Problem 1

Based on studies of livestock, one speculates whether cows exposed to substantial levels of pollution have larger fluoride levels in their intestinal system. If so, this could influence animal health as well as dairy production. Data (Table 1) of urinary fluoride concentration (parts per million) were measured for cows grazing near an industrial area (7 cows) and in an unpolluted area (5 cows).

Polluted	21.3	18.7	23.0	17.1	16.8	20.9	19.7
Unpolluted	14.2	18.3	17.2	18.4	20.0		

Table 1: Data showing urinary fluoride levels in cows grazing in two different areas.

In part a) and b) below, we assume that the data in the two groups are distributed according to $X_i \sim N(\mu_1, \sigma_1^2)$, $i = 1, \ldots, 7$, $Y_j \sim N(\mu_2, \sigma_2^2)$, $j = 1, \ldots, 5$. We further assume independence in the data.

a) From the data, we get pooled variance (sammenslått varians) $s^2_{\text{pool}} = 2.23^2$.

Conduct a one-sided T-test for equality of mean values in the two groups. Use $\alpha = 0.05$ significance level.

b) In this situation with data in Table 1, we get $s_1^2 = \frac{1}{6}\sum_{i=1}^{7}(x_i - \bar{x})^2 = 2.274^2$, $s_2^2 = \frac{1}{4}\sum_{j=1}^{5}(y_j - \bar{y})^2 = 2.157^2$.

Conduct a hypothesis test at significance level $\alpha = 0.05$ for equal variance in the two groups.

Does the result matter for what you did in question a)?

Alternative tests for equality of mean or median in this two-sample situation include the non-parametric Wilcoxon rank-sum test (to-utvalgstest).

c)

Formulate the one-sided Wilcoxon rank-sum test for the current setting. Use an exact test or a normal approximation test to conduct the hypothesis testing at significance level $\alpha = 0.05$. Compare with the result in a).

Problem 2

Billy is in his cabin with friends, celebrating Easter holiday. Easter Sunday is the first Sunday after the first full moon after March equinox (vårjevndøgn). Ignoring equinox time variability and leap-year/time-zone problems, assume that equinox is on 21 March. This means that Easter Sunday can occur on any date between 23 March and 23 April. Billy thinks, like most of his friends, that no date is more likely than others in this interval. But he is doing a statistics class this semester, and he starts comtemplating whether the date of Easter Sunday is uniformly distributed between these limits or not. He looks through old calenders stored in the cabin. Every year since 1971 (54 years in total), he finds the date of Easter Sunday. He makes a list where 1 corresponds to 23 March, 2 corresponds to 24 March, and so on, until 32 which corresponds to 23 April. He decides to bin the entries in the listed data in 4 bins: (1-8), (9-16), (17-24) and (25-32). Data are in Table 2.

Bins (Days after equinox)	(1-8)	(9-16)	(17-24)	(25-32)
Number of years	9	15	15	15

Table 2: Data showing date distribution of Easter Sunday, counting days after equinox. The number of years data in bins represent the times since 1971 (54 years in total) that Easter Sunday was this many days after equinox.

a)

Formulate Billy's reflections about uniformly distributed dates of Easter Sunday as a hypothesis test.

Use a goodness-of-fit statistic to conduct the hypothesis test for this data. Use $\alpha = 0.05$ significance level.

Going through the list of dates, Billy recognizes a kind of pattern in the dates occurring in consecutive years. For each of 53 consecutive year pairs, he marks the first date (Day 1) and the second date (Day 2). He conducts the similar binning for each outcome in these date pairs. The data are summarized in Table 3.

b)

Are Easter Sunday dates of two consecutive years independent? Formulate this as a hypothesis test with the statistical model under the null hypothesis.

Use an approximate test for contingency tables to conduct the hypothesis test for this data, with a significance level of $\alpha = 0.05$. Comment on the validity for the approximation.

	Day 1(1-8)	Day 1(9-16)	Day 1(17-24)	Day 1(25-32)	
Day 2(1-8)	0	0	5	4	9
Day 2(9-16)	4	0	0	11	15
Day 2(17-24)	4	10	0	0	14
Day 2(25-32)	0	5	10	0	15
	8	15	15	15	53

Table 3: Data showing joint date (after equinox) of Easter Sunday in 53 consecutive years.

