Exercises for Linear Statistical Models

1. Suppose we have data $x_1, \dots, x_n$ and $y_i, \dots, y_n$, which we model as

$$Y_i = b_0 + b_1 x_i + \varepsilon_i, \quad \varepsilon_i \overset{ind}{\sim} N(0, \sigma^2).$$

   (a) Write down the formula for the sum of squares criterion.

   (b) Take the partial derivatives of the sum of squares criterion to derive the normal equations.

   (c) Solve the normal equations to show that the least squares estimates are

$$\widehat{b}_1 = \frac{\sum_{i=1}^n (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2}$$
$$\widehat{b}_0 = \overline{y} - \widehat{b}_1 \overline{x}$$

2. In terms of the estimates $\widehat{b}_0$ and $\widehat{b}_1$, write down the formulas for the fitted values, the residuals, and $\widehat{\sigma}^2$ (the variance estimate).

3. Explain the difference between an estimate and an estimator.

4. Write down how you would explain all of the assumptions of the simple linear model for $y_1, \dots, y_n$ given $x_1, \dots, x_n$ to your grandmother (who was a chemical engineer).

5. Suppose we have data $x_1, \dots, x_n$ and $y_1, \dots, y_n$. The quantity sxy is given to the left of the equals sign and is always equal to **exactly one** of the three expressions to the right of the equals sign. Circle the correct expression and show why sxy is equal to it.

$$\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y}) \quad = \quad \sum_{i=1}^n (x_i - \overline{x}) y_i \qquad \sum_{i=1}^n (x_i - \overline{x}) \overline{y} \qquad \sum_{i=1}^n \overline{x}(y_i - x_i)$$

6. Derive a formula for $\mathrm{Cov}(\widehat{B}_0, \widehat{B}_1)$

7. Use the formulas for the variances and covariances of $\widehat{B}_0$ and $\widehat{B}_1$ to derive a simplified expression for the prediction variance $\mathrm{Var}(\widehat{B}_0 + \widehat{B}_1 x_i)$.

8. Suppose we have data $x_1, \dots, x_n$ and $y_1, \dots, y_n$. Which of the following equations qualify as statistical models for $y_1, \dots, y_n$?

$$Y_i = b_0 + b_1 x_i + \varepsilon_i, \quad \varepsilon_i \overset{ind}{\sim} N(0, \sigma^2)$$
$$Y_i = b_0 + b_1 x_i$$
$$Y_i = e^{\cos(x_i) + \varepsilon_i}, \quad \varepsilon_i \overset{ind}{\sim} N(0, \sigma^2)$$
$$0 \le Y_i \le 10$$

9. Under the simple linear model, there is a formula for the least squares estimators of the regression coefficients. Which property of the estimators allows us to conclude that the estimators are normally distributed?

10. Select all that are correct: A confidence interval for the slope b_1 . . .
    (a) contains the slope values that we rejected using a hypothesis test
    (b) contains the slope values that we failed to reject using a hypothesis test
    (c) is a set of plausible values of the slope given the data
    (d) has endpoints that are random variables
11. Name two desirable properties of estimators, and explain why they are desirable.
12. Let

$$\boldsymbol{Y} = X\boldsymbol{b} + \boldsymbol{\varepsilon}, \quad \varepsilon_i \overset{ind}{\sim} N(0, \sigma^2 I)$$

    where $\boldsymbol{Y}$ is $n \times 1$ and X is $n \times p + 1$.

    What are the dimensions of $\boldsymbol{b}$ and $\boldsymbol{\varepsilon}$?

    What is the expectation of $\boldsymbol{Y}$?

    What is the expectation of $X^T\boldsymbol{Y}$?

    What is the expectation of $\widehat{\boldsymbol{B}} = (X^TX)^{-1}X^T\boldsymbol{Y}$?

    What is the covariance matrix for ε?

    What is the covariance matrix for $\widehat{\boldsymbol{B}}$?
13. Recall the simple linear model

$$Y_i = b_0 + b_1 x_i + \varepsilon_i, \quad \varepsilon_i \overset{ind}{\sim} N(0, \sigma^2).$$

    Write down the design matrix X.

    Calculate the entries of X^TX

    Calculate the entries of $(X^TX)^{-1}$

    Calculate $X^T\boldsymbol{y}$.

