

# Tutorial 8 Solutions

## Multifactor (3-way) ANOVA and Simple Linear Regression

STAT 292: Applied Statistics 2A

### R Functions Used in This Tutorial

In this tutorial we use the `aov` function to fit a **three-way ANOVA** and visualise the effects of two of the factors in an **interaction graph**, produced using the `interaction.plot` function. We also use the `lm` function to fit a **simple linear regression**.

R functions that will be emphasised during this tutorial are:

Function	Description	Package
<code>aov</code>	Fit an analysis of variance model	<code>stats</code>
<code>interaction.plot</code>	Plot the mean (or other summary) of the response for two-way combinations of factors	<code>stats</code>
<code>lm</code>	Fit linear models	<code>stats</code>

The actions of the functions in this tutorial are best illustrated by looking at the resulting output given below (and in the solutions), along with other examples given in Chapters 3 and 4 of the Part 2 Lecture Notes. Also refer back to the earlier tutorials.

Recall that the help file for any function can be produced by typing `?<FUNCTION_NAME>` or `help(FUNCTION_NAME)` (where `FUNCTION_NAME` is the name of the function) at the command line in the R console (e.g., `?interaction.plot`, `?lm`).

### Questions and Solutions

1. Suppose a new strain of influenza occurs, and a medical centre records the time to recovery for each of 59 patients. Also information is gathered on the age of the patient (in three age groups), the severity of the symptoms at first diagnosis (Low or High), and whether or not the patient self-medicated with a certain over-the-counter drug (Med Y or Med N). The data follow.

	Low S, Med Y	Low S, Med N	High S, Med Y	High S, Med N
20-29	6, 3	7, 4, 5, 4	7, 7, 8	5, 6, 7, 4, 6, 8
30-59	8, 5, 6, 9, 5, 6, 6	7, 2, 4, 6, 7, 3	9, 5, 8, 12, 3, 6, 8, 7, 5, 6, 6	6, 8, 9, 6
60+	7, 9, 9, 11, 6	7, 8, 6	10, 11, 12, 6, 13, 7,	9, 8

- a. Name the three factors and their levels.

There are three factors: **Age** with levels 20-29 (coded "Y" in the code below), 30-59 (coded "M") and 60+, 'older' (coded "O"), **Severity** with Low and High levels (coded "L", "H") and **SelfMed** with Yes and No levels for self-medication (coded "Y", "N").

- b. State the model equation, and give null and alternative hypotheses for the test of interaction between age group and severity of symptoms.

The model equation, allowing for main effects and two-way interactions (but **not** three-way interactions), is

$$Y_{ijkm} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + E_{ijkm}$$

where  $i, j$  and  $k$  are the levels of factors  $A$  (Age),  $B$  (Severity) and  $C$  (self-medication) respectively.  $Y_{ijkm}$  is the  $m$ -th observation having  $A$  at level  $i$ ,  $B$  at level  $j$  and  $C$  at level  $k$ . Sum-to-zero constraints are needed, as usual. The error term,  $E_{ijkm}$ , is assumed to be a random variable from a  $N(0, \sigma^2)$  distribution, with errors being independent.

The test of interaction between the first two factors has hypotheses:

$\mathcal{H}_0$  : There is no interaction—all  $(\alpha\beta)_{ij} = 0$ , versus

$\mathcal{H}_1$  : There is interaction—at least one  $(\alpha\beta)_{ij} \neq 0$ .

- c. Some R commands to enter the data and do some analysis are given below along with their output. Discuss the commands and results. This type of data would often be entered from a csv file or a spreadsheet, but data entry into such files also needs care.

FYI, these data (and the corresponding Patient number: 1, 2, ..., 59) are available with the tutorial questions, in the csv file "recovery.csv". If desired, they can be read into a data frame from that file.

This is a  $3 \times 2 \times 2$  factorial design. There are 59 observations. Age is at three levels, Y, M and O (younger, middle-aged and older for 20-29, 30-59 and 60+). Severity is at two levels, H (High) and L (Low). SelfMed is at two levels, Y (yes) and N (no). The factors have been entered into the ANOVA model in that order, followed by all two-way interactions.

The diagnostic graphs confirm constant variance (a level band across the Residuals versus Predicted Values graph) and normality of residuals (a straight line on the Q-Q residual plot).

We use the usual (Type I SS) ANOVA, which is a sequential ANOVA table—each term allows for the preceding terms. Start with the interaction tests—none of the three two-way interactions is significant at the 5% level (all  $p$ -values are a long way above 0.05).

Now consider main effects. This is allowed, as no factor is involved in a significant interaction. There is a significant main effect of Age ( $p = 0.0002$ , well below 0.05 and 0.01). After allowing for Age, there is a significant main effect of Severity ( $p = 0.0071$ ). However, after allowing for Age and Severity, there is no significant main effect of SelfMed ( $p = 0.1477 > 0.05$ ).

