Level M Regression Models Lecture 14

We will see several examples and consider the `R` output for the three models discussed in the previous lecture and the interpretation of them.

R code

Please find some R code used to fit and plot some models. You can find most data sets on Moodle. Please download data and try some of the R code as you read through these notes.

Trout Data

When the behaviour of a group of trout is studied, some fish are observed to become dominant and others to become subordinate. Dominant fish (coded as group 1) have freedom of movement whereas subordinate fish (coded as group 2) tend to congregate in the periphery of the waterway to avoid crossing the path of the dominant fish, Metcalfe (1986).

Data on energy expenditure and ration of food obtained were collected as part of a laboratory experiment for 20 trout, and are plotted below. Both variables are measured in units of calories per kilo-calorie per trout per day.

```r
trout <- read.csv("TROUT.csv")
trout$Group=factor(trout$Group)
ggplot(data=trout, aes(y=Energy,x= Ration, group=Group, color=Group,shape= Group)) +
  geom_point(size = 3, alpha = .8) +
  geom_smooth(method="lm", fill=NA, fullrange=TRUE)+
   theme(plot.background = element_rect(
          fill = "transparent",
          colour = NA,
          size = 1),    legend.position = c(0.15,0.85),
        legend.background = element_rect(fill="transparent", size=0.5,
                                          linetype="solid")) +
        scale_shape_discrete(breaks=c("1", "2"),
                             labels=c( "Dominant","Subordinate")) +
        scale_color_discrete(breaks=c("1", "2"),
                             labels=c( "Dominant","Subordinate"))
```
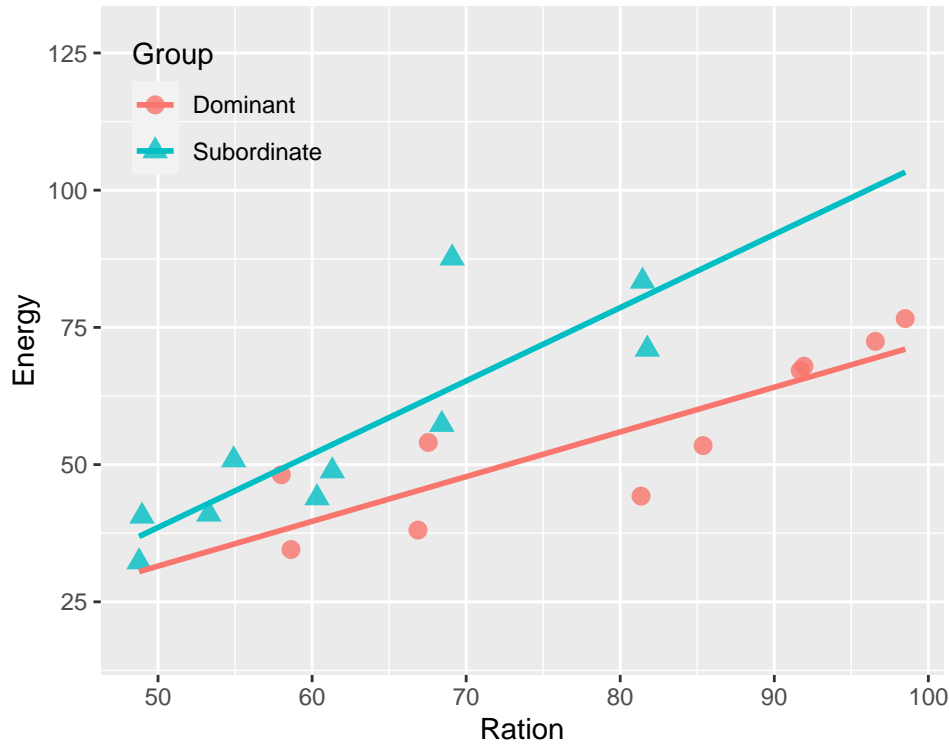
Based on the scatterplot, it appears that a model with parallel lines to be an acceptable description of the data. We will now fit the parallel lines model and examine it. Summary statistics are provided below and have been calculated from the data.

$$\bar{x}_{1.} = 79.644,\ S_{x_1 y_1} = 1772.273,\ \bar{y}_{1.} = 55.665,\ S_{x_1 x_1} = 2193.056$$

$$\bar{x}_{2.} = 62.829,\ S_{x_2 y_2} = 1787.014,\ \bar{y}_{2.} = 55.678,\ S_{x_2 x_2} = 1327.703$$

$$\sum_{i=1}^{2} \sum_{j=1}^{n_i} y_{ij}^2 = 67166.63,\ n_1 = n_2 = 10$$

Calculate a 95% confidence interval that can be used to assess whether the parallel line model should be reduced to a single line.