    Show that $\widehat{\boldsymbol{b}} = (X^TX)^{-1}X^T\boldsymbol{y}$ produces the same least squares estimates that we originally derived. You'll have to use the formula for the inverse of a 2×2 matrix.
14. The residual sum of squares criterion for the multiple linear model is

$$rss(\boldsymbol{b}^*) = \sum_{i=1}^{n}(y_i - \sum_{j=0}^{p} b_j^* x_{ij})^2$$

    Derive the kth normal equation by differentiating the residual sum of squares.
15. Show that the set of $p + 1$ normal equations can be written as $X^T\boldsymbol{y} = X^TX\widehat{\boldsymbol{b}}$

16. Let

$$X = \begin{bmatrix} \boldsymbol{x}_0 & \cdots & \boldsymbol{x}_p \end{bmatrix}$$

And suppose for this problem that $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_p$ are orthonormal, which means that $\boldsymbol{x}_j^T \boldsymbol{x}_k = 0$ if $j \neq k$ and 1 if $j = k$.

Show that the least squares estimates of $\boldsymbol{b}$ are $\widehat{b}_j = \boldsymbol{x}_j^T \boldsymbol{y}$.

17. What are the two defining properties of projection matrices?

18. Show that $X(X^T X)^{-1} X^T$ is a projection matrix.

19. Suppose that

$$U = \begin{bmatrix} 1/\sqrt{3} & 1/\sqrt{2} \\ 1/\sqrt{3} & -1/\sqrt{2} \\ 1/\sqrt{3} & 0 \end{bmatrix} \text{ and } M = UU^T$$

Calculate the entries of M and show that M is a projection matrix.

20. Consider

$$P = \frac{1}{3} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

Show that P is a projection matrix.

Describe the space that P projects onto.

21. Show that the fitted values are in the column space of X. In other words, show that the fitted values can be written as a linear combination of the columns of X.

22. Let $\boldsymbol{b} = (b_0, b_1, b_2)$ and $\widehat{\boldsymbol{B}}$ be the least squares estimator for $\boldsymbol{b}$. Further, let $M = \sigma^2 (X^T X)^{-1}$, and M_{ij} be the (i, j) entry of M.

What is the expected value of $\begin{bmatrix} 0 & 1 & -1 \end{bmatrix} \widehat{\boldsymbol{B}}$?

What is the variance of $\begin{bmatrix} 0 & 1 & -1 \end{bmatrix} \widehat{\boldsymbol{B}}$?

23. Let

$$\boldsymbol{Y} = X\boldsymbol{b} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 I)$$

Recall that if $\boldsymbol{Z} \sim N(\boldsymbol{\mu}, \Sigma)$, then $\boldsymbol{a} + M\boldsymbol{Z} \sim N(\boldsymbol{a} + M\boldsymbol{\mu}, M\Sigma M^T)$.

What is the distribution of $\boldsymbol{Y}$?

What is the distribution of $X^T \boldsymbol{Y}$?

What is the distribution of $\widehat{\boldsymbol{B}} = (X^T X)^{-1} X^T \boldsymbol{Y}$?

Does $\boldsymbol{Y}^T \boldsymbol{Y}$ follow a normal distribution? Why or why not?

24. The following is from a longitudinal study measuring heights, weights, etc. of a group of girls. Height2 = height at age 2, Height9 = height at age 9, Height18 = height at age 18, and LegCirc9 = leg circumference at age 9.

Below is output from a regression of Height18 on Height2, Height9, and LegCirc9

```
Call: lm(formula = Height18 ~ Height2 + Height9 + LegCirc9 )

                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          35.6880    11.2669   3.167  0.00233
Height2               0.2945     0.1827   1.612  0.11163
Height9               0.9016     0.1195   7.547 1.71e-10
LegCirc9             -0.5986     0.2089  -2.865  0.00559

Residual standard error: 3.404 on 66 degrees of freedom
Multiple R-squared:  0.6997,Adjusted R-squared:  0.6861
F-statistic: 51.27 on 3 and 66 DF,  p-value: < 2.2e-16
```

(a) How many individual girls were in this dataset?

