1. (8.5 points) Gian would like to assess how to increase the sales of the yogurts he produces. Specifically, he wants to construct a good model to predict `score` (assumed to be a numerical continuous varible), using the following predictor variables:

   - `sugar`: grams of sugar per 100g of yogurt,
   - `milk`: grams of milk per 100g of yogurt,
   - `fat`: grams of fat per 100g of yogurt,
   - `proteins`: grams of proteins per 100g of yogurt.

   To begin the analysis, Gian looks at the multivariate linear model. The following R output is the R summary of that fitted model.

```
Call:
lm(formula = score ~ ., data = yogurt)

Residuals:
     Min       1Q   Median       3Q      Max
-2.66324 -0.69785  0.03114  0.60478  2.84334

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.807684   0.287953   6.278 3.06e-09 ***
sugar        3.022989   0.042683  70.823  < 2e-16 ***
fat         -0.005531   0.016207  -0.341    0.733
milk         0.030014   0.079475   0.378    0.706
protein      0.093920   0.079267   1.185    0.238
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.042 on 161 degrees of freedom
Multiple R-squared:  0.9696,        Adjusted R-squared:  0.9689
F-statistic:  1286 on 4 and 161 DF,  p-value: < 2.2e-16
```

   (a) (1 point) Write the equation for the estimated model in the form:

   $$\hat{y} = \hat{\beta}_0 + \dots$$

   Use the summary to obtain the estimated coefficients.

   Solution:

   $\hat{y} = 1.808 + 3.023 * \texttt{sugar} - 0.005 * \texttt{fat} + 0.03 * \texttt{milk} + 0.094 * \texttt{protein}$.

   (Or use x_1, x_2,... instead, but it's important to define each of these predictors in that case.)

   (b) (1 point) Provide an interpretation of the coefficient of `fat`.

   Solution: The expected average change in score when the fat content is increased by 1 and all other predictors remain the same.

   (c) (1 point) Provide a 95% confidence interval for the coefficient of `fat`. You can use the following information:

   - If $X \sim t_{165}$, then $P(X \leq 1.974) = 0.975$
   - If $X \sim t_{41}$, then $P(X \leq 2.019) = 0.975$
   - If $X \sim t_{40}$, then $P(X \leq 2.021) = 0.975$
   - If $X \sim t_{161}$, then $P(X \leq 1.975) = 0.975$

Solution: We know that

$$\frac{\hat{\beta}_{fat} - \beta_{fat}}{SE(\hat{\beta}_{fat})} \sim t_{161},$$

we can read off the degrees of freedom from the output of the global F-test. Hence, the 95% confidence interval for the coefficient of `fat` is

$$(\hat{\beta}_{fat} - 1.975 * SE(\hat{\beta}_{fat}), \hat{\beta}_{fat} + 1.975 * SE(\hat{\beta}_{fat})) \approx (-0.0375, 0.0265).$$

(d) (1 point) Based on the summary and/or your confidence interval, would it be reasonable to drop `fat` from the model? Why or why not?

Solution: Yes, the p-value corresponding to `fat` is $\approx .73$, which indicates that under $H_0 : \beta_{fat} = 0$ the data we observed does not seem at all unlikely.

A simple linear regression of `score` on `sugar` leads to the following R output:

```
Call:
lm(formula = score ~ sugar, data = yogurt)

Residuals:
     Min       1Q   Median       3Q      Max
-2.63014 -0.81279  0.03308  0.59263  2.82079

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.98643    0.11607   17.11   <2e-16 ***
sugar        3.01875    0.04193   72.00   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.037 on 164 degrees of freedom
Multiple R-squared:  0.9693,        Adjusted R-squared:  0.9691
F-statistic:  5183 on 1 and 164 DF,  p-value: < 2.2e-16
```

(e) (1 point) What is the root mean squared error of the model?