*Note: We had a choice of which terms were entered in the model first. Here, Age and Severity were the first and second choice, as they were considered to be important, based on previous experience with other strains of influenza. The ANOVA has verified that those factors are important, and that after allowing for them there is no discernible effect from SelfMed, either main effect or interaction.*

```
## Days to recovery from flu

# Store the flu recovery data in separate variables for
# days to recovery (response), age group (coded "Y", "M", "O"),
# severity of symptoms ("L", "H") and self medicated indicator ("Y", "N").

Recovery <- c(6, 3, 7, 4, 5, 4, 7, 7, 8, 5, 6, 7, 4, 6, 8,
8, 5, 6, 9, 5, 6, 6, 7, 2, 4, 6, 7, 3, 9, 5, 8, 12, 3, 6, 8, 7, 5, 6, 6, 6, 8, 9, 6,
7, 9, 9, 11, 6, 7, 8, 6, 10, 11, 12, 6, 13, 7, 9, 8)
Age <- factor(rep(c("Y", "M", "O"), c(15, 28, 16)), levels = c("Y", "M", "O"))
Severity <- factor(c(rep(c("L", "H"), c(6, 9)), rep(c("L", "H"), c(13, 15)), rep(c("L", "H"), c(8, 8))),
levels = c("L", "H"))
SelfMed <-
factor(c(rep(c("Y", "N"), c(2, 4)), rep(c("Y", "N"), c(3, 6)), rep(c("Y", "N"), c(7, 6)), rep(c("Y", "N"), c(11, 4)), rep(c(
"Y", "N"), c(5, 3)), rep(c("Y", "N"), c(6, 2))),
levels = c("Y", "N"))
head(cbind(Age, Severity, SelfMed))
```

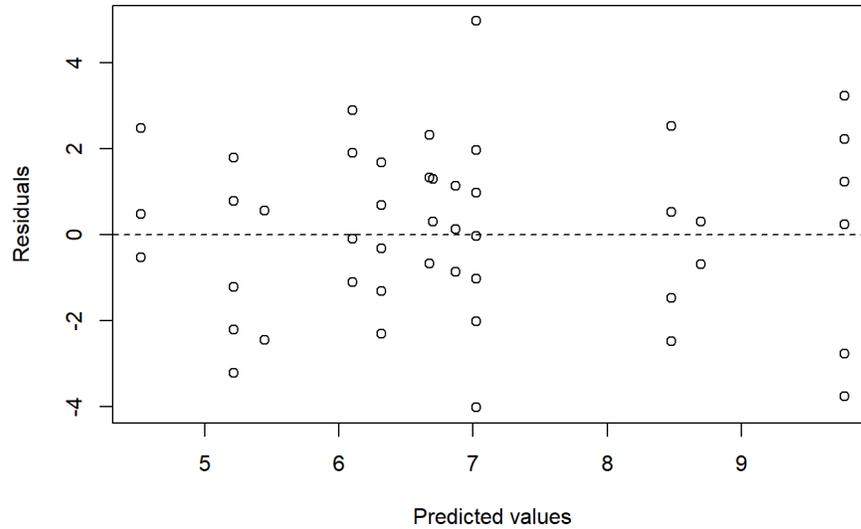
```
      Age Severity SelfMed
[1,]  1         1         1
[2,]  1         1         1
[3,]  1         1         2
[4,]  1         1         2
[5,]  1         1         2
[6,]  1         1         2
```

```
# Fit a three-way ANOVA to the flu recovery data,
# including all main effects and two-way interactions.
recover.ANOVA <- aov(Recovery ~ Age * Severity + Age * SelfMed + Severity * SelfMed)
summary(recover.ANOVA)
```

```
              Df Sum Sq Mean Sq F value    Pr(>F)
Age              2  77.36   38.68  10.150 0.000205 ***
Severity          1  30.11   30.11   7.903 0.007078 **
SelfMed           1   8.25    8.25   2.164 0.147693
Age:Severity      2   0.86    0.43   0.112 0.893859
Age:SelfMed       2   1.40    0.70   0.184 0.832266
Severity:SelfMed  1   0.92    0.92   0.241 0.625898
Residuals       49 186.73    3.81
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

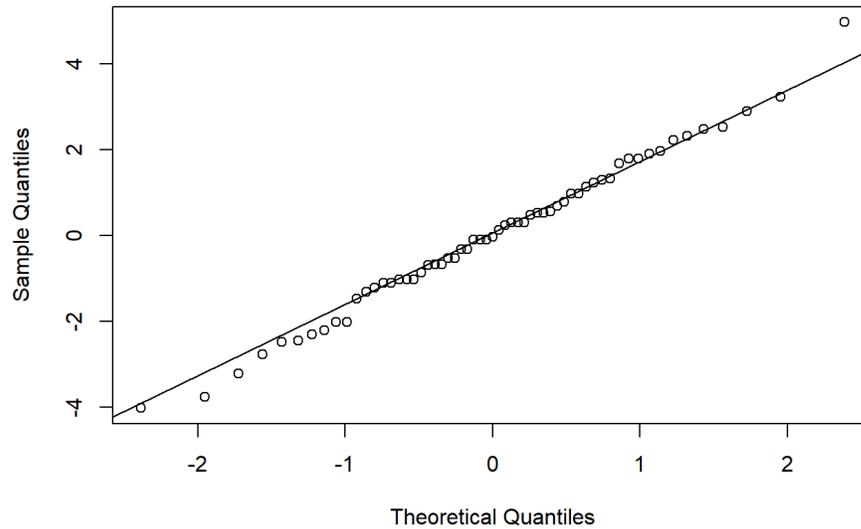
```
# Scatterplot of residuals vs. fitted values for flu recovery data.
plot(x = recover.ANOVA$fitted.values, y = recover.ANOVA$residuals,
     main = "Residuals vs. fitted values\n ANOVA for flu recovery data",
     xlab = "Predicted values", ylab = "Residuals")
abline(h = 0, lty = 2)
```

### Residuals vs. fitted values ANOVA for flu recovery data



```
# Normal Q-Q plot of residuals for flu recovery data.
qqnorm(recover.ANOVA$residuals,
       main = "Normal Q-Q plot of residuals\n ANOVA for flu recovery data")
qqline(recover.ANOVA$residuals)
```

### Normal Q-Q plot of residuals ANOVA for flu recovery data



d. The `R` commands given below produce multiple comparisons of the age groups. May we use these? If so, summarise and interpret the results. If not, why not?

Yes, we may use the multiple comparisons for the three levels of Age, since this factor was not involved in any significant interactions. We may look at the main effects, and do all pairwise comparisons of the three levels. The output shows the 60+ age group is significantly different from each of the other two. So the 60+ age group differs from 30-59 ( $p = 0.0015$ ) and from 20-29 ( $p = 0.0006$ ). However, the 30-59 and 20-29 age groups do not differ significantly in their effect on days to recovery ( $p = 0.6660$ ).