Based on the summary statistics

2

$$n_1 = n_2 = 10$$

$$\hat{\beta} = \frac{S_{x_1y_1} + S_{x_2y_2}}{S_{x_1x_1} + S_{x_2x_2}} = 1.010943$$

$$\hat{\alpha}_1 = \bar{y}_{1.} = 55.665$$

$$\hat{\alpha}_2 = \bar{y}_{2.} = 55.678$$

$$b = \begin{pmatrix} 1 \\ -1 \\ 62.829 - 79.644 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ -16.815 \end{pmatrix}$$

$$RSS = \sum_i \sum_j y_{ij}^2 - n_1\bar{y}_{1.}^2 - n_2\bar{y}_{2.}^2 - \frac{(S_{x_1y_1} + S_{x_2y_2})^2}{S_{x_1x_1} + S_{x_2x_2}}$$

$$= 67166.63 - 10 \times 55.665^2 - 10 \times 55.678^2 - \frac{(1772.273 + 1787.014)^2}{(2193.056 + 1327.703)}$$

$$= 1582.074$$

$$\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b} = \begin{pmatrix} 1 & -1 & -16.815 \end{pmatrix} \begin{pmatrix} \frac{1}{10} & 0 & 0 \\ 0 & \frac{1}{10} & 0 \\ 0 & 0 & \frac{1}{5180.314} \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ -16.815 \end{pmatrix}$$

$$= \frac{1}{10} + \frac{1}{10} + \frac{(-16.815)^2}{3520.759}$$

$$= 0.2803$$

The 95% C.I. is therefore

$$\hat{\alpha}_1 - \hat{\alpha}_2 + \hat{\beta}(\bar{x}_{2.} - \bar{x}_{1.}) \pm t(n - p, 0.975)\sqrt{\left(\frac{RSS}{n - p}\right)\left(\frac{1}{n_1} + \frac{1}{n_2} + \frac{(\bar{x}_{2.} - \bar{x}_{1.})^2}{S_{x_1x_1} + S_{x_2x_2}}\right)}$$

$$55.665 - 55.678 + 1.0109(-16.815) \pm t(17; 0.975)\sqrt{\frac{1582.074}{17}0.2803}$$

$$-17.0113 \pm 10.7757$$

$$(-27.79, -6.24)$$

The above interval does not contain zero so we should keep the parallel line models. This interval in entirely negative. Recall this is an interval estimate for the intercept term for group 1 (dominant) minus group 2 (subordinate).

Therefore the intercept term for regression line for group 1 (dominant) is significantly lower than that of group 2 (subordinate). This suggests that the regression line for group 1 (dominant) lies significantly lower than that of group 2 (subordinate).

This suggests that for any given amount of food obtained, x, dominant fish on average expend less energy than subordinate fish.

Now, we see how we can use R output to come to the same conclusion.

```
trout.lm<-lm(Energy~Ration+Group,data=trout)
trout.lm.single<-lm(Energy~Ration,data=trout)
summary(trout.lm)
```

```
##
## Call:
## lm(formula = Energy ~ Ration + Group, data = trout)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.130  -5.139  -0.870   2.199  25.622
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -24.8506    13.3031  -1.868  0.07910 .
## Ration        1.0109     0.1626   6.218 9.36e-06 ***
## Group2       17.0120     5.1075   3.331  0.00396 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.647 on 17 degrees of freedom
## Multiple R-squared:  0.6946, Adjusted R-squared:  0.6587
## F-statistic: 19.33 on 2 and 17 DF,  p-value: 4.182e-05
```

The first indicator that we should keep the two parallel lines is the fact that the term Group is significant in the summary of the output of model trout.lm<-lm(Energy~Ration+Group,data=trout). In particular the p-value is 0.00396 which is less than 0.05.

