Solutions for 10.5 Exercises

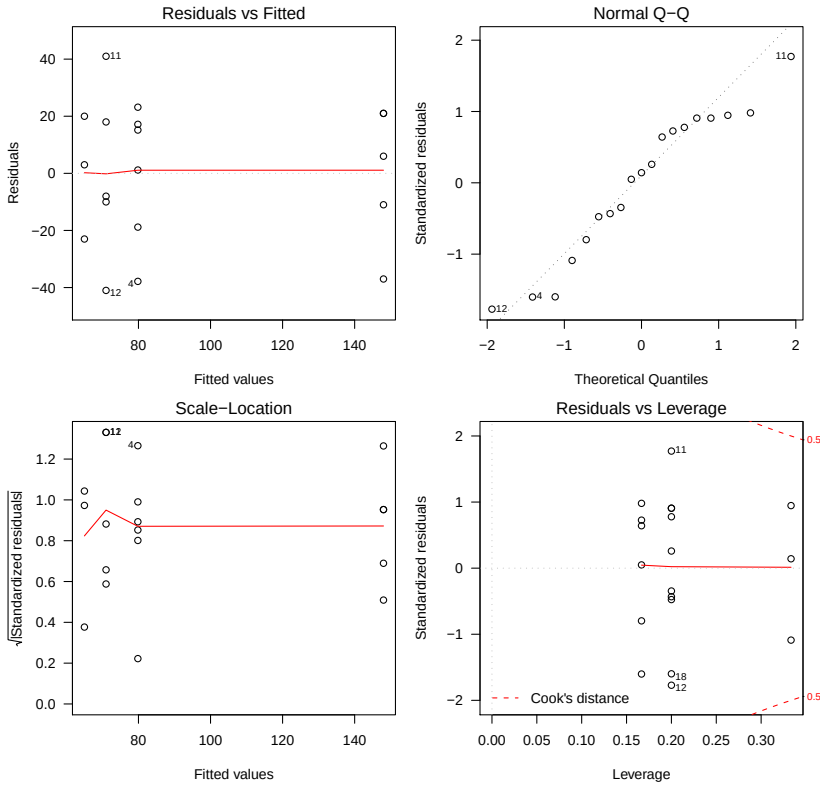
1. We use a one-way ANOVA model:

$$y_{ij} = \mu_i + e_{ij}; \qquad e_{ij} \sim \mathrm{N}(0, \sigma), \quad i = \text{A, B, C, D (or 1,2,3,4). j is different for each } i.$$

```
> chicks <- data.frame(feed = rep(c("A","B","C","D"),
+                           times=c(3,6,5,5)),
+                      weight_gain = c(
+                           42, 68, 85,
+                           42, 97, 81, 95, 61, 103,
+                           61, 112, 30, 89, 63,
+                           169, 137, 169, 111, 154))
> chicks.lm <- lm(weight_gain~feed, data=chicks)
```

To obtain linear model diagnostics, we use:

```
> par(mfrow = c(2, 2), las = 1, mar = c(4, 4, 2, 1))
> plot(chicks.lm)
```



the `las` argument of the `par` function determines the style of axis labels; 1="always horizontal". The `mar` argument determines the margins, with the four numbers corresponding to (bottom, left, top, right). Type `?par` for more details.