```
# Fit a one-way ANOVA to the flu recovery data,
# including only the Age factor.
recover2.ANOVA <- aov(Recovery ~ Age)
summary(recover2.ANOVA)
```

```

      Df Sum Sq Mean Sq F value    Pr(>F)
Age      2  77.36   38.68    9.489 0.000282 ***
Residuals 56 228.27    4.08
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```
TukeyHSD(recover2.ANOVA)
```

```

Tukey multiple comparisons of means
 95% family-wise confidence level

Fit: aov(formula = Recovery ~ Age)

$Age
      diff      lwr      upr      p adj
M-Y 0.5571429 -0.9981564 2.112442 0.6659994
O-Y 2.8875000  1.1405535 4.634447 0.0005806
O-M 2.3303571  0.8070351 3.853679 0.0014890

```

**e. The `rx` commands below fit a model with only one factor, `SelfMed`. Why does it give a result for the `SelfMed` factor which is different from the one given in the part (c) output?**

The one-way ANOVA below ignores Age and Severity and only tests `SelfMed`. This is not a sensible model, since those factors are non-ignorable. `SelfMed` now appears to be significant, as it is the only factor being given an opportunity to explain the variability in the data. The earlier model, with the age and severity factors included, gave a better explanation of the variability in time to recovery. Note too the higher MSE with the overly simple model. MSE is our estimate of unexplained variability. The one-way ANOVA has increased the unexplained variance by excluding two useful predictors, Age and Severity, from the model.

All ANOVAs are a certain type of regression model, where the explanatory variables are factors having discrete levels. In regression models we can quantify the explained variation in the response variable by the value of  $R^2$ . To get the  $R^2$  values for our ANOVA models, we need to refit them using the `lm` function, then look at the Multiple R-squared value in the penultimate line of the output from the `summary` function. The R code below does that for both models. We see that  $R^2 = 0.389$  for the model with three factors, but  $R^2$  is much lower, only 0.084, for the one-way ANOVA. With the simplified model we have switched from explaining 38.9% of the variance down to 8.4%.

In summary, with the over-simplified analysis, there is an apparent effect of `SelfMed` ( $p = 0.0263$ ), but when we did the full analysis and controlled for Age and Severity, we saw `SelfMed` was not significant, given that Age and Severity were already in the model.

*Note: This kind of confusion can't happen with a balanced design, but here we had observational data together with an imbalance (unequal numbers of observations in the 12 cells). With unbalanced data, the order of entry of terms to the model matters—the effects can change, depending on what has already been allowed for. The researcher must judge the best order of entry of terms to the model. In this example, the older people were more likely to self-medicate, which means that if we put age group first into the model, self-medication becomes less important because some of its explanatory power has already been provided by age group. We really can't tell whether the slower recovery is "caused" by age group or self-medication or a combination of the two, or by some other factor which we didn't measure. In doing the analysis, we decided to allow for age group first, but this was a judgment call. We can say that age group and severity are risk factors, with some power to predict the outcome. We can also say that after controlling for age group and severity, self medication had no significant effect in improving the model.*

```

# Fit a one-way ANOVA to the flu recovery data,
# including only the SelfMed factor.
recover3.ANOVA <- aov(Recovery ~ SelfMed)
summary(recover3.ANOVA)

```

```

      Df Sum Sq Mean Sq F value    Pr(>F)
SelfMed  1  25.55   25.552     5.2 0.0263 *
Residuals 57 280.08    4.914
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

# Fit the full ANOVA using the lm function,
# to see R-squared in the penultimate line of the summary.
recover.lm <- lm(Recovery ~ Age * Severity + Age * SelfMed + Severity * SelfMed)
summary(recover.lm)

```

```

Call:
lm(formula = Recovery ~ Age * Severity + Age * SelfMed + Severity *
    SelfMed)

Residuals:
    Min       1Q   Median       3Q      Max
-4.0263 -1.0639 -0.0263  1.1815  4.9737

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.44847    1.15565   4.715 2.04e-05 ***
AgeM            0.65302    1.26773   0.515  0.6088
AgeO            3.03006    1.35786   2.231  0.0303 *
SeverityH       1.25255    1.26202   0.992  0.3258
SelfMedN       -0.92270    1.25534  -0.735  0.4658
AgeM:SeverityH -0.32772    1.32010  -0.248  0.8050
AgeO:SeverityH  0.03681    1.47837   0.025  0.9802
AgeM:SelfMedN   0.03614    1.33415   0.027  0.9785
AgeO:SelfMedN  -0.68671    1.51704  -0.453  0.6528
SeverityH:SelfMedN 0.53784    1.09628   0.491  0.6259
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.952 on 49 degrees of freedom
Multiple R-squared:  0.389, Adjusted R-squared:  0.2768
F-statistic: 3.467 on 9 and 49 DF,  p-value: 0.00223

```

```

# Also re-fit the one-way ANOVA with lm, to see R-squared.
recover3.lm <- lm(Recovery ~ SelfMed)
summary(recover3.lm)

```

```

Call:
lm(formula = Recovery ~ SelfMed)

Residuals:
    Min       1Q   Median       3Q      Max
-4.412 -1.412 -0.080  1.588  5.588

