression formulation of the general model (7.1):

$$\text{IQF} = \beta_0 + \beta_1 I_1 + \beta_2 I_2 + \beta_3 \text{ IQN} + \beta_4 I_1 \text{ IQN} + \beta_5 I_2 \text{ IQN} + e. \quad (7.2)$$

If you think of β_0 as a grand mean, $\beta_1 I_1 + \beta_2 I_2$ as the status effect (i.e., the two indicators I_1 and I_2 allow you to differentiate among social classes), β_3 IQN as the IQN effect and $\beta_4 I_1$ IQN $+ \beta_5 I_2$ IQN as the status by IQN interaction, then

you can represent the model as

$$\begin{aligned} \text{IQF} &= \text{Grand Mean} + \text{Status Effect} + \text{IQN effect} \\ &\quad + \text{Status} \times \text{IQN interaction} + \text{Residual}. \end{aligned} \tag{7.3}$$

This representation has the same form as a two-factor ANOVA model with interaction, except that IQN is a quantitative effect rather than a qualitative (i.e., categorical) effect. The general model has the same structure as a two-factor interaction ANOVA model because the plot of the population means allows non-parallel profiles. However, the general model is a special case of the two-factor interaction ANOVA model because it restricts the means to change linearly with IQN.

The ANCOVA model has main effects for status and IQN but no interaction:

$$\text{IQF} = \text{Grand Mean} + \text{Status Effect} + \text{IQN effect} + \text{Residual}. \tag{7.4}$$

The ANCOVA model is a special case of the additive two-factor ANOVA model because the plot of the population means has parallel profiles, but is not equivalent to the additive two-factor ANOVA model.

The model with equal slopes and intercepts has no main effect for status nor an interaction between status and IQN:

$$\text{IQF} = \text{Grand Mean} + \text{IQN effect} + \text{Residual}. \tag{7.5}$$

The one-way ANOVA model has no main effect for IQN nor an interaction between status and IQN:

$$\text{IQF} = \text{Grand Mean} + \text{Status Effect} + \text{Residual}. \tag{7.6}$$

I will expand on these ideas later, as they are useful for understanding the connections between regression and ANOVA models.

7.4 Choosing Among Models

I will suggest a backward sequential method to select which of models (7.1), (7.4), and (7.5) fits best. You would typically be interested in the one-way

ANOVA model (7.6) only when the effect of IQN was negligible.

Step 1: Fit the full model (7.1) and test the hypothesis of equal slopes $H_0 : \beta_4 = \beta_5 = 0$. (aside: t-tests are used to test **either** $\beta_4 = 0$ or $\beta_5 = 0$.) To test H_0, eliminate the predictor variables I_1 IQN and I_2 IQN associated with β_4 and β_5 from the full model (7.1). Then fit the reduced model (7.4) with equal slopes. Reject $H_0 : \beta_4 = \beta_5 = 0$ if the increase in the Residual SS obtained by deleting I_1 IQN and I_2 IQN from the full model is significant. Formally, compute the F-statistic:

$$F_{obs} = \frac{(\text{ERROR SS for reduced model} - \text{ERROR SS for full model})/2}{\text{ERROR MS for full model}}$$

and compare it to an upper-tail critical value for an F-distribution with 2 and *df* degrees of freedom, where *df* is the Residual *df* for the full model. The F-test is a direct extension of the single degree-of-freedom F-tests in the stepwise fits. A p-value for F-test is obtained from `library(car)` with `Anova(aov(LMOBJECT), type=3)` for the interaction. If H_0 is rejected, stop and conclude that the population regression lines have different slopes (and then I do not care whether the intercepts are equal). Otherwise, proceed to step 2.

Step 2: Fit the equal slopes or ANCOVA model (7.4) and test for equal intercepts $H_0 : \beta_1 = \beta_2 = 0$. Follow the procedure outlined in Step 1, treating the ANCOVA model as the full model and the model $IQF = \beta_0 + \beta_3$ IQN $+ e$ with equal slopes and intercepts as the reduced model. See the intercept term using `library(car)` with `Anova(aov(LMOBJECT), type=3)`. If H_0 is rejected, conclude that that population regression lines are parallel with unequal intercepts. Otherwise, conclude that regression lines are identical.

Step 3: Estimate the parameters under the appropriate model, and conduct a diagnostic analysis. Summarize the fitted model by status class.

A comparison of regression lines across $k > 3$ groups requires $k-1$ indicator variables to define the groups, and $k - 1$ interaction variables, assuming the model has a single predictor. The comparison of models mimics the discussion above, except that the numerator of the F-statistic is divided by $k - 1$ instead

of 2, and the numerator *df* for the F-test is $k-1$ instead of 2. If $k=2$, the F-tests for comparing the three models are equivalent to $t-$tests given with the parameter estimates summary. For example, recall how you tested for equal intercepts in the tools problems.