The model fitted using R is specifed slightly different to that defined in the previous lecture.

95% Confidence interval for parallel lines model

Recall how we specified the parallel regression model in order to obtain a 95% Confidence interval for parallel lines model.

Data: (y_{ij}, x_{ij}); $i = 1, 2$; $j = 1, \dots, n_i$.

Model : $E(Y_{ij}) = \alpha_i + \beta(x_{ij} - \bar{x}_{i.})$

Paralle lines model in R

Compare this to the specific model we have specifed in R.

Data: $(y_j, x_j, \text{group}_j)$; $j = 1, \dots, n_1 + n_2$.

Model : $E(Y_j) = a + bx_j + cI_{\text{group}_j}$

where

$$I_{group_j} = \begin{cases} 1 & \text{if group}_j = 2 \\ 0 & \text{if group}_j = 1 \end{cases}$$

From the R output

$$\hat{a} = -24.8506$$
$$\hat{b} = 1.0109$$
$$\hat{c} = 17.0120$$

and from there, we can deduce that $\hat{\beta} = 1.0109$ and the intercept term in relation to the dominant group (1)is -24.8506 and the subordinate group (2) intercept term for the is -24.9506+17.0120=-7.9386. Therefore

$$\hat{\alpha_1} = -24.8506 + 1.0109 \times 55.665 = 31.42115$$
$$\hat{\alpha_2} = -7.9386 + 1.0109 \times 55.678 = 48.34629$$
$$\hat{\beta} = 1.0109$$

Notice that the difference between the two groups from the R output is 17.0120. The coding Group2 tells us that group 2 is signficantly higher than group 1 (since the p-value < 0.05). Therefore, accounting for Ration group 2 fish on average expend more energy that group 1 fish.

We can obtain the same conclusion using the anova function. In this case we have two competing models. Model 1 is the model including Ration and Group and Model 2 is the model including only Ration. When given a sequence of models in this way, anova tests the models against one another. The null model is the simpler of the two (Model 2). Given the p-value of 0.003958, we have evidence to reject this model.

```
print(anova(trout.lm,trout.lm.single))
```

```
## Analysis of Variance Table
##
## Model 1: Energy ~ Ration + Group
## Model 2: Energy ~ Ration
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     17 1582.1
## 2     18 2614.5 -1   -1032.5 11.094 0.003958 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Let's now plot the data again. This time, we will include 95% confidence interval bands for each group.

```
fit<-predict(trout.lm, interval="confidence")
trout.fit<-data.frame(trout,fit)

ggplot(data=trout.fit, aes(x=Ration,y= Energy, colour = Group, shape= Group))+
```