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.4118    0.3802  19.50 <2e-16 ***
SelfMedN       -1.3318    0.5840  -2.28  0.0263 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.217 on 57 degrees of freedom
Multiple R-squared:  0.0836, Adjusted R-squared:  0.06753
F-statistic:  5.2 on 1 and 57 DF,  p-value: 0.02634

```

f. The `R` commands given below produce an interaction graph for Age by Severity. Note that Age was coded “Y” for ages 20-29, “M” for ages 30-59 and “O” for ‘older’ ages, 60+, when the data was entered above. Why is it valid to produce this interaction graph? What is the plot illustrating?

This interaction graph is valid, as the third factor, SelfMed, had no significant interactions or main effects, so it is fine to average over its levels of Yes/No and only consider the other two factors. Age has been entered as the factor on the  $x$  axis, since it has ordered levels.

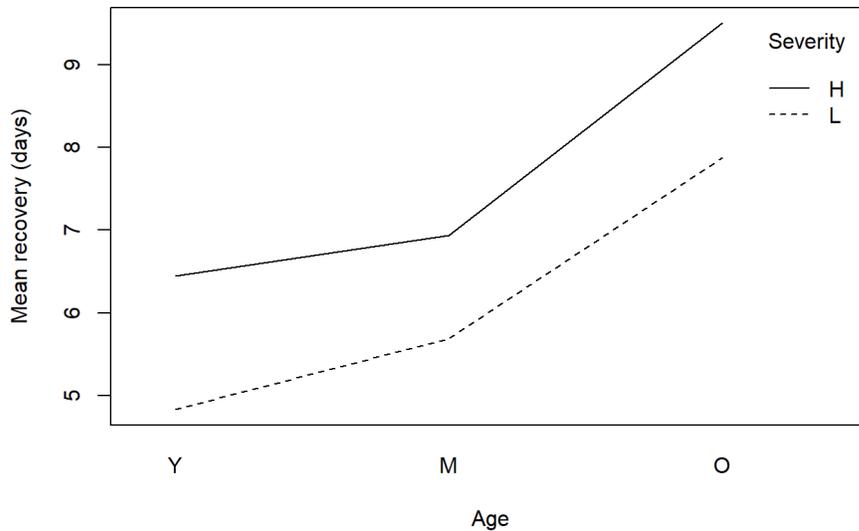
Interpreting the interaction plot, the non-significant interaction between Age and Severity appears as (almost) parallel traces. The main effect of Age is shown by at least some lines being non-horizontal, with essentially horizontal lines between Y and M (similar recovery times for Young and Middle-aged), but an upward slope between M and O (older people taking longer to recover), as seen in the Tukey multiple comparisons in part (d). The significant main effect of Severity is seen in the vertical separation of the lines, with greater severity associated with longer recovery times, which is certainly expected.

```

# Interaction graph for flu recovery data.
interaction.plot(x.factor = Age,
               trace.factor = Severity,
               response = Recovery,
               fun = mean,
               ylab = "Mean recovery (days)",
               main = "Interaction graph for flu recovery",
               legend = TRUE, xpd=TRUE)

```

Interaction graph for flu recovery



2. This dataset gives characteristics of water samples taken at  $n = 53$  Florida lakes (Lange, Royals and Connor, 2004, "Mercury accumulation in largemouth bass (*Micropterus salmoides*) in a Florida Lake", *Archives of Environmental Contamination and Toxicology*, 27(4): 466-471). Two of the variables are Mercury = average mercury levels in the fish, and water pH (acidity level,  $\text{pH} < 7$  for acid,  $\text{pH} = 7$  for neutral,  $\text{pH} > 7$  for alkaline). This type of data would often be entered from a csv file or a spreadsheet.

FYI, these data (including the Lake number: 1, 2, ..., 53) are available with the tutorial questions, in the csv file "floridaBass.csv". If desired, they can be read into a data frame from that file.

Lake	Mercury	pH	Lake	Mercury	pH	Lake	Mercury	pH
1	1.23	6.1	19	1.08	5.8	37	0.19	6.8
2	1.33	5.1	20	0.98	6.7	38	0.04	8.4
3	0.04	9.1	21	0.63	4.4	39	0.49	7.0
4	0.44	6.9	22	0.56	6.7	40	1.10	7.5
5	1.20	4.6	23	0.41	6.1	41	0.16	7.0
6	0.27	7.3	24	0.73	6.9	42	0.10	6.8
7	0.48	5.4	25	0.34	5.5	43	0.48	5.9
8	0.19	8.1	26	0.59	6.9	44	0.21	8.3
9	0.83	5.8	27	0.34	7.3	45	0.86	6.7
10	0.81	6.4	28	0.84	4.5	46	0.52	6.2
11	0.71	5.4	29	0.50	4.8	47	0.65	6.2
12	0.50	7.2	30	0.34	5.8	48	0.27	8.9
13	0.49	7.2	31	0.28	7.8	49	0.94	4.3
14	1.16	5.8	32	0.34	7.4	50	0.40	7.0
15	0.05	7.6	33	0.87	3.6	51	0.43	6.9
16	0.15	8.2	34	0.56	4.4	52	0.25	5.2
17	0.19	8.7	35	0.17	7.9	53	0.27	7.9
18	0.77	7.8	36	0.18	7.1			

Some R output from a simple linear regression with  $x = \text{pH}$  and  $Y = \text{average mercury level}$  follow.