The plot of the twins data shows fairly linear relationships within each social class. The linear relationships appear to have similar slopes and similar intercepts. The p-value for testing the hypothesis that the slopes of the population regression lines are equal is essentially 1. The observed data are consistent with the reduced model of equal slopes.

The p-value for comparing the model of equal slopes and equal intercepts to the ANCOVA model is 0.238, so there is insufficient evidence to reject the reduced model with equal slopes and intercepts. The estimated regression line, regardless of social class, is:

$$\text{Predicted IQF} = 9.21 + 0.901\text{*IQN}.$$

There are no serious inadequacies with this model, based on a diagnostic analysis (not shown).

An interpretation of this analysis is that the natural parents' social class has no impact on the relationship between the IQ scores of identical twins raised apart. What other interesting features of the data would be interesting to explore? For example, what values of the intercept and slope of the population regression line are of intrinsic interest?

7.4.1 Simultaneous testing of regression parameters

In the twins example, we have this full interaction model,

$$\text{IQF} = \beta_0 + \beta_1 I_1 + \beta_2 I_2 + \beta_3\ \text{IQN} + \beta_4 I_1\ \text{IQN} + \beta_5 I_2\ \text{IQN} + e, \quad (7.7)$$

where $I_1 = 1$ indicates H, and $I_2 = 1$ indicates M, and L is the baseline status.

Consider these two specific hypotheses:

1. H_0 : equal regression lines for status M and L
2. H_0 : equal regression lines for status M and H

That is, the intercept and slope for the regression lines are equal for the pairs of status groups.

First, it is necessary to formulate these hypotheses in terms of testable parameters. That is, find the β values that make the null hypothesis true in terms of the model equation.

1. $H_0 : \beta_2 = 0$ and $\beta_5 = 0$
2. $H_0 : \beta_1 = \beta_2$ and $\beta_4 = \beta_5$

Using linear model theory, there are methods for testing these multiple-parameter hypothesis tests.

One strategy is to use the Wald test of null hypothesis $\mathbf{r}\underset{\sim}{\beta} = \underset{\sim}{r}$, where $\mathbf{r}$ is a matrix of contrast coefficients (typically $+1$ or -1), $\underset{\sim}{\beta}$ is our vector of regression β coefficients, and $\underset{\sim}{r}$ is a hypothesized vector of what the linear system $\mathbf{r}\underset{\sim}{\beta}$ equals. For our first hypothesis test, the linear system we're testing in matrix notation is

$$\begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Let's go about testing another hypothesis, first, using the Wald test, then we'll test our two simultaneous hypotheses above.

- H_0 : equal slopes for all status groups
- $H_0 : \beta_4 = \beta_5 = 0$

```
lm.f.n.s.ns <- lm(IQF ~ IQN*status, data = twins)
library(car)
Anova(aov(lm.f.n.s.ns), type=3)
```

```
## Anova Table (Type III tests)
##
## Response: IQF
##              Sum Sq Df F value   Pr(>F)
## (Intercept)   11.61  1  0.1850   0.6715
## IQN         1700.39  1 27.1035 3.69e-05 ***
```

```
## status         8.99  2  0.0716   0.9311
## IQN:status     0.93  2  0.0074   0.9926
## Residuals   1317.47 21
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# beta coefficients (term positions: 1, 2, 3, 4, 5, 6)
coef(lm.f.n.s.ns)
```

```
## (Intercept)         IQN     statusH     statusM IQN:statusH
##  7.20460986  0.94842244 -9.07665352 -6.38858548  0.02913971
## IQN:statusM
##  0.02414450
```

The test for the interaction above (IQN:status) has a p-value=0.9926, which indicates that common slope is reasonable. In the Wald test notation, we want to test whether those last two coefficients (term positions 5 and 6) both equal 0. Here we get the same result as the ANOVA table.

```
library(aod) # for wald.test()
# Typically, we are interested in testing whether individual parameters or
#   set of parameters are all simultaneously equal to 0s
# However, any null hypothesis values can be included in the vector coef.test.values.
coef.test.values <- rep(0, length(coef(lm.f.n.s.ns)))
wald.test(b = coef(lm.f.n.s.ns) - coef.test.values
        , Sigma = vcov(lm.f.n.s.ns)
        , Terms = c(5,6))
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 0.015, df = 2, P(> X2) = 0.99
```