(b) The model fit to the height data was

$$Y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + \varepsilon_i, \quad \varepsilon_i \overset{ind}{\sim} N(0, \sigma^2),$$

where the covariates are listed in the same order as in the output above. What is the estimate of σ?

(c) Given an interpretation for the result of the t-test for the Height2 regression coefficient. Why does this make sense for height data?

(d) Can we conclude that height at age 2 is not important for predicting height at age 18?

(e) What is the residual sum of squares for this model?

(f) What is syy for this dataset, and what is the standard deviation of the response?

25. Show that $X^T \widehat{\boldsymbol{e}} = \mathbf{0}$, that is, the observed residual vector is orthogonal to every column of the design matrix.

26. Let $T = Z/\sqrt{W/m}$. T has a t distribution with m degrees of freedom if what 3 things are true?

27. Suppose we have a hypothesis for the multiple linear model that can be written as

$$H_0 : \boldsymbol{c}^T \boldsymbol{b} = a$$

where $\boldsymbol{c}$ is a $p+1$ by 1 vector, a is a known (hypothesized) scalar, and, as usual, $\boldsymbol{b}$ is the vector of regression coefficients. For example, the hypothesis $H_0 : b_2 - b_1 = 0$ can be written in this form.

(a) Write down the t statistic for this test in terms of $\widehat{\boldsymbol{b}}$, X, and $\widehat{\sigma}^2$.

(b) Show that the random version of the t statistic has a T distribution.

28. Here is a dataset with results from the pinewood derby:

i	racer	lane	heat	time
1	Mary	1	1	3.123
2	Suzy	2	1	3.147
3	June	3	1	3.201
4	Mary	3	2	3.133
5	Suzy	1	2	3.168
6	June	2	2	3.192
7	Mary	2	3	3.118
8	Suzy	3	3	3.146
9	June	1	3	3.225

(a) Write down the design matrix for the one-factor model with **racer** as the factor.

(b) Write down the design matrix for the one-factor model with **lane** as the factor.

(c) Write down the design matrix for the one-factor model with **heat** as the factor.

29. Suppose we have the factor model:

$$Y_i = b_0 + b_{j(i)} + \varepsilon_i \quad \varepsilon_i \overset{ind}{\sim} N(0, \sigma^2)$$

where $j(i)$ is the factor level of the ith observation. Our dataset has the following design matrix:

$$X = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

(a) How many levels of the factor are in our dataset?

(b) What is the rank of the design matrix?

(c) Compute $X^T X$.

(d) State which of the following are estimable functions. For those that are, give a linear combination of observations whose expected value is the estimable function. $b_0 + b_1$, $b_1 + b_2$, $b_2 - b_3$, $b_1 - 0.5(b_2 + b_3)$, $2b_3 - b_1 - b_2$

(e) Suppose we pick a solution to the normal equations for which $\widehat{b}_0 = 0$. Then how do we interpret $\widehat{b}_2$?

(f) Suppose we pick a solution to the normal equations for which $\widehat{b}_1 = 0$. Then how do we interpret $\widehat{b}_2$?

30. Suppose that we have six dogs. Dogs 1 and 2 are greyhounds, dogs 3 and 4 are whippets, and dogs 5 and 6 are Italian greyhounds. We model their weights as:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 I_6).$$

Dog 3 is a whippet that weights 26.0 pounds and is the cutest dog you'll ever meet.

(a) Write these quantities in terms of the notation above:

$E(Y_1) =$

$E(Y_3) =$

$E(Y_5) =$

(b) Give interpretations in words of the following quantities. If no interpretation exists, write "no interpretation". Hint: use your answers from the previous part.

b_0

$b_0 + b_1$

$b_3 - b_2$

(c) Using mathematical notation, write down the hypothesis that all three breeds of the dogs have the same expected weight.

Should we use a t test or an F test? Give the degrees of freedom for the test.

(d) Using mathematical notation, write down the hypothesis that we expect a whippet to weigh 15 pounds more than an Italian greyhound.

Should we use a t test or an F test? Give the degrees of freedom for the test.

