Solutions to the Exercises

Estimates of parameters guessed at in the lab:

- Soil pH: $\hat{\mu} = 6.68$,$\hat{\sigma} = 0.45$
- Fuel economy: $\hat{\beta} = 0.0080$, $\hat{\alpha} = -0.82$
- Potato yield: $\hat{\mu}_1 = 746.8$, $\hat{\mu}_2 = 625.8$, $\hat{\mu}_3 = 656.0$, $\hat{\mu}_4 = 611.0$, $\hat{\sigma} = 33.5$

1.
```
> salinity <- c(9.3, 10.7, 5.5, 9.6, 12.2, 16.6, 9.2, 10.5, 7.9,
+     13.2, 11, 8.8, 13.7, 12.1, 9.8)

> hist(salinity)
```

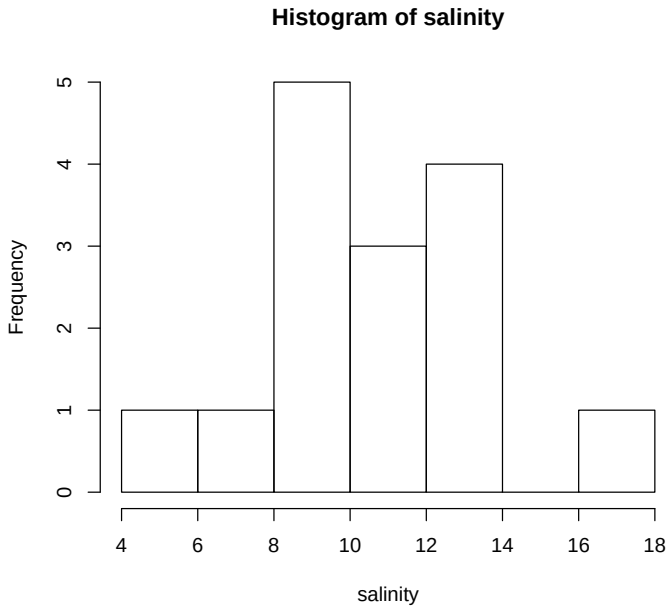

Figure 1: Plot of salinity

A possible model for this is given by:

$$y_i = \mu + e_i,$$

where the y_i are the salinity values.

Parameters to be estimated: μ; and σ, the standard deviation of the error distribution.

```
> mean(salinity)

[1] 10.67333

> sd(salinity)

[1] 2.660684
```

Therefore $\hat{\mu} = 10.7$ and $\hat{\sigma} = 2.7$.

2.
```
> pollution <- data.frame(location = rep(c("Above", "Below"), each = 15),
+     d.o. = c(5.2, 4.8, 5.1, 5, 4.9, 4.8, 5, 4.7, 4.7, 5, 4.7,
+         5.1, 5, 4.9, 4.9, 4.2, 4.4, 4.7, 4.9, 4.6, 4.8, 4.9,
+         4.6, 5.1, 4.3, 5.5, 4.7, 4.9, 4.8, 4.9))

> plot(pollution)
```

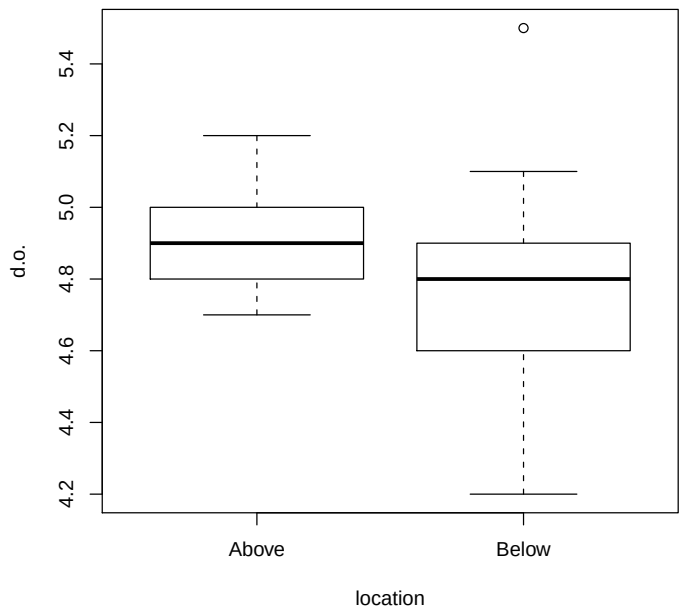

Figure 2: Plot of pollution

A possible model for this is given by:

$$\text{oxygen}_{ij} = \text{location}_i + e_{ij},$$

OR

$$y_{ij} = \mu_i + e_{ij},$$

for $i = 1, 2$ and $j = 1, \ldots, 15$.

Parameters to be estimated: μ_1, μ_2, σ.

Assumptions: the errors e_{ij} are separate draws from a normal distribution with mean 0 and standard deviation σ, i.e. $e_{ij} \sim N(0, \sigma)$.

Estimate μ_1:

```
> mean(pollution[pollution$location == "Above", ]$d.o.)
```

```
[1] 4.92
```

Estimate μ_2:

```
> mean(pollution[pollution$location == "Below", ]$d.o.)
```

```
[1] 4.753333
```

Note the use of square brackets to select part of the data frame. The condition listed before the comma determines which rows to select: in the first case here, those for which `location` is "Above", and in the second case, those for which `location` is "Below". The condition listed after the comma determines which columns to select: in both cases here, no columns are to be excluded, so there is no condition – it is left blank.