Now to our two simultaneous hypotheses. In hypothesis 1 we are testing $\beta_2 = 0$ and $\beta_5 = 0$, which are the 3rd and 6th position for coefficients in our original equation (7.7). *However, we need to choose the correct positions based on the* `coef()` *order, and these are positions 4 and 6.* The large p-value=0.55 suggests that M and L can be described by the same regression line, same slope and intercept.

```
library(aod) # for wald.test()
coef.test.values <- rep(0, length(coef(lm.f.n.s.ns)))
wald.test(b = coef(lm.f.n.s.ns) - coef.test.values
        , Sigma = vcov(lm.f.n.s.ns)
        , Terms = c(4,6))
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 1.2, df = 2, P(> X2) = 0.55
```

```
# Another way to do this is to define the matrix r and vector r, manually.
mR <- as.matrix(rbind(c(0, 0, 0, 1, 0, 0), c(0, 0, 0, 0, 0, 1)))
mR
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]    0    0    0    1    0    0
## [2,]    0    0    0    0    0    1
```

```
vR <- c(0, 0)
vR
```

```
## [1] 0 0
```

```
wald.test(b = coef(lm.f.n.s.ns)
        , Sigma = vcov(lm.f.n.s.ns)
        , L = mR, H0 = vR)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 1.2, df = 2, P(> X2) = 0.55
```

In hypothesis 2 we are testing $\beta_1 - \beta_2 = 0$ and $\beta_4 - \beta_5 = 0$ which are the difference of the 2nd and 3rd coefficients and the difference of the 5th and 6th coefficients. *However, we need to choose the correct positions based on the* `coef()` *order, and these are positions 3 and 4, and 5 and 6.* The large p-value=0.91 suggests that M and H can be described by the same regression line, same slope and intercept.

```
mR <- as.matrix(rbind(c(0, 0, 1, -1, 0, 0), c(0, 0, 0, 0, 1, -1)))
mR
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]    0    0    1   -1    0    0
## [2,]    0    0    0    0    1   -1
```

```
vR <- c(0, 0)
vR
```

```
## [1] 0 0
```

```
wald.test(b = coef(lm.f.n.s.ns)
        , Sigma = vcov(lm.f.n.s.ns)
        , L = mR, H0 = vR)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 0.19, df = 2, P(> X2) = 0.91
```

The results of these tests are not surprising, given our previous analysis where we found that the status effect is not significant for all three groups.

Any simultaneous linear combination of parameters can be tested in this

way.

7.5 Comments on Comparing Regression Lines

In the twins example, I defined two indicator variables (plus two interaction variables) from an ordinal categorical variable: status (H, M, L). Many researchers would assign numerical codes to the status groups and use the coding as a predictor in a regression model. For status, a "natural" coding might be to define NSTAT=0 for L, 1 for M, and 2 for H status families. This suggests building a multiple regression model with a single status variable (i.e., single df):

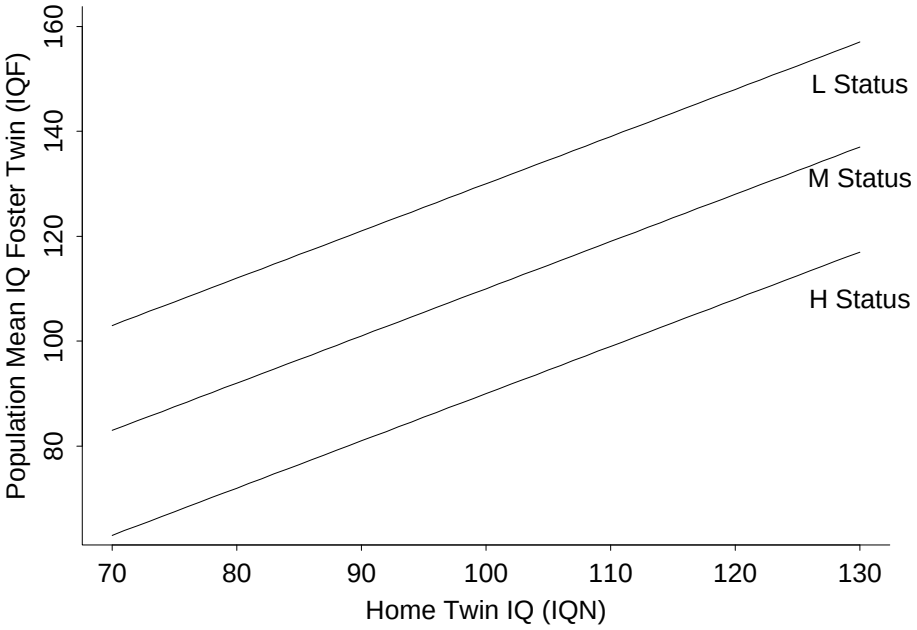
$$\text{IQF} = \beta_0 + \beta_1 \text{ IQN} + \beta_2 \text{NSTAT} + e.$$