Solution: The root MSE is $\hat{\sigma} = 1.037$

(f) (1 point) What is the absolute difference in the residual variance estimates between the two models?

Solution: $|1.037^2 - 1.042^2| = |1.075369 - 1.085764| = 0.010395$

(g) (1 point) Carefully state the null hypothesis for a partial F test (ANOVA test) comparing this model to the model above.

Solution: H_0: $\beta_{fat} = \beta_{milk} = \beta_{protein} = 0$.

(h) (1.5 points) What is the F statistic for testing whether any of the predictors `sugar, protein, milk, fat` are significant? Specify the distribution this test statistic follows if none of the predictors `sugar, protein, milk, fat` are significant. What is the p-value of this F-statistic?

Solution:

F statistic:

$$F = \frac{(RSS_0 - RSS_4)(161)}{4 * RSS_4},$$

where RSS_0 is the residual sum of squares after fitting an empty model (just the intercept) and RSS_4 is the the residual sum of squares after fitting the model with `sugar, protein, milk, fat` as predictors (that is the model which was fit at the beginning of this question).

If H_0 is true (that is if none of the mentioned predictors are significant), the F statistic follows $F_{4,161}$ distribution and the p-value of it is: $< 2.2 \cdot 10^{-16}$ (see summary at the beginning of the question).

2. (8.5 points) Andreas would like to assess how to improve the quality of the coffee in his office. He wants to model the response `score` (assumed to be a numerical continuous varible), using the following predictor variables:

   - `water quality`: numerical continuous scale,
   - `roasting`: numerical continuous scale,
   - `strength`: numerical continuous scale,
   - `tempearature`: numerical continuous scale,
   - `origin`: 0, 1 indicating the region of origin, a factor

   To begin the analysis, we look at the linear model with interaction terms. The following R output is the R summary of that fitted model.

```
lm(formula = score ~ (water + roasting + strength + temeprature + origin)^2,
data = coffee)

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            8.915226   1.119475   7.963 3.31e-12 ***
water                  1.330888   0.109961  12.103  < 2e-16 ***
roasting              -0.003289   0.216295  -0.015   0.9879
strength               0.079470   0.246080   0.323   0.7472
temperature           -0.153715   0.519494  -0.296   0.7678
origin1                0.012086   0.155129   0.078   0.9380
water:roasting        -0.034419   0.185652  -0.185   0.8532
water:strength        -0.120053   0.190112  -0.631   0.5288
water:temperature      0.409800   0.374226   1.095   0.2755
water:origin1          0.123756   0.150112   0.824   0.4112
roasting:strength     -0.026629   0.070166  -0.380   0.7049
roasting:temperature   0.071436   0.800531   0.089   0.9290
roasting:origin1       0.173919   0.333485   0.522   0.6029
strength:temperature  -0.274557   0.980973  -0.280   0.7800
strength:origin1       1.703945   0.391175   4.356 2.61e-05 ***
temperature:origin1    1.577546   0.740012   2.132   0.0348 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9024 on 134 degrees of freedom
Multiple R-squared:  0.8085,Adjusted R-squared:  0.7871
F-statistic: 37.73 on 15 and 134 DF,  p-value: < 2.2e-16
```

   (a) (1 point) Write only the predictor variables significant at the 1% level according to the individual hypothesis tests reported above.

   Solution: `water`, and the interaction `strength`·`origin`.

   (b) (2 points) If you were to fit the model using only the predictors you wrote down in a, this model would be dependent on the additive changes in scale of certain predictors. Why? How could you solve this issue?

Solution: The issue is that we are including the interaction `strength` · `origin` without the main effects of strength and origin. So if we transform strength as, $strength' = strength + a$ and then fit the model

$$\begin{aligned}\underline{score} &= \beta_0 + \beta_1 \underline{water} + \beta_2 \underline{strength'} \cdot \underline{origin} \\ &= \beta_0 + \beta_1 \underline{water} + \beta_2 (\underline{strength} + a) \cdot \underline{origin} \\ &= \beta_0 + \beta_1 \underline{water} + \beta_2 \cdot a \cdot \underline{origin} + \beta_2 \underline{strength} \cdot \underline{origin},\end{aligned}$$

we suddenly have a linear dependence on origin (we will be fitting different intercepts for different values of origin) which was not present before!