31. The SAT data contains average SAT scores from each state. Each state was grouped into 1 of 9 regions: ENC, ESC, MA, MTN, NE, PAC, SA, WNC, WSC. Consider the following model for the average SAT math score from the ith state:

$$Y_i = b_0 + b_{j(i)} + \varepsilon_i, \quad \varepsilon_i \overset{ind}{\sim} N(0, \sigma^2),$$

where $j(i)$ is the region number (1-9) of the ith state. Below is R output from fitting the model:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  551.000      8.093  68.088  < 2e-16 ***
regionESC      1.750     12.139   0.144 0.886059
regionMA     -52.333     13.215  -3.960 0.000284 ***
```

```
regionMTN    -11.500     10.316  -1.115 0.271285
regionNE     -49.167     10.957  -4.487 5.52e-05 ***
regionPAC    -36.200     11.445  -3.163 0.002899 **
regionSA     -61.333     10.093  -6.077 3.08e-07 ***
regionWNC     29.857     10.596   2.818 0.007339 **
regionWSC    -11.750     12.139  -0.968 0.338599

Residual standard error: 18.1 on 42 degrees of freedom
Multiple R-squared:  0.7733,Adjusted R-squared:  0.7302
F-statistic: 17.91 on 8 and 42 DF,  p-value: 2.832e-11
```

(a) Which region's coefficient did R set to zero (the reference region)?

(b) What is the estimated expected SAT score in the reference region?

(c) What is the estimated expected SAT score in the NE region?

(d) Which region has the highest estimated expected SAT score, and what is the estimate?

(e) What is the t statistic for testing whether MA has the same expected SAT score as the reference region?

(f) How many degrees of freedom in the t statistic from the previous part?

(g) In terms of b_j's, write down the hypothesis tested by the F test in the output.

(h) In the F test, what is the rss for the full model?

(i) What is the rss for the reduced model? You can use knowledge of the F statistic or the R^2 statistic to solve this. Try it both ways and make sure you get the same answer.

(j) Think carefully: how many "states" are in this dataset?

32. Consider the following model for 3pm temperatures:

$$Y_i = b_0 + b_{j(i)} + c_1 x_i + \varepsilon_i, \quad \varepsilon_i \overset{ind}{\sim} N(0, \sigma^2),$$

where $j(i)$ is the month associated with the ith observation, and x_i is the 7am temp (atobstemp). Below is output from fitting the model to data from months 1 through 3.

```
                  Estimate Std. Error t value Pr(>|t|)
b0 - (Intercept) 15.06816    1.62505   9.272   <2e-16
b2 - mon02        2.90206    1.64191   1.767   0.0789
b3 - mon03        3.36943    1.56601   2.152   0.0328
c1 - atobstemp    1.00450    0.04983  14.471   <2e-16

Residual standard error: 8.704 on 173 degrees of freedom
Multiple R-squared:  0.5712,Adjusted R-squared:  0.5639
```

```
F-statistic: 78.14 on 3 and 176 DF,  p-value = 0.0459
```

(a) Is this an additive model or an interaction model?

(b) Give interpretations for all 4 regression coefficients.

(c) Draw a plot by hand with the three fitted regression lines.

(d) It goes without saying that the month 3 temperatures (March) are higher on average than the month 1 temperatures (January). This model says something slightly different. Explain what this model says and why it is also plausible.

(e) I altered some numbers in the R output. See how many inconsistencies you can find, and explain why.

(f) Write down the reduced model for the F-test that appears in the R output.

33. Below is an incomplete table of estimated expected values for a two factor additive model for levels $j = 1, 2, 3$ and $k = 1, 2, 3$. Complete the table with numbers and write out what estimates R would produce for b_0, b_1, b_2, b_3, c_1, c_2, and c_3.

$j \backslash k$	1	2	3
1		1	
2	0	7	8
3		5	

34. In our housing example, we have a dataset with sale prices for houses in the St. Louis area. Let y_i be the sale price of house i, let x_i be its size in square feet, and let $j(i)$ be a factor variable for the zip code. Each zip code is given a label 1 through J.

We model the data as:

$$Y_i = b_0 + b_{j(i)} + (c_0 + c_{j(i)})x_i + \varepsilon_i, \quad \varepsilon_i \overset{ind}{\sim} N(0, \sigma^2)$$

If we apply R's convention that $b_1 = 0$ and $c_1 = 0$,

(a) Explain in words what this model assumes about the relationship between expected sale price and zip code and size.

(b) What is the interpretation of $b_0 + b_j$?

(c) What is the interpretation of $c_0 + c_j$?

(d) What is the interpretation of $c_3 - c_2$?

(e) Under R's default constraint that $b_1 = 0$ and $c_1 = 0$, give interpretations for the following parameters:

b_0:

c_0:

b_2:

c_3:

(f) In terms of the parameters, what is the expected sale price for a 2000 square foot house in the fourth zip code?

35. In a memory experiment, subjects of different ages were given different strategies for remembering words from a list. The number of words memorized by subject i is y_i. We fit the following model to the data:

$$Y_i = b_0 + b_{j(i)} + c_{k(i)} + (bc)_{j(i),k(i)} + \varepsilon_i, \quad \varepsilon_i \overset{ind}{\sim} N(0, \sigma^2),$$

where $j(i)$ is the age group of observation i (Older, Younger), and $k(i)$ is the memory process of observation i (Adjective, Counting, Imagery, Intentional, Rhyming). Below is output from fitting the model.