```
geom_point(size = 3, alpha = .8) +
geom_line(aes(x=Ration,y=fit),size=1.5)+
geom_ribbon(aes(ymin = lwr, ymax = upr), linetype = "dotted", alpha = I(0.2)) + #For CI
 theme(plot.background = element_rect(
        fill = "transparent",
        colour = NA,
        size = 1),    legend.position = c(0.15,0.85),
      legend.background = element_rect(fill="transparent", size=0.5, linetype="solid")) +
guides(color = guide_legend(label.position = "right") )
```
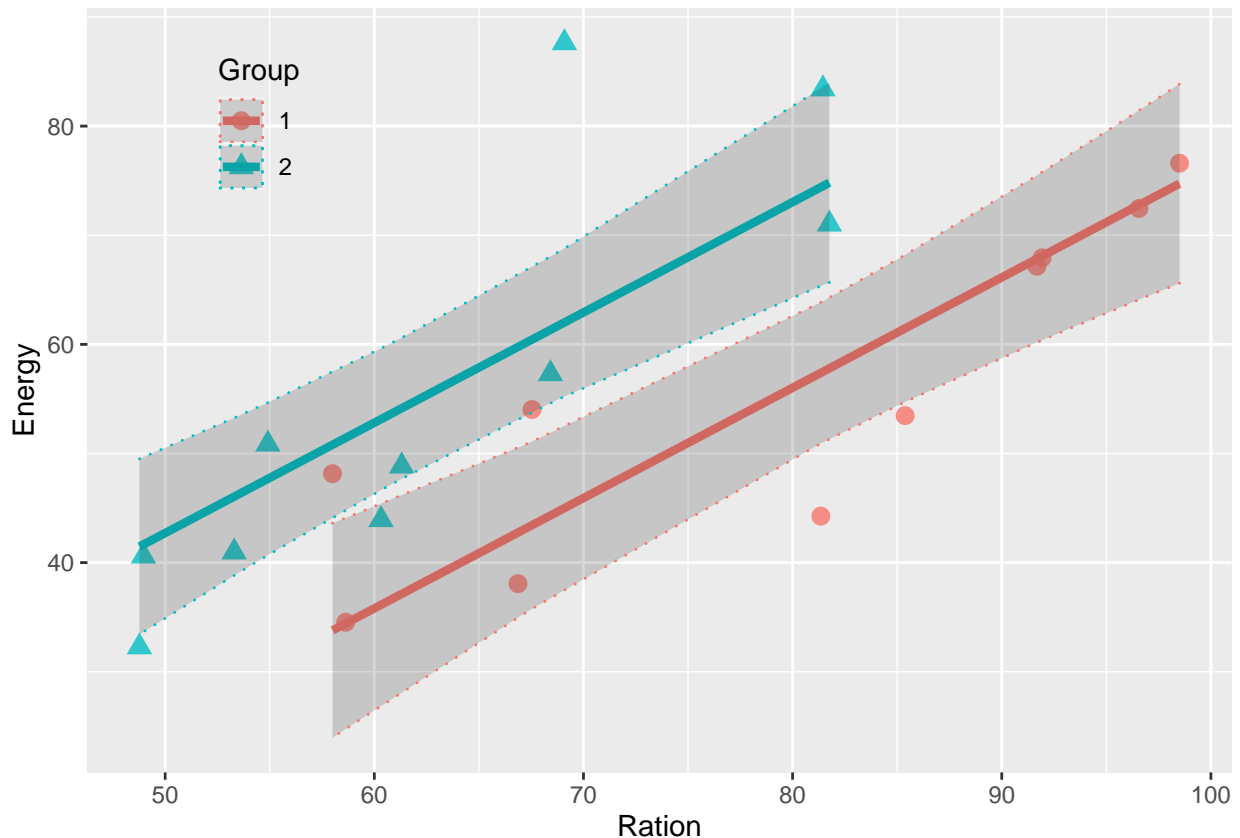

Notice the two CI's do not overlap over the whole range. So our conclusion would be to keep both lines.

Respiratory Distress Syndrome

Premature babies often suffer from a variety of problems and respiratory distress syndrome (RDS) is a common, and serious, lung problem. It is thought that the occurrence of this syndrome might be related to a property of the blood called red cell deformability. This refers to the ability of red cells to change shape to pass through small pores. The rate (Lrate, on a log scale) of blood flow through a set of 3μm pores is recorded for two groups of babies, some of whom suffer from respiratory distress syndrome (RDS) and some who do not (No RDS). The gestational age (GA) in weeks of each baby is also recorded. These data were kindly provided by Queen Mother's Hospital, Glasgow.
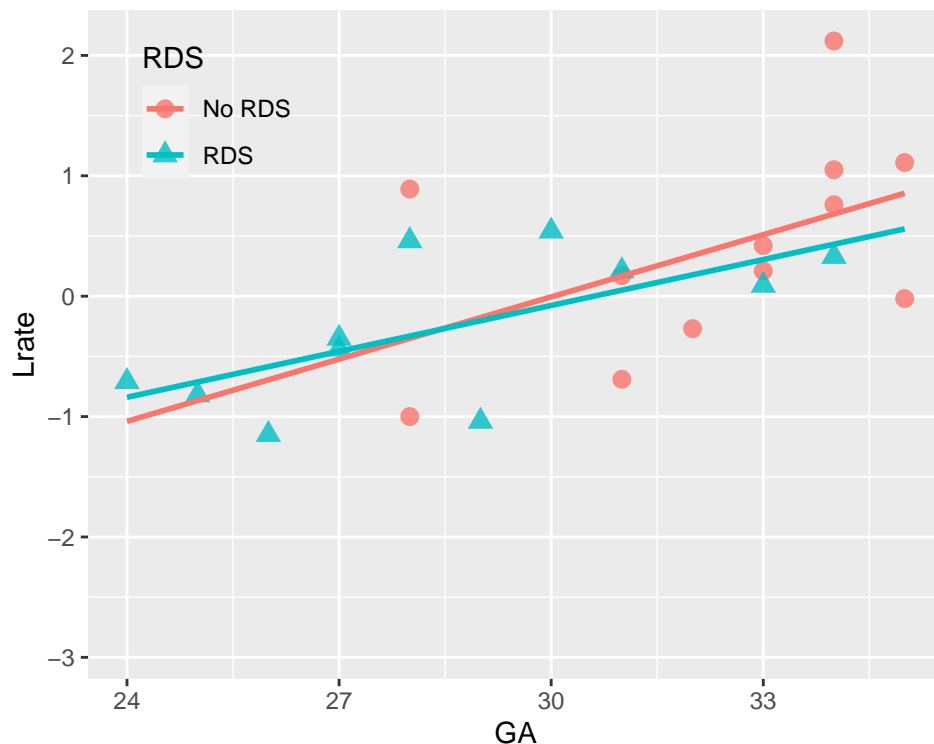
We proposed that the relationship between `Lrate` and `GA` does not appear to differ by whether the baby did or didn't suffer from respiratory distress syndrome (RDS). To examine this formally we can use the following summary numbers for the data.

$$\bar{y}_{1.} = 0.395833 \quad \bar{y}_{2.} = -0.2618182$$
$$\bar{x}_{1.} = 32.333 \quad \bar{x}_{2.} = 28.54545$$
$$S_{y_1y_1} = 8.117929 \quad S_{y_2y_2} = 3.806964$$
$$S_{x_1x_1} = 64.66667 \quad S_{x_2x_2} = 102.7273$$
$$S_{x_1y_1} = 11.13667 \quad S_{x_2y_2} = 13.06091$$
$$n_1 = 12 \quad n_2 = 11$$

For the "different lines" model we have

$$\hat{\beta}_1 = 0.1722$$

$$\hat{\beta}_2 = 0.1271$$

$$RSS = 6.199374 + 2.146379 = 8.345753$$

A C.I. for the difference in slopes is

$$0.0451 \pm 2.09 \sqrt{\frac{8.345753}{19} \left(\frac{1}{64.66667} + \frac{1}{102.7273} \right)}$$

i.e. (0.0451 ± 0.22) and hence (-0.17, 0.27).

Fitting a model in R for a response of rate of blood flow, with a continuous covariate for gestational age (GA), a factor covariate for RDS group (RDS) and an interaction between gestational age and RDS group provides the output on the following page.