We should additionally include main effects of strength and origin to solve this issue.

(c) (1 point) What is the fitted value of score for water=1, roasting =5, strength=10, temperature = 100 and origin=0?

Solution:

$$\begin{aligned}\widehat{score} &= 8.915 + 1.330 + (-0.003) * 5 + 0.079 * 10 + (-0.154) * 100 + \\ &\quad + (-0.034) * 1 * 5 + (-0.12) * 1 * 10 + 0.409 * 1 * 100 + 0.071 * 10 * 100 \\ &= 106.15\end{aligned}$$

(d) (2 points) Andreas now decides to fit the following model:

```
Call:
lm(formula = score ~ strength + origin + strength:origin, data = coffee)

Residuals:
     Min       1Q   Median       3Q      Max
-12.0677  -2.6464  -0.0875   2.7378  10.1776

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        8.8657     2.0577   4.309 3.00e-05 ***
strength           2.8045     0.2382  11.776  < 2e-16 ***
origin1           -6.1207     2.6303  -2.327   0.0213 *
strength:origin1   1.7821     0.3130   5.694 6.59e-08 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.074 on 146 degrees of freedom
Multiple R-squared:  0.8331,        Adjusted R-squared:  0.8296
F-statistic: 242.8 on 3 and 146 DF,  p-value: < 2.2e-16
```

The estimated model represents two lines with `score` as the dependent variable and `strength` as the independent variable. Specify the intercept and slope for these two fitted lines.

Solution: The intercept and the slope for the fitted line when `origin` = 0: intercept - $\widehat{\beta_0} = 8.866$ and slope - $\widehat{\beta_1} = 2.804$.

The intercept and the slope for the fitted line when `origin` = 1: intercept - $\widehat{\beta_0} + \widehat{\beta_3} = 8.866 - 6.121 = 2.745$ and slope - $\widehat{\beta_1} + \widehat{\beta_4} = 2.804 + 1.782 = 4.586$.

(e) (1 point) What is the difference in the estimated `score` for a coffee of strength 5 compared to coffee of strength 10 when origin = 0.

Solution: This difference is equal to $\widehat{\beta_1} * 5 = 2.804 * 5 = 14.02$.

(f) (1.5 points) What does the model which was fit in (d) assume about the distribution of errors for the data on coffees of origin =0 and coffees of origin = 1 (Answer in terms of mean, variance, and correlation).

Solution: The OLS regression assumes that the errors have mean 0, are uncorrelated and have identical (constant) variance. (0.5 points for each correct assumption)

3. (5 points) Five independent hypothesis tests were performed and the following p-values were obtained:

- For null hypothesis A - 0.011
- For null hypothesis B - 0.027
- For null hypothesis C - 0.017
- For null hypothesis D - 0.008
- For null hypothesis E - 0.023

(a) (1 point) Suppose each test was performed at the significance level 0.05 without any correction. Which null hypotheses would be rejected with this criterion?

Solution: Since all p-values are smaller than 0.05 all of them would be rejected in this case.

(b) (1 point) Without applying any error correction and performing each of the above 5 tests at the significance level of 0.05, what is the probability of having at least one false rejection?

Solution: Probability of a false rejection in one of the tests is .05.

The probability of at least one false rejection assuming that all null hypothesis A, B, C, D, E are true (and that the tests are independent as indicated above) is: $1-(1-0.05)^5 = 1-.95^5 \approx .226$. (The final calculation is not necessary).

(c) (1 point) Suppose you apply Bonferroni correction procedure with FWER control at 5%. Which null hypotheses would be rejected in this case?