The diagnostic plots look acceptable. The variance in the four groups is similar, and the residuals appear to be consistent with a normal distribution.

```
> summary(chicks.lm)

Call:
lm(formula = weight_gain ~ feed, data = chicks)

Residuals:
   Min     1Q Median     3Q    Max
-41.00 -14.92   3.00  19.00  41.00

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    65.00      14.94   4.351 0.000571 ***
feedB          14.83      18.30   0.811 0.430247
feedC           6.00      18.90   0.317 0.755251
feedD          83.00      18.90   4.392 0.000525 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 25.88 on 15 degrees of freedom
Multiple R-squared: 0.6758,        Adjusted R-squared: 0.6109
F-statistic: 10.42 on 3 and 15 DF,  p-value: 0.0005872

> anova(chicks.lm)

Analysis of Variance Table

Response: weight_gain
          Df Sum Sq Mean Sq F value    Pr(>F)
feed       3  20937  6979.0  10.422 0.0005872 ***
Residuals 15  10045   669.7
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The small P-value indicates substantial evidence against the null hypothesis of equal means: $\mu_A = \mu_B = \mu_C = \mu_D$. The table for coefficients shows that β_3 is significantly different from β_0 but β_1 and β_2 are not. This means that feed B and feed C are not statistically significantly different from feed A, but feed D is. (Recall the parameterisation that R uses: feed A is the intercept, and the other three parameters are the difference between feed A and feeds B, C, and D respectively.) So to promote weight gain we would employ feed D.

2. (a)
```
> pine <- data.frame(condition = factor(c(1, 1, 1, 1, 1, 2, 2,
+     2, 3, 3, 3)), moisture = c(7.3, 8.3, 7.6, 8.6, 8.3, 5.4,
+     7.4, 7.1, 8.5, 9.5, 10))
> pine.lm <- lm(moisture ~ condition, data = pine)
> anova(pine.lm)
Analysis of Variance Table

Response: moisture
          Df  Sum Sq Mean Sq F value   Pr(>F)
```

```
condition  2 10.9387  5.4693  9.3466 0.008068 **
Residuals  8  4.6813  0.5852
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

$P = 0.008$, so the null hypothesis of equal means is clearly rejected.

(b) `> qt(0.975, 8)`

```
[1] 2.306004
```

$8.02 - 6.63 \pm t_8^{0.975} \times \sqrt{0.5852\left(\frac{1}{5} + \frac{1}{3}\right)} = 1.39 \pm 2.306 \times 0.559 = 1.39 \pm 1.29 = (0.10, 2.68)$.

3. (a)
```
> cloud <- read.csv("cloud seeding.csv")
> tapply(cloud$rain, cloud$seeding, mean)
```

```
 control   seeded
164.6000 441.9923
```

```
> tapply(cloud$rain, cloud$seeding, median)
```

```
control  seeded
   44.2   221.6
```

```
> tapply(cloud$rain, cloud$seeding, sd)
```

```
 control   seeded
278.4263 650.7832
```

It appears that the distribution of the rainfall from seeded clouds has a larger mean and larger variability than the rainfall from control clouds. In both groups, the sample mean is much larger than the sample median, and the standard deviation is larger than the mean, suggesting substantial skewness in both populations.

(b)
```
> cloud.lm <- lm(rain ~ seeding, data = cloud)
> anova(cloud.lm)
```

```
Analysis of Variance Table

Response: rain
          Df   Sum Sq Mean Sq F value  Pr(>F)
seeding    1  1000304 1000304  3.9929 0.05115 .
Residuals 50 12526000  250520
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The formal comparison of means ($P = 0.051$) is not significant at the 0.05 level, but it would be bad science to just ignore the difference. In a report, it would be appropriate to use words such as "some evidence of a difference".

(c)
```
> t.test(rain ~ seeding, data = cloud, var = TRUE)
```

```
    Two Sample t-test
```

```
data:  rain by seeding
t = -1.9982, df = 50, p-value = 0.05115
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -556.218867    1.434252
sample estimates:
mean in group control  mean in group seeded
             164.6000              441.9923
```

The equal P-values show that this analysis is equivalent to the ANOVA.

ANOVA assumes equal variances for the two groups, so the equivalent t-test needs this assumption incorporated. Note that the assumption is questionable here, and so a t-test with unequal variances would actually be more appropriate in this situation.

(d)
```
> cloud$lograin <- log(cloud$rain)
> tapply(cloud$lograin, cloud$seeding, mean)
 control   seeded
3.990476 5.134252
> tapply(cloud$lograin, cloud$seeding, median)
 control   seeded
3.786259 5.396406
> tapply(cloud$lograin, cloud$seeding, sd)
 control   seeded
1.641897 1.599498
```

The standard deviations for the two groups are now very similar, so an analysis of the log transformed data would better meet the assumptions of ANOVA.