```
## Mercury accumulation data

# Store the largemouth bass data in separate variables for
# mercury accumulation and pH level.
# Could be read in from, e.g., a csv file instead.

mercury <- c(1.23, 1.33, 0.04, 0.44, 1.2, 0.27, 0.48, 0.19, 0.83,
            0.81, 0.71, 0.5, 0.49, 1.16, 0.05, 0.15, 0.19, 0.77,
            1.08, 0.98, 0.63, 0.56, 0.41, 0.73, 0.34, 0.59, 0.34,
            0.84, 0.5, 0.34, 0.28, 0.34, 0.87, 0.56, 0.17, 0.18,
            0.19, 0.04, 0.49, 1.1, 0.16, 0.1, 0.48, 0.21, 0.86,
            0.52, 0.65, 0.27, 0.94, 0.4, 0.43, 0.25, 0.27)
pH <- c(6.1, 5.1, 9.1, 6.9, 4.6, 7.3, 5.4, 8.1, 5.8,
       6.4, 5.4, 7.2, 7.2, 5.8, 7.6, 8.2, 8.7, 7.8,
       5.8, 6.7, 4.4, 6.7, 6.1, 6.9, 5.5, 6.9, 7.3,
       4.5, 4.8, 5.8, 7.8, 7.4, 3.6, 4.4, 7.9, 7.1,
       6.8, 8.4, 7, 7.5, 7, 6.8, 5.9, 8.3, 6.7,
       6.2, 6.2, 8.9, 4.3, 7, 6.9, 5.2, 7.9)
head(cbind(mercury,pH))
```

```
      mercury pH
[1,]    1.23 6.1
[2,]    1.33 5.1
[3,]    0.04 9.1
[4,]    0.44 6.9
[5,]    1.20 4.6
[6,]    0.27 7.3
```

```
# Fit a linear regression of mercury level on lake pH
# for the Florida lakes largemouth bass data.
bass.lm <- lm(mercury ~ pH)
# ANOVA table output for the linear regression.
anova(bass.lm)
```

```
Analysis of Variance Table

Response: mercury
      Df Sum Sq Mean Sq F value    Pr(>F)
pH      1  2.0024  2.00236   25.243 6.573e-06 ***
Residuals 51  4.0455  0.07932
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Summary output for the linear regression.
summary(bass.lm)
```

```
Call:
lm(formula = mercury ~ pH)

Residuals:
    Min       1Q   Median       3Q      Max
-0.48895 -0.19188 -0.05774  0.09456  0.71134

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.53092     0.20349   7.523 8.14e-10 ***
pH          -0.15230     0.03031  -5.024 6.57e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2816 on 51 degrees of freedom
Multiple R-squared:  0.3311,    Adjusted R-squared:  0.318
F-statistic: 25.24 on 1 and 51 DF,  p-value: 6.573e-06
```

```

# Produce a scatterplot of mercury level vs. water pH
# for the largemouth bass data.
plot(x = pH, y = mercury,
main = "Scatterplot of mercury in largemouth bass vs. water pH\n with fitted regression line",
xlab = "Water pH", ylab = "Average mercury")
# Overlay the line of best fit from the linear regression.
abline(bass.lm)

# Produce a scatterplot of studentized residuals versus fitted values
# for the largemouth bass data.
plot(x = bass.lm$fitted.values, y = rstudent(bass.lm),
main = "Studentized residuals vs. fitted values\n regression for largemouth bass data",
xlab = "Predicted value", ylab = "Studentized residual")
abline(h = 0)
abline(h = c(-2, 2), lty = 2)

# Produce a normal Q-Q plot of residuals
# for the largemouth bass data - doesn't look great.
qqnorm(bass.lm$residuals,
main = "Normal Q-Q plot of residuals\n regression for largemouth bass data")
qqline(bass.lm$residuals)

# Try the normal Q-Q plot again, for a model fitted to
# logged mercury levels regressed on lake pH - looks better
bass2.lm <- lm(log(mercury) ~ pH)
# Summary output for the linear regression using logged response.
summary(bass2.lm)

qqnorm(bass2.lm$residuals, main = "Normal Q-Q plot of residuals\n regression using logged
response")
qqline(bass2.lm$residuals)

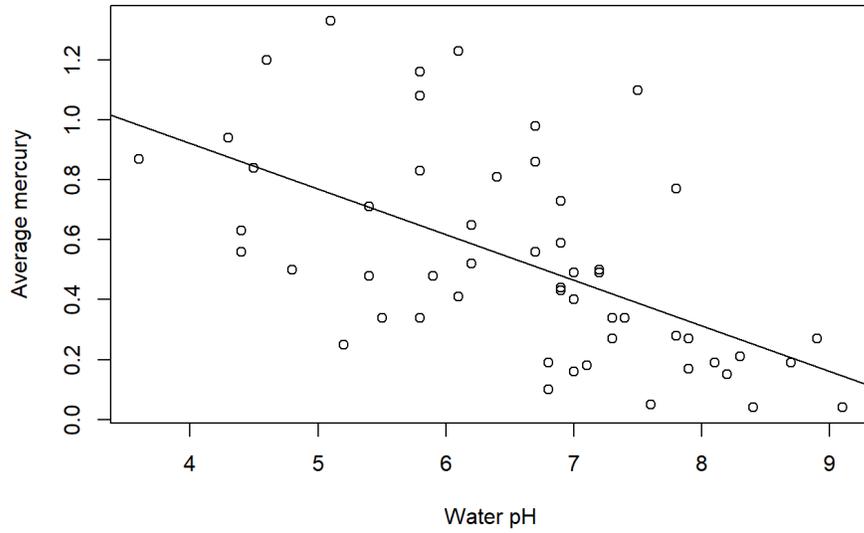
# For completeness, here's a scatterplot of logged mercury level vs. water pH
# for the largemouth bass data.
plot(x = pH, y = log(mercury),
main = "Scatterplot of logged mercury in largemouth bass vs. water pH\n with fitted regression
line",
xlab = "Water pH", ylab = "Average logged mercury")
# Overlay the line of best fit from the linear regression.
abline(bass2.lm)