Solution: Bonferroni correction compares each of the above p-values with a corrected significance level $.05/5 = 0.01$. Based on the Bonferroni procedure only null hypothesis D would be rejected. (1 P.)

(d) (1 point) Suppose you apply Holm correction procedure with FWER control at 5%. Which null hypotheses would be rejected in this case?

Solution: We first sort the p-values:

$$0.008 < 0.011 < 0.017 < 0.023 < 0.027.$$

We form the appropriate adjusted significance level $0.05/(6-i), i \in \{1, \ldots, 5\}$:

$$0.01 < 0.0125 < 0.0167 < 0.025 < 0.05.$$

Then we find the smallest p-value that is larger than its corresponding adjusted significance level. In this case that is p-value: 0.017. So we can only reject the null hypotheses corresponding to p-values 0.008 and 0.011, that is D and A. (1 P.)

(e) (1 point) Suppose you apply Benjamini-Hochberg correction procedure with FDR control at 5%. Which null hypotheses would be rejected in this case?

Solution: We first sort the p-values:

$$0.008 < 0.011 < 0.017 < 0.023 < 0.027.$$

We form the appropriate adjusted significance level $0.01 * i, i \in \{1, \ldots, 5\}$:

$$0.01 < 0.02 < 0.03 < 0.04 < 0.05.$$

Then we find the largest p-value that is smaller than its corresponding adjusted significance level. In this case the largest p-values is smaller than 0.05. Hence, we reject all null hypotheses A, B, C, D, E. Notice how this corresponds to question b). (1 P.)

4. (4 points) Niklas has a data set of n=100 observation of one response variable `Y` and one predictor variable `X`. He fits a simple linear regression model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ i = 1, \ldots, n$ and performs some model diagnostics.

   (a) (1 point) Niklas decides to take a look at the Tukey-Anscombe plot first. Which of the following could be **detected** using the Tukey-Anscombe plot?

   ◯ Heteroskedasticity (non-constant variance) of the errors.

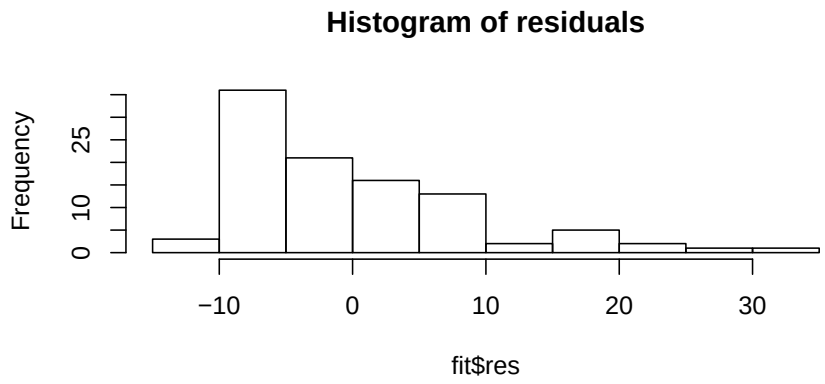
   ◯ Whether the maximum difference of the observed response and the fitted response is larger than 2.

   ◯ The presence of systematic error (non-zero mean).