```
                                      Estimate Std. Error t value Pr(>|t|)
b0 - (Intercept)                       11.0000     0.8959  12.279  < 2e-16
b2 - AgeYounger                         3.8000     1.2669   2.999  0.00350
c2 - ProcessCounting                   -4.0000     1.2669  -3.157  0.00217
c3 - ProcessImagery                     2.4000     1.2669   1.894  0.06139
c4 - ProcessIntentional                 1.0000     1.2669   0.789  0.43201
c5 - ProcessRhyming                    -4.1000     1.2669  -3.236  0.00170
(bc)22 - AgeYounger:ProcessCounting    -4.3000     1.7917  -2.400  0.01846
(bc)23 - AgeYounger:ProcessImagery      0.4000     1.7917   0.223  0.82385
(bc)24 - AgeYounger:ProcessIntentional  3.5000     1.7917   1.953  0.05387
(bc)25 - AgeYounger:ProcessRhyming     -3.1000     1.7917  -1.730  0.08702
----------------------------------------------------------------
Residual standard error: 2.833 on 90 degrees of freedom
Multiple R-squared:  0.7293,Adjusted R-squared:  0.7022
F-statistic: 26.93 on 9 and 90 DF,  p-value: < 2.2e-16
```

(a) In 20 words or fewer, give an interpertation of the estimate 11.000 in the first row.

(b) Find the p-value 0.00217. Explain in words what ages and processes are being compared in that test.

(c) What is the estimated expected number of words memorized by Older subjects using the Rhyming process?

(d) What is the estimated expected number of words memorized by Younger subjects using the Imagery process?

(e) The $(bc)_{22}$ coefficient is significant. What is the interpretation of this coefficient, and what do we learn about memory strategies from this significant result?

(f) Using information supplied here, write a formula for the missing F statistic. Use only numbers in your formula.

```
Analysis of Variance Table
------------------------------------------------
Model 1: Words ~ Age + Process
Model 2: Words ~ Age * Process
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1     94 912.6
2     90 722.3  4     190.3 ?????? 0.0002793 ***
```

36. There is a weather station in Ithaca that measures hourly temperatures. Suppose you would like to build a model for predicting the 3pm temperature (`temp3p`) from the 7am temperature (`temp7a`) and the day of year (`doy`). We expect a sinusoidal pattern in `doy` for a full year of data, but we'll analyze data from day 121 to 269 of the year, which we can approximate with a quadratic. Here is the full quadratic model in `temp7a` and `doy`:

```
Call:
lm(formula = temp3p ~ temp7a + doy + I(doy^2) )

Residuals:
    Min      1Q  Median      3Q     Max
-10.953  -2.557   0.263   2.821  12.587

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.240e+00  3.020e+00  -2.729  0.00646 **
temp7a       4.771e-01  2.952e-02  16.160  < 2e-16 ***
doy          2.322e-01  3.405e-02   6.818 1.57e-11 ***
I(doy^2)    -5.415e-04  8.635e-05  -6.271 5.27e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.889 on 1028 degrees of freedom
Multiple R-squared:  0.396,Adjusted R-squared:  0.3942
F-statistic: 224.6 on 3 and 1028 DF,  p-value: < 2.2e-16
```

(a) Write out the statistical model that was assumed, defining all of your notation.

(b) The estimated quadratic coefficient on day of year is negative. Why does this make sense?

(c) How do you interpret the intercept? Should we trust our estimate of the intercept?

(d) What is our estimate of the day of the year with the maximum 3pm temperature?

(e) Here is a model that adds an interaction between 7am temperature and day of year. Write out the statistical model that was assumed, defining all of your notation.