```
> t.test(lograin ~ seeding, data = cloud, var = TRUE)

    Two Sample t-test

data:  lograin by seeding
t = -2.5443, df = 50, p-value = 0.01408
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.0467013 -0.2408495
sample estimates:
mean in group control  mean in group seeded
             3.990476              5.134252
```

The difference in means is now significant at the 0.05 level.

(e) From this analysis, cloud seeding would be recommended. The analysis of the log transformed data (which better meets the assumptions) shows a significant increase in rainfall due to seeding.

4. (a)
```
> metal.lm.1 <- lm(hardness ~ strip + tip, data = metal)
> anova(metal.lm.1)
```

```
Analysis of Variance Table

Response: hardness
          Df Sum Sq  Mean Sq F value    Pr(>F)
strip      4  0.843 0.210750  29.753 3.809e-06 ***
tip        3  0.460 0.153333  21.647 3.928e-05 ***
Residuals 12  0.085 0.007083
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The very small P-value implies a highly significant difference between tips. Strip is a blocking factor, and so the small P-value is not of primary interest, except that it suggests that the blocking has been worthwhile.

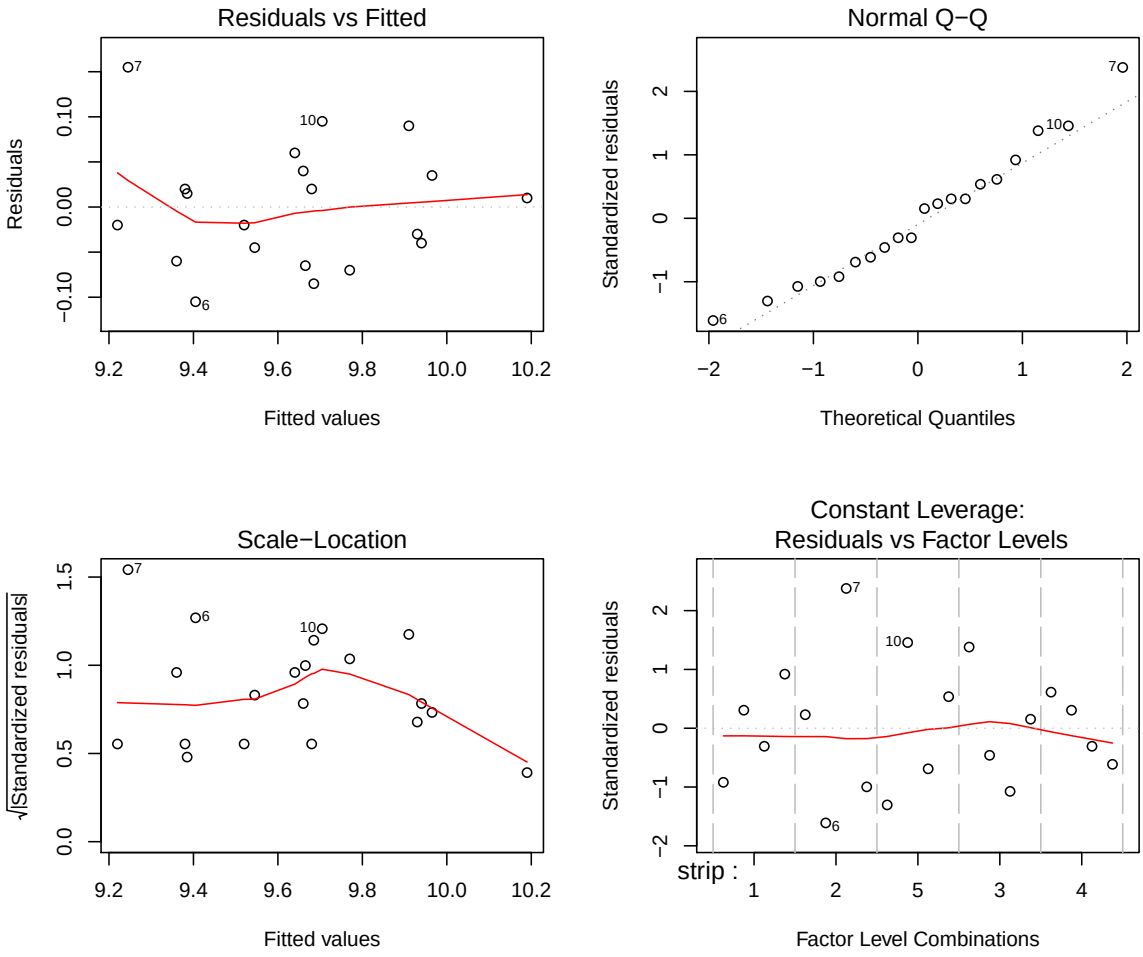
(b) `> qt(0.975, 12)`

```
[1] 2.178813
```