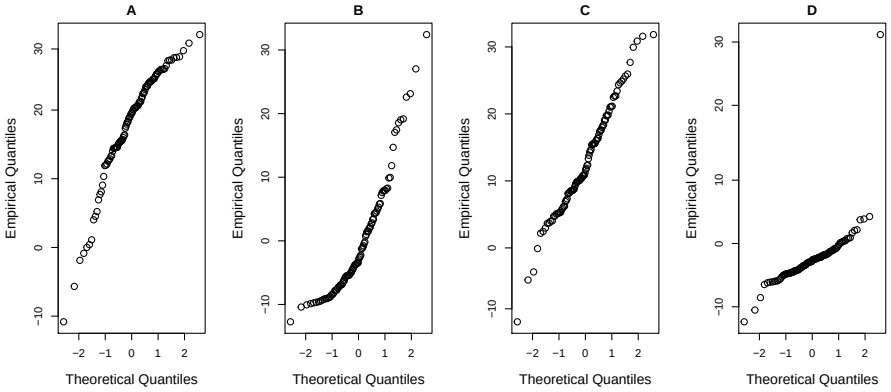
   √ **All of the above.**

   Solution: The Tukey-Anscombe plot plots residuals against the fitted values of the response. It is used as a model diagnostic tool to check for non-zero mean and non-constant variance of the errors.

   (b) (1 point) The histogram of the residuals $r_i = y_i - \hat{y}_i,\ i = 1, \ldots, n$ in the above model is the following:



   Which of the QQ-plots given below would you expect to correspond to the residuals based on the histogram above?



   ◯ A, because the residuals are left skewed.

√ **B, because the residuals are right skewed.**

◯ C, because the residuals are right skewed.

◯ D, because the residuals are left skewed.

Solution: B, The distribution of the residuals appears to be right skewed.

(c) (1 point) Which of the following assumptions on the errors is **not** needed for the Gauss-Markov theorem to hold in the case of ordinary least squares regression?

◯ The errors have expectation zero.

√ **The errors are normally distributed.**

◯ The errors are uncorrelated.

◯ The errors have a constant variance.

Solution: For the OLS regression to be the best linear unbiased estimator, that is for the OLS to satisfy the Gauss-Markov theorem, we only need the assumptions the errors to have expectation zero, constant variance and to be uncorrelated. The assumption of iid Gaussian errors is often made to calculate p-values, confidence intervals etc., however this assumption is not necessary for the Gauss-Markov theorem to hold.

(d) (1 point) Which of the following statements concerning a QQ-plot of the residuals is **true**?

◯ It plots the empirical quantiles of the studentized residuals against the corresponding theoretical quantiles of a normal random variable.

◯ It plots the raw residuals against their empirical quantiles.

◯ It plots the studentized residuals against their empirical quantiles.

√ **It plots the empirical quantiles of the standardized residuals against the corresponding theoretical quantiles of a normal random variable.**

5. (5 points) Are the following statements true or false? Explain your reasonsing.

(a) (3 points) Below you see the summary output of a linear model which was fitted to an artificial data set of $n = 101$ observations of the response variable `y` and the predictor variables `x1` and `x2`. Assume that the corresponding Tukey-Anscombe plot does not show any model violations.

```
Call:
lm(formula = y ~ x1 + x2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.85158 -0.53488  0.01632  0.56072  2.46917

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.3176     0.6136  -2.147   0.0342 *
x1            0.8290     0.1533   5.409 4.48e-07 ***
x2           -0.1622     0.1551  -1.046   0.2980
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9048 on 98 degrees of freedom
Multiple R-squared:  0.2366,        Adjusted R-squared:  0.221
F-statistic: 15.18 on 2 and 98 DF,  p-value: 1.802e-06
```

i. (1 point) The null hypothesis $H_0 : \beta_1 = 0$ corresponding to the predictor `x1` is certainly wrong.

Solution: False, we cannot make such definite statements about whether H_0 is true or not. We can only say that we reject the H_0 hypothesis at the 5% level.

ii. (1 point) The null hypothesis $H_0 : \beta_2 = 0$ corresponding to the predictor `x2` is certainly true.
Solution: Again, false. See above.

iii. (1 point) The test for the parameter corresponding to `x2` does *not* reject the null hypothesis $\beta_2 = 0$ at a 30% level.
Solution: False, the p-value for β_2 is smaller than 0.3 and hence, the corresponding test rejects H_0.