```
Call:
lm(formula = temp3p ~ temp7a * doy + I(doy^2) )

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.159e+01  3.119e+00  -3.715 0.000214 ***
temp7a       9.064e-01  1.137e-01   7.969 4.24e-15 ***
doy          2.402e-01  3.388e-02   7.089 2.51e-12 ***
I(doy^2)    -4.971e-04  8.651e-05  -5.746 1.20e-08 ***
temp7a:doy  -2.171e-03  5.557e-04  -3.906 9.98e-05 ***

Residual standard error: 3.863 on 1027 degrees of freedom
Multiple R-squared:  0.4048,Adjusted R-squared:  0.4025
F-statistic: 174.6 on 4 and 1027 DF,  p-value: < 2.2e-16
```

(f) What are the slopes with respect to 7am temperature on days 121 and 269? Are these numbers practically different from the slope in the first model?

(g) What are the intercepts with respect to 7am temperature on days 121 and 169?

(h) Without doing any calculations, can you find the p-value for the F test that compares these two models?

37. For the following model:

$$Y_i = b_0 + b_1 x_i + z_i \varepsilon_i, \quad \varepsilon_i \overset{ind}{\sim} N(0, \sigma^2)$$

Write down the loglikelihood function, takes its derivative with respect to σ^2, and derive the maximum likelihood estimate for σ^2, written in terms of b_0, b_1, the x_i's, and the z_i's.

38. We have a drug trial with four subjects. Subjects 1 and 2 are siblings, and subjects 3 and 4 are siblings. The first set of siblings is unrelated to the second set. Each subject is assigned either a placebo or a drug treatment, and $j(i) = 1$ if subject i got the placebo, and $j(i) = 2$ if subject i got the drug. We model the responses as

$$\begin{aligned} &Y_i = b_0 + b_{j(i)} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \\ &\text{Cov}(\varepsilon_i, \varepsilon_j) = \sigma^2/2 \text{ if subjects } i \text{ and } j \text{ are siblings} \\ &\text{Cov}(\varepsilon_i, \varepsilon_j) = 0 \text{ if subjects } i \text{ and } j \text{ are not siblings} \end{aligned}$$

(a) Write out the covariance matrix $\Sigma = \text{Cov}(\boldsymbol{\varepsilon})$.

(b) Calculate Σ^{-1}. (Hint, look up "block diagonal matrix")

(c) Calculate $\det(\Sigma)$ (your answer should depend on σ^2).

(d) Suppose that subjects 1 and 3 got the placebo, and subjects 2 and 4 got the drug. Write out the design matrix, dropping the column for b_1.

(e) Calculate the generalized least squares estimates for b_0 and b_2. Your answer should depend on the responses y_1, y_2, y_3, y_4. Check your answers by asking yourself if they "make sense" in terms of how you interpret b_0 and b_2, and how you would expect to estimate those quantities.

(f) Calculate the covariance matrix for $(\widehat{B}_0, \widehat{B}_2)$.

39. Repeat (d)-(f) from the previous problem, except assume that subjects 1 and 2 got the placebo, and subjects 3 and 4 got the drug. Which experiment is better, and why?

40. What are two good reasons for making the effects from a factor random, as opposed to fixed?

41. Suppose we have the following dataset:

i	y_i	x_i	$j(i)$
1	4.3	2	1
2	6.2	4	1
3	5.9	3	2
4	3.1	2	2

and we model the responses y_i as

$$
\begin{aligned}
&Y_i = b_0 + b_1 x_i + C_{j(i)} + D_{j(i)} x_i + \varepsilon_i \\
&C_j \overset{ind}{\sim} N(0, \sigma_1^2), \quad D_j \overset{ind}{\sim} N(0, \sigma_2^2), \quad \varepsilon_i \overset{ind}{\sim} N(0, \sigma_3^3)
\end{aligned}
$$

What is the mean vector and covariance matrix for (Y_1, Y_2, Y_3, Y_4)?