$2.179 \times \sqrt{0.00708\left(\frac{1}{5}+\frac{1}{5}\right)} = 0.12.$

95% confidence interval $= 9.88 - 9.46 \pm 0.12 = (0.30, 0.54)$.

(c)
```
> par(mfrow = c(2, 2))
> plot(metal.lm.1)
```



The plot of residuals vs fitted values is consistent with the assumption of constant variance, and the Q-Q plot is consistent with the assumption of normality of the errors.

(d) Pool degrees of freedom and sum of squares for strip and residuals:

Df: $4 + 12 = 16$.

Sum Sq: $0.843 + 0.085 = 0.928$.

Residual Mean Sq $= 0.928/16 = 0.058$.

$F = 0.1533/0.058 = 2.64$.

This results in:

```
          Df Sum Sq  Mean Sq F value
tip        3  0.460 0.153333    2.64
Residuals 16  0.928 0.058
```

Analysis in R:

```
> metal.lm.2 <- lm(hardness ~ tip, data = metal)
> anova(metal.lm.2)
Analysis of Variance Table

Response: hardness
          Df Sum Sq Mean Sq F value  Pr(>F)
tip        3  0.460 0.15333  2.6437 0.08464 .
Residuals 16  0.928 0.05800
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

(e) Blocking by strip in the experiment has been very worthwhile. It has substantially decreased the residual MS by removing a large amount of the random variation, allowing much more precise inference on the factor of interest (tip). This is reflected in a much more significant effect of tip.

5. (a)
```
> turnip <- read.csv("turnip.csv")
> tapply(turnip$weight, turnip$variety, mean)
        A         B         C         D         E         F
0.9958333 0.7866667 0.3716667 0.9660000 0.6100000 0.6503333
```

(b)
```
> turnip.lm.1 <- lm(weight ~ factor(row) + factor(col) + variety,
+     data = turnip)
> anova(turnip.lm.1)
Analysis of Variance Table

Response: weight
            Df  Sum Sq Mean Sq F value    Pr(>F)
factor(row)  5 0.16445 0.03289  1.0002    0.4429
factor(col)  5 0.08307 0.01661  0.5053    0.7688
variety      5 1.67234 0.33447 10.1712 5.839e-05 ***
Residuals   20 0.65768 0.03288
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The P-value is very small, so the (null) hypothesis is decisively rejected.

(c) Combine `col` and residual into new residual:

Df $= 5 + 20 = 25$;

Sum Sq $= 0.0831 + 0.6577 = 0.7408$;

new residual Mean Sq $= 0.7408/25 = 0.0296$.