(b) (2 points) Consider the following two statements about the p-values in the column indicated by `Pr(>|t|)` in the above summary output.

i. (1 point) The t-values stay the same if a level α different from 0.05 is chosen.
Solution: True. The significance level α is not used to calculate the t-values.

ii. (1 point) The p-values change if a level α different from 0.05 is chosen.
Solution: False. The significance level α is not used to calculate the p-values. They are compared to the significance level and the test decision is made based on whether the p-value is smaller or larger than α.

6. (Bonus: 5 points) (Linear transformations preserve information) You have regressed y on predictors $x_1, x_2, \ldots, x_p$. Your colleague, Bob, has regressed Y on the variables $z_1, z_2, \ldots, z_p$, where

$$z_j = c_{j0} + \sum_{k=1}^{p} c_{jk} x_k$$

That is, Bob has applied a linear transformation to the predictors (but not to the response).

You can use the following: For two constant matrices A and B:

- $(AB)' = B'A'$

For two constant and **square** matrices A and B:

- $(AB)^{-1} = B^{-1}A^{-1}$

(a) (2 points) Show that Bob's $n \times (p+1)$ design matrix Z is related to yours via

$$Z = XT$$

for some $(p+1) \times (p+1)$ matrix T.
Explain how the entries in T are related to Bob's coefficients c.
Solution: The design matrix X is:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & x_{23} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \ldots & x_{np} \end{bmatrix}.$$

The design matrix Z is

$$Z = \begin{bmatrix} 1 & z_{11} & z_{12} & z_{13} & \dots & z_{1p} \\ 1 & z_{21} & z_{22} & z_{23} & \dots & z_{2p} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & z_{n1} & z_{n2} & z_{n3} & \dots & z_{np} \end{bmatrix}$$
$$= \begin{bmatrix} 1 & c_{10} + \sum_{k=1}^{p} c_{1k} x_{1k} & c_{20} + \sum_{k=1}^{p} c_{2k} x_{1k} & \dots & c_{p0} + \sum_{k=1}^{p} c_{pk} x_{nk} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & c_{10} + \sum_{k=1}^{p} c_{1k} x_{nk} & c_{20} + \sum_{k=1}^{p} c_{2k} x_{nk} & \dots & c_{p0} + \sum_{k=1}^{p} c_{pk} x_{nk} \end{bmatrix}.$$

Hence,

$$T = \begin{bmatrix} 1 & c_{10} & c_{20} & \dots & c_{p0} \\ 0 & c_{11} & c_{21} & \dots & c_{p1} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & c_{1p} & c_{2p} & \dots & c_{pp} \end{bmatrix}.$$

(b) (2 points) Using the hat matrices of the two regressions, show that your fitted values and Bob's fitted values are exactly equal, if T is invertible.

Solution:
Using $H = X(X'X)^{-1}X'$, $\underline{\hat{y}} = H\underline{y}$ and above, we have that for Bob's regression the hat matrix is

$$\begin{aligned} Z(Z'Z)^{-1}Z' &= XT((XT)'(XT))^{-1}(XT)' \\ &= XT(T'X'XT)^{-1}T'X' \\ &= XT(X'XT)^{-1}(T')^{-1}T'X' \\ &= XTT^{-1}(X'X)^{-1}X' \\ &= X(X'X)^{-1}X' = H. \end{aligned}$$

So the fitted values from both regressions are the same.

(c) (1 point) Show that, if $\underline{\hat{\beta}}$ is your vector of coefficients, and if T is invertible, then Bob's vector of coefficient estimates is exactly

$$T^{-1}\underline{\hat{\beta}}$$

.

Solution: We use part of the calculation from above. Bob's vector of coefficient estimates is exactly

$$\begin{aligned} (Z'Z)^{-1}Z' &= ((XT)'(XT))^{-1}(XT)' \\ &= (T'X'XT)^{-1}T'X' \\ &= (X'XT)^{-1}(T')^{-1}T'X' \\ &= T^{-1}(X'X)^{-1}X' \\ &= T^{-1}\underline{\hat{\beta}}. \end{aligned}$$