# Produce a scatterplot of studentized residuals versus fitted values
# for the logged response.
plot(x = bass2.lm$fitted.values, y = rstudent(bass2.lm),
main = "Studentized residuals vs. fitted values\n regression for logged response",
xlab = "Predicted value", ylab = "Studentized residual")
abline(h = 0)
abline(h = c(-2, 2), lty = 2)

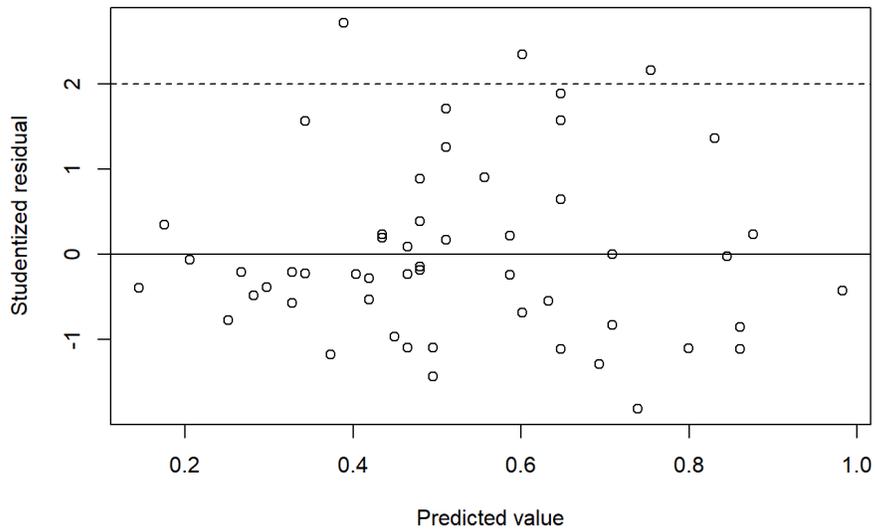
# Load the "olsrr" add-on package.
library(olsrr)
# Plot Cook's distances for the largemouth bass data.
ols_plot_cooksd_chart(bass.lm)
# And now plot Cook's distances for the logged response model.
ols_plot_cooksd_chart(bass2.lm)

```

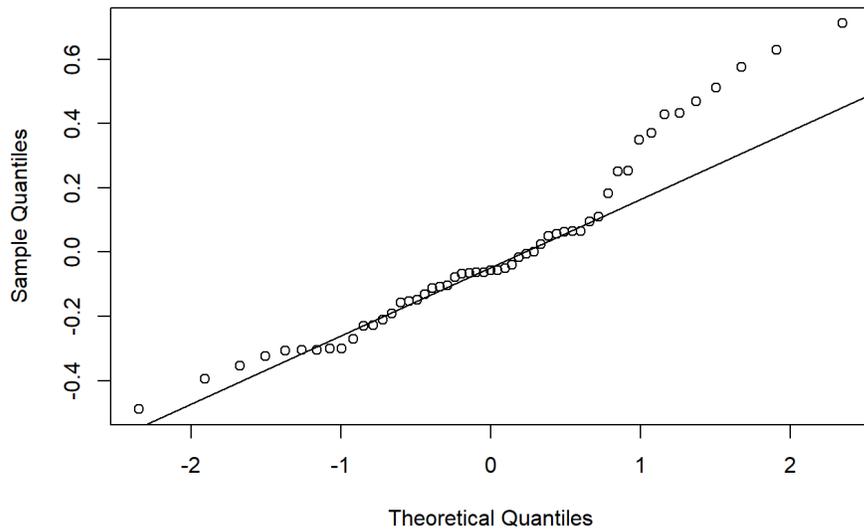
**Scatterplot of mercury in largemouth bass vs. water pH with fitted regression line**



**Studentized residuals vs. fitted values regression for largemouth bass data**



**Normal Q-Q plot of residuals regression for largemouth bass data**



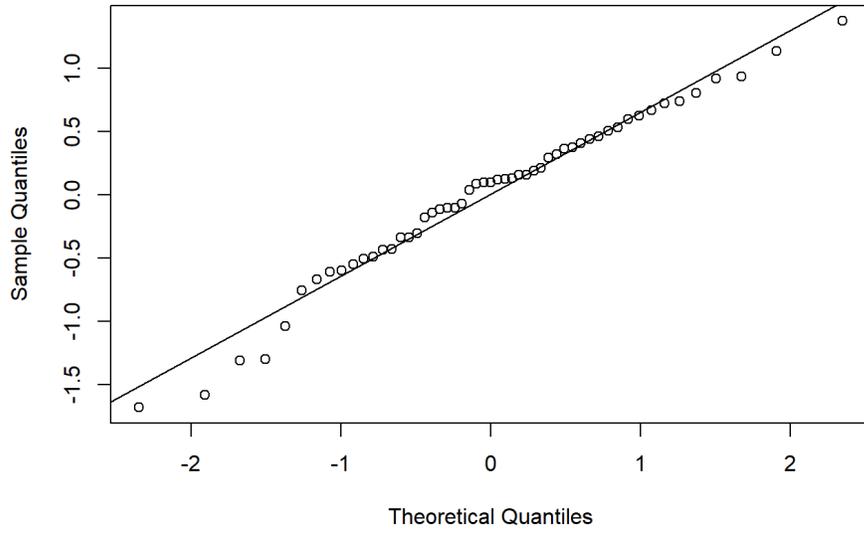
```
Call:
lm(formula = log(mercury) ~ pH)