It is not very different.

```
> turnip.lm.2 <- lm(weight ~ factor(row) + variety, data = turnip)
> anova(turnip.lm.2)
Analysis of Variance Table

Response: weight
            Df  Sum Sq Mean Sq F value   Pr(>F)
factor(row)  5 0.16445 0.03289   1.110   0.3802
variety      5 1.67234 0.33447  11.288 9.01e-06 ***
Residuals   25 0.74076 0.02963
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

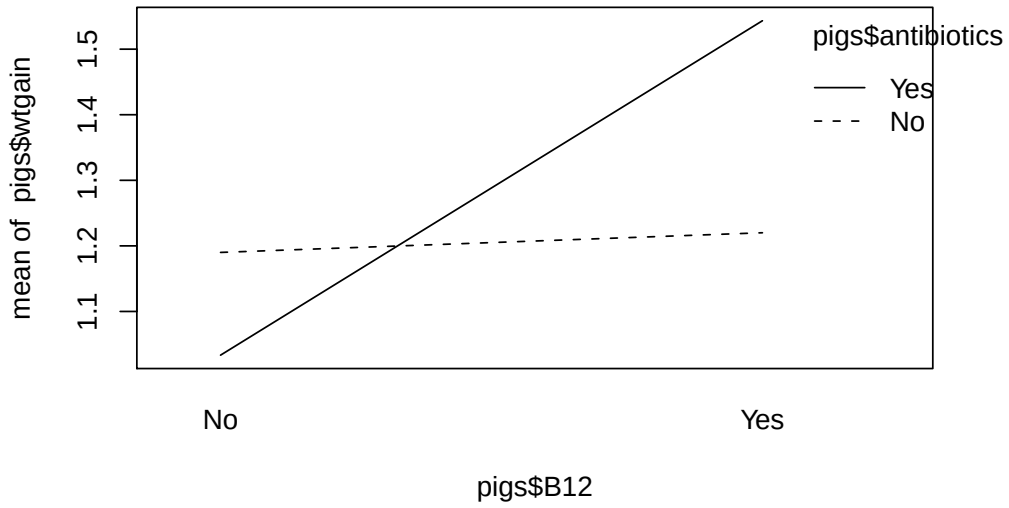
Not very different—the F statistics, P-values and residual Mean Sq do not change much.

6. (a) Randomised block design, with litters as blocks.

   (b)
```
> pigs <- data.frame(B12 = rep(c("No", "Yes"), each = 6), antibiotics = rep(c("No",
+     "Yes"), 6), litter = factor(c(1, 1, 2, 2, 3, 3, 1, 1, 2,
+     2, 3, 3)), wtgain = c(1.3, 1.05, 1.19, 1, 1.08, 1.05, 1.26,
+     1.52, 1.21, 1.56, 1.19, 1.55))
> tapply(pigs$wtgain, pigs[, 1:2], mean)
     antibiotics
B12     No      Yes
  No  1.19 1.033333
  Yes 1.22 1.543333
```

library(stats)

```
> interaction.plot(pigs$B12, pigs$antibiotics, pigs$wtgain)
```



There appears to be substantial interaction.

(c)
```
> pigs.lm.1 <- lm(wtgain ~ litter + B12 * antibiotics, data = pigs)
> anova(pigs.lm.1)
Analysis of Variance Table

Response: wtgain
                Df   Sum Sq  Mean Sq F value    Pr(>F)
litter           2 0.008717 0.004358  1.2684 0.3471931
B12              1 0.218700 0.218700 63.6475 0.0002066 ***
antibiotics      1 0.020833 0.020833  6.0631 0.0489646 *
B12:antibiotics  1 0.172800 0.172800 50.2894 0.0003946 ***
Residuals        6 0.020617 0.003436
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

$P < 0.001$, confirming the significant interaction.