```
resp.separate <- lm(Lrate~GA*RDS,data=rds)
summary(resp.separate)
```

```
##
## Call:
## lm(formula = Lrate ~ GA * RDS, data = rds)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.87508 -0.43260  0.00379  0.20733  1.43714
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.17250    2.67167  -1.936   0.0679 .
## GA           0.17222    0.08242   2.090   0.0503 .
## RDSRDS       1.28137    3.26526   0.392   0.6991
## GA:RDSRDS   -0.04507    0.10521  -0.428   0.6731
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6628 on 19 degrees of freedom
## Multiple R-squared:  0.4207, Adjusted R-squared:  0.3292
## F-statistic: 4.599 on 3 and 19 DF,  p-value: 0.01392
```

```
print(anova(resp.separate))
```

```
## Analysis of Variance Table
##
## Response: Lrate
##           Df Sum Sq Mean Sq F value   Pr(>F)
## GA         1 5.9335  5.9335 13.5081 0.001608 **
## RDS        1 0.0466  0.0466  0.1062 0.748129
## GA:RDS     1 0.0806  0.0806  0.1836 0.673147
## Residuals 19 8.3458  0.4393
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Neither the confidence interval and p-value for the interaction terms provide enough evidence to conclude that the two seperate regressions model is appropriate, so we will instead try to fit the "parallel lines" model to the data.