Residuals:
    Min       1Q   Median       3Q      Max
-1.67936 -0.43150  0.09943  0.44216  1.37147

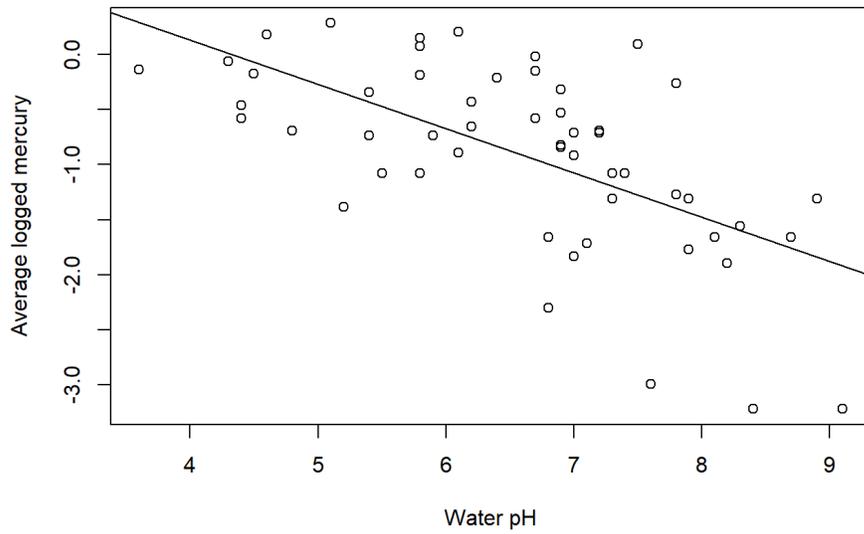
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.73999    0.48187   3.611 0.000696 ***
pH           -0.40215    0.07178  -5.602 8.54e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6669 on 51 degrees of freedom
Multiple R-squared:  0.381, Adjusted R-squared:  0.3688
F-statistic: 31.39 on 1 and 51 DF,  p-value: 8.544e-07
```

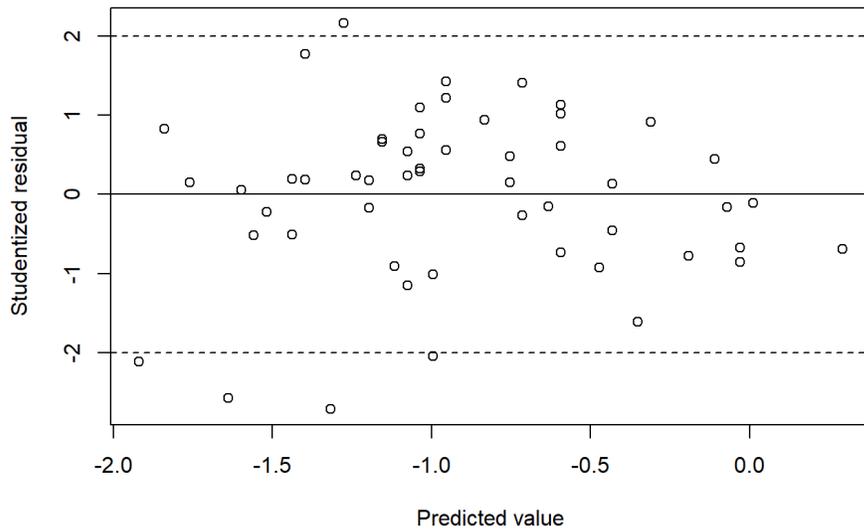
**Normal Q-Q plot of residuals  
regression using logged response**

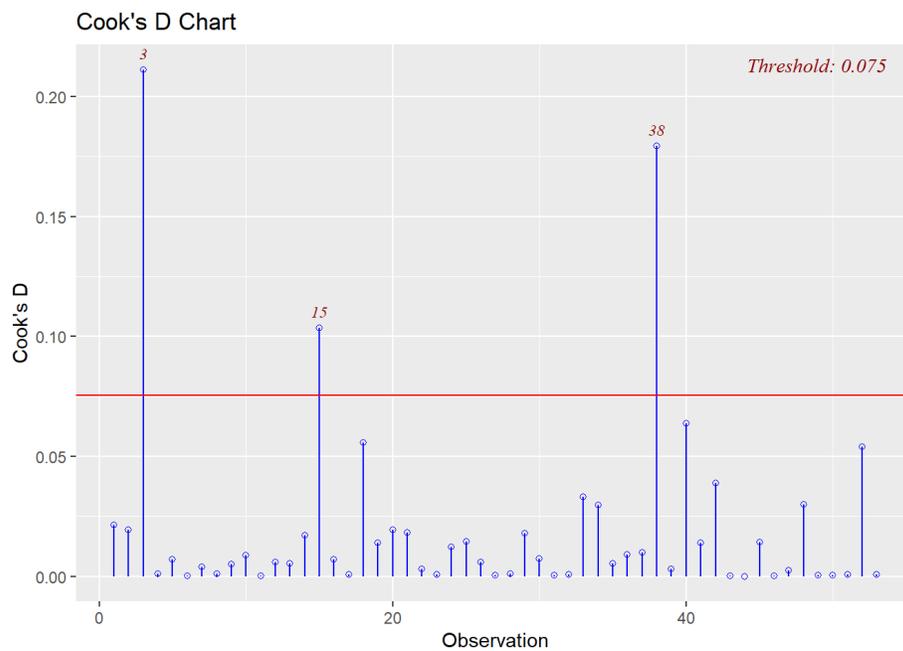
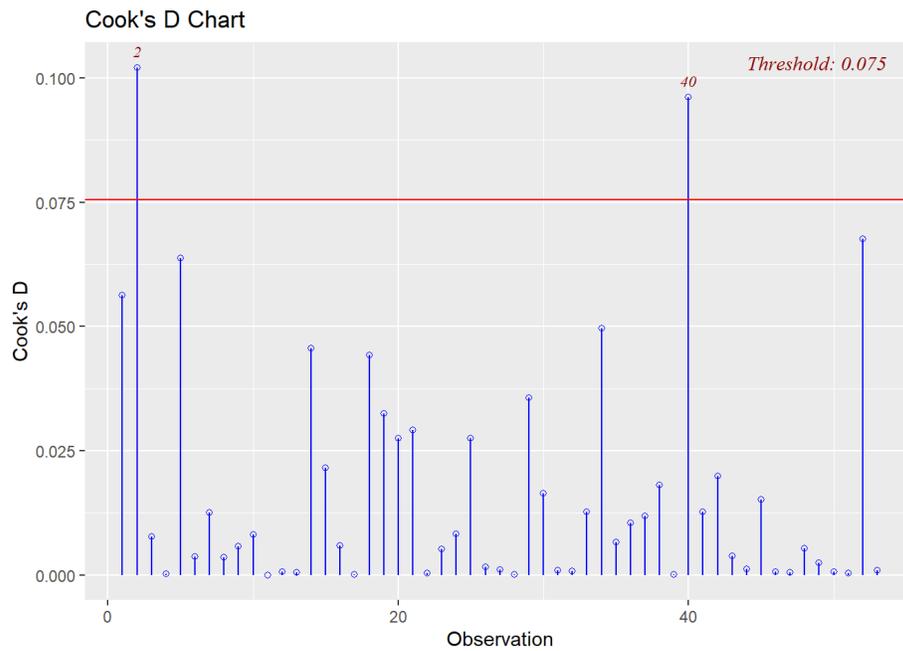


**Scatterplot of logged mercury in largemouth bass vs. water pH  
with fitted regression line**



**Studentized residuals vs. fitted values  
regression for logged response**





a. Check that you understand how all the output was obtained.

Look through the code and the corresponding output. Code can be copied and pasted into R to reproduce the output.

b. What is the theoretical model equation? What is the fitted model equation? Use the model fitted to the original data, rather than the logged data— but discuss both models briefly in part (c).

The theoretical model is

$$Y = \beta_0 + \beta_1 x + E$$

where  $Y$  = Mercury and  $x$  = pH.

The fitted model is the line

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 x \quad \text{or}$$

$$\widehat{Y} = 1.53092 - 0.15230x$$

using the estimates from the R output.

c. What are the assumptions? Are they satisfied?

The assumptions are that the errors come independently from a  $N(0, \sigma^2)$  distribution, and that the errors are also independent of  $x$ .

The constant variance assumption looks valid, as there is a level band in the plot of Studentized residuals versus Predicted Values. Note, however, that the data is clearly not symmetrically distributed around the fitted line—you can see that in the scatterplot, as well as in the residual plot: there are some quite big positive residuals and more, smaller, negative ones. That suggests we might take logs of the response variable, as does the shape of residual Q-Q plot, which is not a very straight line—so there is doubt about the normality assumption. That log transformation has been done as part of the given R code and output.

All the Cook's distances are well below 1 with the untransformed data, so no outliers or points of high leverage have been detected.