(d) Treatment Mean Sq $= (0.2187 + 0.0208 + 0.1728)/(1 + 1 + 1) = 0.1374$.
$F = 0.1374/0.003436 = 40.0$, which is highly significant.

```
> pigs$treatment <- factor(c(1, 2, 1, 2, 1, 2, 3, 4, 3, 4, 3, 4))
> pigs.lm.2 <- lm(wtgain ~ litter + treatment, data = pigs)
> anova(pigs.lm.2)
```

```
Analysis of Variance Table

Response: wtgain
          Df  Sum Sq  Mean Sq F value    Pr(>F)
litter     2 0.00872 0.004358  1.2684 0.3471931
treatment  3 0.41233 0.137444 40.0000 0.0002319 ***
Residuals  6 0.02062 0.003436
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

7.

SOURCE	DF	SS	MS	F	P
Thinning	2	140	70	8.75	0.0030
Provenance	4	1200	300	37.5	<0.0001
Interaction	8	430	53.75	6.72	0.0008
Error	15	120	8		
Total	29	1890			

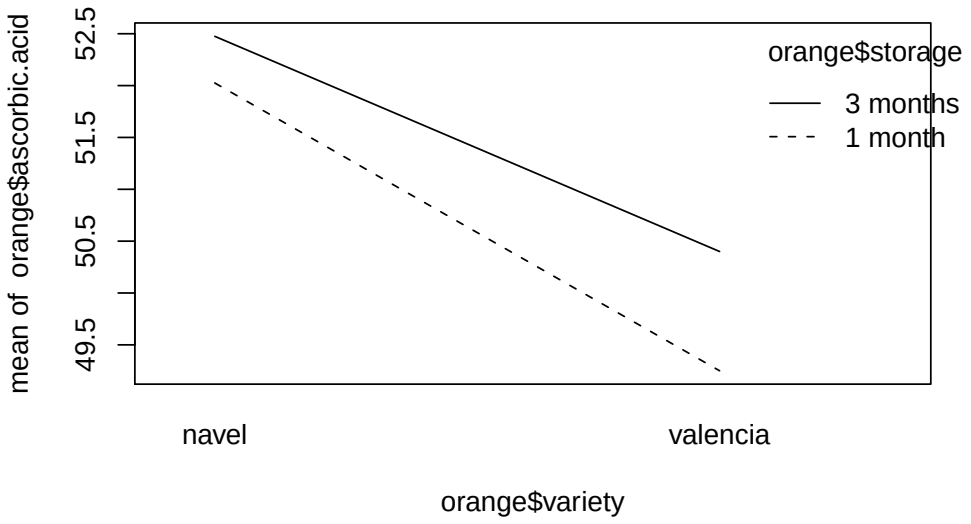
Calculation of P-values:

```
> 1 - pf(8.75, 2, 15)

[1] 0.003030814

> 1 - pf(37.5, 4, 15)

[1] 1.209653e-07

> 1 - pf(6.72, 8, 15)

[1] 0.000819122
```

8. (a)
```
> orange <- read.csv("orange juice.csv")
> tapply(orange$ascorbic.acid, orange[, 2:3], mean)
```

```
          storage
variety    1 month 3 months
  navel     52.025   52.475
  valencia  49.250   50.400
```

```
> interaction.plot(orange$variety, orange$storage, orange$ascorbic.acid)
```



There appears to be not much interaction.

(b)
```
> orange.lm <- lm(ascorbic.acid ~ variety * storage, data = orange)
> anova(orange.lm)
Analysis of Variance Table

Response: ascorbic.acid
                Df Sum Sq Mean Sq F value   Pr(>F)
variety          1 23.523 23.5225 11.9784 0.004709 **
storage          1  2.560  2.5600  1.3036 0.275823
variety:storage  1  0.490  0.4900  0.2495 0.626444
Residuals       12 23.565  1.9638
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The interaction and the main effect of storage time are not significant. The main effect of variety is highly significant, with the mean for navels 2 to 3 mg/litre higher than for valencia.

(c) $\hat{\sigma} = \sqrt{1.9638} = 1.40$.

(d) Recommendation: use navel oranges if possible; it doesn't matter whether the juice is stored for 1 month or 3 months.