If you consider the status classes separately, the model implies that

$$\begin{aligned}
\text{IQF} &= \beta_0 + \beta_1 \text{ IQN} + \beta_2(0) + e = \beta_0 + \beta_1 \text{ IQN} + e && \text{for L status,} \\
\text{IQF} &= \beta_0 + \beta_1 \text{ IQN} + \beta_2(1) + e = (\beta_0 + \beta_2) + \beta_1 \text{ IQN} + e && \text{for M status,} \\
\text{IQF} &= \beta_0 + \beta_1 \text{ IQN} + \beta_2(2) + e = (\beta_0 + 2\beta_2) + \beta_1 \text{ IQN} + e && \text{for H status.}
\end{aligned}$$

The model assumes that the IQF by IQN regression lines are parallel for the three groups, and are separated by a constant β_2. This model is more restrictive (and less reasonable) than the ANCOVA model with equal slopes but arbitrary intercepts. Of course, this model is a easier to work with because it requires keeping track of only one status variable instead of two status indicators.

A plot of the population regression lines under this model is given above, assuming $\beta_2 < 0$.



7.6 Three-way interaction

In this example, a three-way interaction is illustrated with two categorical variables and one continuous variable. Let a take values 0 or 1 (it's an indicator variable), b take values 0 or 1, and c be a continuous variable taking any value.

Below are five models:

(1) Interactions: ab. All lines parallel, different intercepts for each (a, b) combination.

(2) Interactions: ab, ac. (a, c) combinations have parallel lines, different intercepts for each (a, b) combination.

(3) Interactions: ab, bc. (b, c) combinations have parallel lines, different intercepts for each (a, b) combination.

(4) Interactions: ab, ac, bc. All combinations may have different slope lines with different intercepts, but difference in slope between $b = 0$ and $b = 1$ is similar for each a group (and vice versa).

(5) Interactions: ab, ac, bc, abc. All combinations may have different slope lines with different intercepts.