Problem 3

In a study of North American bluegill fish, scientists gathered length data of 20 fish. In addition to the length of the fish, they registered the age of the fish and the altitude and pH level in the lake that the fish were caught. For multiple linear regression modeling, the response is length (y) and covariates are age (x_1), altitude (x_2) and pH level (x_3).

a)

```
Call:
lm(formula = y ~ age + alt + ph)

Residuals:
    Min      1Q  Median      3Q     Max
-37.555 -13.028  -1.005  17.676  22.535

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 172.31786   85.76092   2.009 0.061686 .
age          15.30603   [redacted]  4.647 0.000269 ***
alt           0.04854    0.07843  [redacted] 0.544686
ph          -20.42423   12.52085  -1.631 [redacted]
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.44 on 16 degrees of freedom
Multiple R-squared:  0.7051,    Adjusted R-squared:  0.6498
F-statistic: 12.75 on 3 and 16 DF,  p-value: 0.000164
```

b)

```
Call:
lm(formula = y ~ age)

Residuals:
    Min      1Q  Median      3Q     Max
-35.444 -15.480  -0.918  17.442  28.257

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   48.802     12.569   3.883  0.00109 **
age           16.200      2.807   5.771 1.81e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.99 on 18 degrees of freedom
Multiple R-squared:  0.6492,    Adjusted R-squared:  0.6297
F-statistic: 33.31 on 1 and 18 DF,  p-value: 1.806e-05
```

Figure 1: a) R print-out of the multiple linear regression fit for length with covariates age, altitude and pH level. b) R print-out of the linear regression fit for length (mm) with covariate age (year).

a)

Write down the multiple linear regression model for this situation.

Describe the statistical modeling assumptions (modellantakelsene).

Figure 1 a) shows a standard print-out from R. Find approximate values for the numbers that are marked out with gray boxes in this display.

b) We will now focus on length [mm] as a function of the age [year] covariate alone, with model $y_i = \beta_0 + \beta_i x_{1,i} + \epsilon_i$ and independent $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, 20$. Figure 1 b) shows a standard print-out of the linear regression fit from R.

Use the print-out to construct a 90 % confidence interval for the length growth of a fish per year.

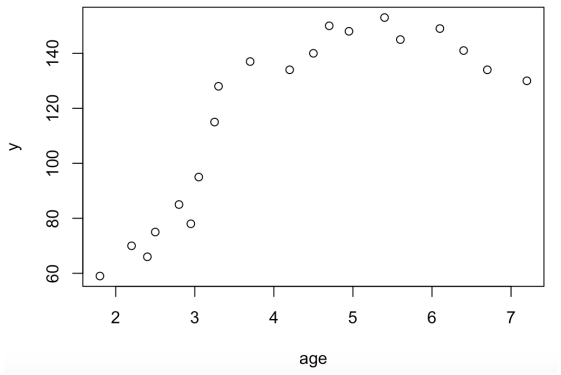
Predict the length for fish of age 2, 4 and 7 year. Sketch the fitted line with the data (without details). Comment on the residuals one gets from this model.

c) Figure 2 a) shows a cross plot of $(x_{1,i}, y_i)$, $i = 1, \ldots, 20$. Based on this plot, researchers switched to a quadratic model for length as a function of age.

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{1,i}^2 + \epsilon_i, \quad i = 1, \ldots, 20.$$

Figure 2 b) shows the R print-out for this situation.

a) b)




```
Call:
lm(formula = y ~ age + age2)

Residuals:
     Min       1Q   Median       3Q      Max
-21.2681  -4.8541  -0.1554   4.3487  16.0225

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -81.6882    17.2046  -4.748 0.000186 ***
age          83.7151     8.4675   9.887 1.83e-08 ***
age2         -7.5844     0.9397  -8.071 3.23e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.356 on 17 degrees of freedom
Multiple R-squared:  0.9274,    Adjusted R-squared:  0.9189
```

Figure 2: Left) Plot of length (mm, second axis) and age (year, first axis) of bluegill fish. Right) R-print out of the quadratic regression model for length with covariate age and age^2.

How are R^2 and R^2_{adj} defined? Compare and discuss these values in the three different models in a), b) and c).

Relying on the quadratic model for length as a function of age, next:

Estimate the age at which the expected length of a bluegill fish is the largest.

Use linearization to approximate the variance of this age estimate. (The covariance between $\hat{\beta}_1$ and $\hat{\beta}_2$ is -7.86.)

Problem 4

Data of production efficiency y in a 2^{4-1} fractional factorial design are shown in Table 4.

A	B	C	D	y
-1	-1	-1	-1	88.1
1	-1	-1	1	92.2
-1	1	-1	1	90.1
1	1	-1	-1	93.7
-1	-1	1	1	92.1
1	-1	1	-1	92.7
-1	1	1	-1	93.1
1	1	1	1	92.9

Table 4: A fractional factorial 2^{4-1} design with factors A, B, C and D along with response y.

a)

What is the estimated effect of factor A?

What is the generator and the confounding (alias) structure of this fractional factorial experiment?

Ignoring any interaction effects, we can estimate the variance in this experiment. We can then study the significance of main effects.

b)

Specify the variance of the estimated effects.

Is the effect of A significant at level $\alpha = 0.05$?