Note also the use of the double equals == in the above statements, rather than the single = which R uses for arguments in functions.

Estimate σ:

```
> sd1 <- sd(pollution[pollution$location == "Above", ]$d.o.)
> sd2 <- sd(pollution[pollution$location == "Below", ]$d.o.)
> wsd <- sqrt((sd1^2 + sd2^2)/2)
> wsd
```

```
[1] 0.2536402
```

Using the average of the variances gives an appropriate estimate of σ because it represents the within groups variation, after adjusting for location.

3. A possible model is:

$$\text{attacks}_{ij} = \text{aspirin}_i + e_{ij},$$

where $i = 1, 2$ for aspirin or placebo and $j = 1, \ldots, 11000$.

We are not told what values the heart attack response variable could take.

4. A possible model is:

$$\text{weight}_{ijk} = \text{corn}_i + \text{protein}_j + e_{ijk},$$

where $i = 1, 2, 3$ for corn varieties, $j = 1, 2, 3$ for protein levels and $k = 1, \ldots, 10$.

Note that a more complex model would consider if there was any interaction between corn and protein in their effect on weight, i.e. if the effect of protein was different for different corn types.

5. A possible model is:

$$\text{worms}_{ij} = \text{treatment}_i + e_{ij},$$

where $i = 1, 2$ for treated and untreated and $j = 1, \ldots, 7$.

6.
```
> salinity <- c(7.6, 7.7, 4.3, 5.9, 5, 6.5, 8.3, 8.2, 13.2, 12.6,
+     10.4, 10.8, 13.1, 12.3, 10.4)
> afterflow <- c(23, 24, 26, 25, 30, 24, 23, 22, 22, 24, 25, 22,
+     22, 22, 24)
```

A possible model is:

$$y_i = \alpha + \beta x_i + e_i,$$

where y refers to the salinity, x refers to the afterflow and $i = 1, \ldots, 15$. We can fit this using the R linear model function.

```
> model1 <- lm(salinity ~ afterflow)
> summary(model1)

Call: lm(formula = salinity ~ afterflow)

Residuals:
     Min       1Q   Median       3Q      Max
-2.84435 -2.21527  0.01381  1.91381  3.63473

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  30.8163     7.0248   4.387 0.000735 ***
afterflow    -0.9105     0.2932  -3.105 0.008369 **
---
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Ś Š 1

Residual standard error: 2.341 on 13 degrees of freedom
Multiple R-squared: 0.4258, Adjusted R-squared: 0.3816
F-statistic:  9.64 on 1 and 13 DF,  p-value: 0.008369
```

```
> plot(salinity ~ afterflow)
> curve(30.8 - 0.91 * x, add = TRUE)
```

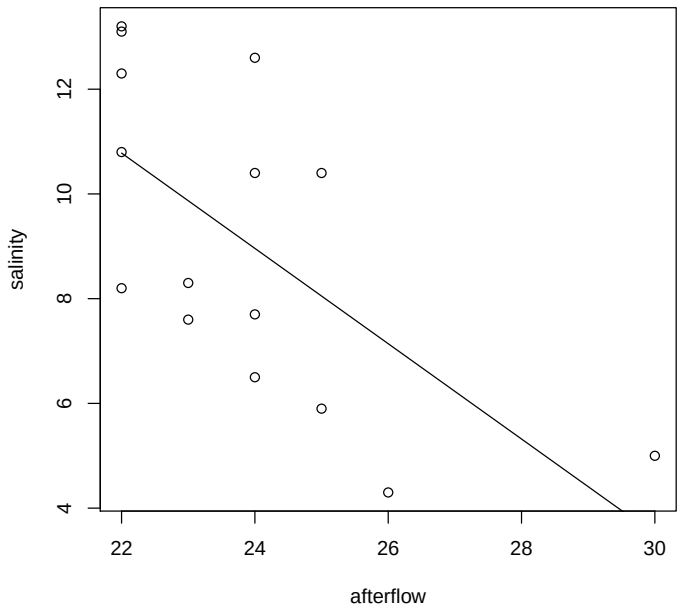

Figure 3: Plot of salinity against waterflow with linear model fit

Appropriate statistical models for the exercises in Lab 1:

- $\text{bp}_{ij} = \text{user}_i + e_{ij}$, for $i = 1, 2$
- $\text{pulse2}_{ijk} = \text{pulse1}_i + \text{ran}_j + e_{ijk}$, for $j = 1, 2$
- $\text{score}_i = \mu + e_i$
- $\text{evaporation}_i = \alpha + \beta\ \text{air.speed}_i + e_i$
- $\text{yield}_{ij} = \text{variety}_i + e_{ij}$, for $i = 1, 2, 3, 4$
- $\text{height}_{ijk} = \mu + \beta\ \text{dbh}_i + \text{species}_j + e_{ijk}$, for $i = 1, 2, 3, 4$
- $\text{count}_{ij} = \text{litter}_i + \text{drug}_j + e_{ij}$, for $i = 1, 2, 3, 4, 5$ and $j = 1, 2, 3, 4$