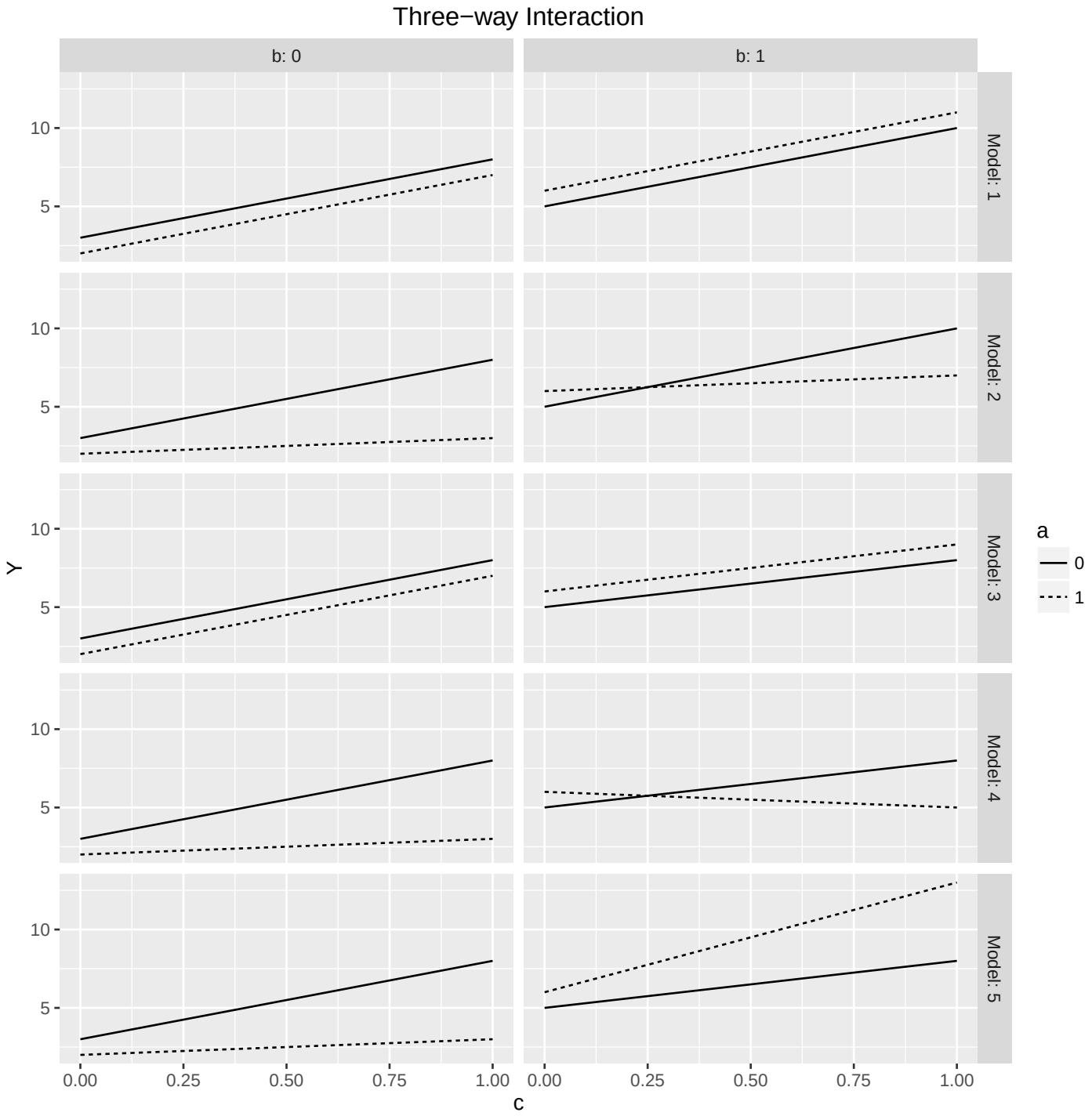
Model		Intercepts				Slopes for c			
(1)	$y =$	β_0	$+\beta_1 a$	$+\beta_2 b$	$+\beta_3 ab$	$+\beta_4 c$			
(2)	$y =$	β_0	$+\beta_1 a$	$+\beta_2 b$	$+\beta_3 ab$	$+\beta_4 c$	$+\beta_5 ac$		
(3)	$y =$	β_0	$+\beta_1 a$	$+\beta_2 b$	$+\beta_3 ab$	$+\beta_4 c$		$+\beta_6 bc$	
(4)	$y =$	β_0	$+\beta_1 a$	$+\beta_2 b$	$+\beta_3 ab$	$+\beta_4 c$	$+\beta_5 ac$	$+\beta_6 bc$	
(5)	$y =$	β_0	$+\beta_1 a$	$+\beta_2 b$	$+\beta_3 ab$	$+\beta_4 c$	$+\beta_5 ac$	$+\beta_6 bc$	$+\beta_7 abc$

```
X <- expand.grid(c(0,1),c(0,1),c(0,1))
X <- cbind(1, X)
colnames(X) <- c("one", "a", "b", "c")
X$ab  <- X$a * X$b
X$ac  <- X$a * X$c
X$bc  <- X$b * X$c
X$abc <- X$a * X$b * X$c
X <- as.matrix(X)
X <- X[,c(1,2,3,5,4,6,7,8)] # reorder columns to be consistent with table above
#£

vbeta <- matrix(c(3, -1, 2, 2, 5, -4, -2, 8), ncol = 1)
rownames(vbeta) <- paste("beta", 0:7, sep="")

Beta <- matrix(vbeta, nrow = dim(vbeta)[1], ncol = 5)
rownames(Beta) <- rownames(vbeta)

# Beta vector for each model
Beta[c(6,7,8), 1] <- 0
Beta[c(  7,8), 2] <- 0
Beta[c(6,  8), 3] <- 0
Beta[c(    8), 4] <- 0

colnames(Beta) <- 1:5 #paste("model", 1:5, sep="")

# Calculate response values
Y <- X %*% Beta

library(reshape2)
YX <- data.frame(cbind(melt(Y), X[,"a"], X[,"b"], X[,"c"]))
colnames(YX) <- c("obs", "Model", "Y", "a", "b", "c")
YX$a <- factor(YX$a)
YX$b <- factor(YX$b)
```

These are the β values used for this example.

beta0	beta1	beta2	beta3	beta4	beta5	beta6	beta7
3	−1	2	2	5	−4	−2	8

```
library(ggplot2)
p <- ggplot(YX, aes(x = c, y = Y, group = a))
#p <- p + geom_point()
p <- p + geom_line(aes(linetype = a))
p <- p + labs(title = "Three-way Interaction")
p <- p + facet_grid(Model ~ b, labeller = "label_both")
print(p)
```


Three−way Interaction
b: 0
b: 1
Model: 1
Model: 2
Model: 3
Model: 4
Model: 5
Y
c
a
0
1
10
5
0.00
0.25
0.50
0.75
1.00