We now have

$$\hat{\beta} = 0.1446$$

$$RSS = 8.426383$$

A C.I. for the vertical separation between the regression lines is

$$0.110 \pm 2.09 \sqrt{\frac{8.426383}{20} \left(\frac{1}{12} + \frac{1}{11} + \frac{3.7879^2}{64.66667 + 102.7273} \right)}$$

i.e. (0.110 ± 0.692)
i.e. $(-0.58, 0.80)$.

Fitting a model in R with only a continuous covariate of gestational age and a factor for RDS group provides the following output:

```
resp.parallel<- lm(Lrate~GA+RDS,data=rds)
summary(resp.parallel)
```

```
##
## Call:
## lm(formula = Lrate ~ GA + RDS, data = rds)
##
## Residuals:
##      Min      1Q   Median       3Q      Max
## -0.89309 -0.40617 -0.03309  0.26879  1.48324
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.27810    1.63292  -2.620  0.01640 *
## GA           0.14455    0.05017   2.881  0.00923 **
## RDSRDS      -0.11010    0.33095  -0.333  0.74284
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6491 on 20 degrees of freedom
## Multiple R-squared:  0.4151, Adjusted R-squared:  0.3566
## F-statistic: 7.097 on 2 and 20 DF,  p-value: 0.004686
```

```
print(anova(resp.parallel))
```

```
## Analysis of Variance Table
##
## Response: Lrate
##           Df Sum Sq Mean Sq F value   Pr(>F)
## GA         1 5.9335  5.9335 14.0830 0.001253 **
## RDS        1 0.0466  0.0466  0.1107 0.742844
## Residuals 20 8.4264  0.4213
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is therefore no evidence, from both the confidence interval (-0.58, 0.8) and p-value for RDS (0.743), of a difference between the regression lines. We conclude that there is insufficient evidence of a difference in the rate of blood flow through pores between babies who suffer from RDS and those who do not and finally conclude that a single line is a good model for the data.

9

```
resp.single <- lm(Lrate~GA,data=rds)
print(anova(resp.separate,resp.parallel,resp.single))
```

```
## Analysis of Variance Table
##
## Model 1: Lrate ~ GA * RDS
## Model 2: Lrate ~ GA + RDS
## Model 3: Lrate ~ GA
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     19 8.3458
## 2     20 8.4264 -1 -0.080630 0.1836 0.6731
## 3     21 8.4730 -1 -0.046627 0.1062 0.7481
```

Weight changes in Herring Gulls

The data in the below table and figure are part of a large sample collected by the Department of Zoology at the University of Glasgow in a study of the weight changes in herring gulls throughout the year. Some birds were caught in June (coded as month 1) and others in December (month 2). Since weight is dependent on the size of the bird this information is recorded in the form of the head and bill length, HAB (in mm), the distance from the back of the head to the tip of the bill.