Taking logs fixes the normality 'problem'—but it makes things somewhat worse in other ways. With the logged response variable there are more studentized residuals outside the  $\pm 2$  'guidelines', and the biggest Cook's distances are larger—although still none are close to unity. Also the Studentized residuals versus Predicted Values plot now shows some inverse funnelling—there is greater variation in the studentized residuals at low pH values. Since the underlying theory in regression and ANOVA is identical, we generally prefer data with residuals that show a more-constant variance, over residuals that are better approximated by a normal distribution. (Ideally we want both, so that all our model assumptions hold, but a compromise is often necessary!) So we continue to use the regression model fitted to the untransformed mercury accumulation data.

- d. Find a point prediction for the mercury level of a largemouth bass fish in neutral water ( $pH = 7$ ). (Again use the model fitted to the original data here.)

For neutral water, the predicted average mercury level is  $\hat{Y} = 1.53092 - 0.15230 \times 7 = 0.465$ .

- e. State the null and alternative hypotheses for the test of whether pH is a useful predictor.

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0.$$

- f. What is the test statistic? Give the statistical conclusion and your interpretation (using the model fitted to the original data).

There is a choice of test.

The  $F$  statistic is 25.24 on (1, 51) df, while the  $t$  statistic is  $-5.02$  on 51 df. The df for  $t$  are determined by the estimate of  $\sigma^2$ , from the MSE in the ANOVA Table, which has 51 df. The  $F$  statistic is equal to the  $t$  statistic squared, since there is only one numerator df; any minor difference is due to rounding. The  $p$ -values are the same from either statistic,  $p = 6.57 \times 10^{-6}$ , so we reject  $H_0$  at any reasonable significance level. Water pH is a useful predictor for the average mercury level in the fish.

- g. Interpret the 95% confidence interval for the slope that is given below.

The 95% confidence interval for the slope is  $(-0.21, -0.09)$ . The negative sign shows decreased mercury with increased alkalinity. The 95% confidence interval does not include zero, showing the slope is non-zero at the 5% significance level.

```
# Obtain 95% confidence intervals for beta_0 and beta_1.
confint.default(bass.lm)
```

```
(Intercept)  2.5 %      97.5 %
              1.1320800  1.92975741
pH           -0.2117138 -0.09288796
```