```
ggplot(data=hab, aes(x=HAB.length,y= Weight, colour = Month, shape= Month))+
  geom_point(size = 3, alpha = .8) +
  geom_smooth(method="lm", fill=NA, fullrange=TRUE) +
    theme(plot.background = element_rect(
        fill = "transparent",
        colour = NA,
        size = 1),    legend.position = c(0.15,0.85),
      legend.background = element_rect(fill="transparent", size=0.5, linetype="solid")) +
  guides(color = guide_legend(label.position = "right") )
```
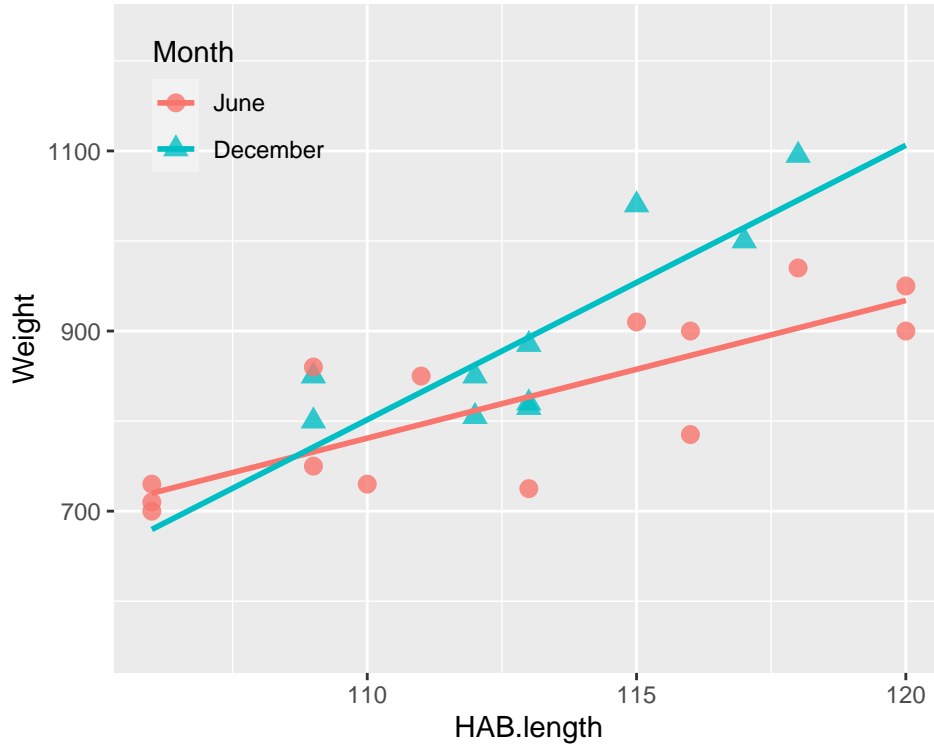
A model with two separate lines appears to be plausible. We will now fit the separate lines model and examine it and conclude whether to keep two separate lines or reduce to parallel lines or even further to a single line. The summary numbers for the data are:

$$\bar{y}_{1.} = 819.2857 \quad \bar{y}_{2.} = 896$$
$$\bar{x}_{1.} = 112.5 \quad \bar{x}_{2.} = 113.1$$
$$S_{y_1 y_1} = 120642.9 \quad S_{y_2 y_2} = 105340.0$$
$$S_{x_1 x_1} = 333.5 \quad S_{x_2 x_2} = 78.9$$
$$S_{x_1 y_1} = 5100 \quad S_{x_2 y_2} = 2404$$
$$n_1 = 14 \quad n_2 = 10$$
$$p = 4$$

We first fit the model with different slopes,

$$E(Y_{ij}) = \alpha_i + \beta_i(x_{ij} - \bar{x}_{i.})$$

We already have the formulae to do this:

$$\hat{\alpha}_1 = \bar{y}_{1.}$$
$$\hat{\alpha}_2 = \bar{y}_{2.}$$
$$\hat{\beta}_1 = \frac{S_{x_1 y_1}}{S_{x_1 x_1}}$$
$$\hat{\beta}_2 = \frac{S_{x_2 y_2}}{S_{x_2 x_2}}$$
$$RSS = S_{y_1 y_1} - \frac{(S_{x_1 y_1})^2}{S_{x_1 x_1}} + S_{y_2 y_2} - \frac{(S_{x_2 y_2})^2}{S_{x_2 x_2}} = RSS_1 + RSS_2$$

11

A 95% C.I. for $(\beta_1 - \beta_2)$ is

$$\hat{\beta}_1 - \hat{\beta}_2 \pm t(n - p; 0.975)\sqrt{\frac{RSS}{n-p}\left(\frac{1}{S_{x_1 x_1}} + \frac{1}{S_{x_2 x_2}}\right)}$$

Using the summary statistics for the "different lines" model we have

$$\hat{\beta}_1 = 15.2924$$

$$\hat{\beta}_2 = 30.4689$$

$$RSS = 42651.896 + 32092.65 = 74744.5$$

A C.I. for the difference in slopes is

$$\hat{\beta}_1 - \hat{\beta}_2 \pm t(n_1 + n_2 - 4; 0.975)\sqrt{\left(\frac{RSS_1 + RSS_2}{n_1 + n_2 - 4}\right)\left(\frac{1}{S_{x_1 x_1}} + \frac{1}{S_{x_2 x_2}}\right)}$$

$$-15.1765 \pm 2.06\sqrt{\frac{74744.5}{20}\left(\frac{1}{333.5} + \frac{1}{78.9}\right)}$$

i.e. (15.17 ± 15.77) or $(-30.9, 0.6)$.

In R, the model above (that allows two completely separate linear regression lines for a response of weight, one for each month) can be achieved by fitting a linear model with a continuous covariate for HAB length, a factor covariate for group and a covariate that allows an interaction between HAB and month. The R code and analysis of variance table are provided below:

```
hab <- read.csv("hab.csv")
hab.separate <- lm(Weight~HAB.length*Month,data=hab)
print(anova(hab.separate))
```

```
## Analysis of Variance Table
##
## Response: Weight
##                  Df Sum Sq Mean Sq F value    Pr(>F)
## HAB.length        1 145746  145746 38.9985 4.248e-06 ***
## Month             1  25126   25126  6.7231   0.01739 *
## HAB.length:Month  1  14696   14696  3.9324   0.06127 .
## Residuals        20  74745    3737
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The term HAB.length:Month is the interaction term. If this has a small p-value (i.e. typically $<$ 0.05) then it suggests that the model with completely separate regression lines is the most appropriate i.e. there is a statistically significant interaction. If this p-value is not < 0.05 then we should consider the parallel lines model.

Since the confidence interval contains zero (although only just!), we cannot reject the parallel lines model. It therefore seems that a separate slope is not required for each month.

The parallel lines model produces

$$\hat{\beta} = 18.1959$$

$$RSS = 89440.6$$

A C.I. for the vertical separation between the regression lines is

$$\hat{\alpha}_1 - \hat{\alpha}_2 + \hat{\beta}(\bar{x}_{2.} - \bar{x}_{1.}) \pm t(n-p, 0.975) \sqrt{\left(\frac{RSS}{n-p}\right)\left(\frac{1}{n_1} + \frac{1}{n_2} + \frac{(\bar{x}_{2.} - \bar{x}_{1.})^2}{S_{x_1 x_1} + S_{x_2 x_2}}\right)}$$

$$-66.0 \pm 2.08 \sqrt{\frac{89440.6}{21}\left(\frac{1}{10} + \frac{1}{14} + \frac{0.6^2}{333.5 + 78.9}\right)}$$

i.e. (-66.0 ± 51.5) or (-117.5, -14.5).

This model can be fitted in R by removing the interaction term from the linear model i.e. we fit a model that only has a continuous covariate of HAB length and a factor covariate of month:

```
hab.parallel <- lm(Weight~HAB.length+Month,data=hab)
summary(hab.parallel)
```

```
##
## Call:
## lm(formula = Weight ~ HAB.length + Month, data = hab)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -103.384  -46.786   -3.306   36.083  109.840
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1293.553    361.884  -3.574  0.00179 **
## HAB.length     18.196      3.214   5.662 1.28e-05 ***
## Month          65.797     27.090   2.429  0.02421 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 65.26 on 21 degrees of freedom
## Multiple R-squared:  0.6564, Adjusted R-squared:  0.6237
## F-statistic: 20.06 on 2 and 21 DF,  p-value: 1.344e-05
```

If the p-value for the factor (which is Month here) is small (typically < 0.05) then we conclude that the parallel lines model is the most appropriate. However, if the p-value is > 0.05 it suggests that there is insufficient evidence of a difference between the two regression lines and one line may be appropriate i.e. the relationship is the same regardless of the group.

Since the confidence interval does not contain zero, the parallel lines model appears to be appropriate. It therefore seems that December weights are systematically higher than June weights i.e. group 2 weights are higher than group 1 weights since the interval is entirely negative. It is highly likely that the December weights are higher by between 14.5 and 117.5 grams.

We could have also considered the model selection using the `anova` function and compare the two separate, two parallel and single line. and as the p-value < 0.05 we do not have evidencew that the interaction term is significant, therefore we drop the interaction term and use the simpler parallel lines model.

```
hab.single <- lm(Weight~HAB.length,data=hab)
print(anova(hab.separate,hab.parallel,hab.single))
```

```
## Analysis of Variance Table
##
## Model 1: Weight ~ HAB.length * Month
## Model 2: Weight ~ HAB.length + Month
## Model 3: Weight ~ HAB.length
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     20  74745
## 2     21  89441 -1    -14696 3.9324 0.06127 .
## 3     22 114566 -1    -25126 6.7231 0.01739 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```