Homework 1 Solutions

Eugene Katsevich

August 25, 2021

1 Instructions

Setup. Pull the latest version of this assignment from Github and set your working directory to `stat-961-fall-2021/homework/homework-1`. Consult the getting started guide if you need to brush up on `R`, `LaTeX`, or `Git`.

Collaboration. The collaboration policy is as stated on the Syllabus:

> "Students are permitted to work together on homework assignments, but solutions must be written up and submitted individually. Students must disclose any sources of assistance they received; furthermore, they are prohibited from verbatim copying from any source and from consulting solutions to problems that may be available online and/or from past iterations of the course."

In accordance with this policy,

Please list anyone you discussed this homework with:

Please list what external references you consulted (e.g. articles, books, or websites):

Writeup. Use this document as a starting point for your writeup, adding your solutions between `\begin{sol}` and `\end{sol}`. See the preparing reports guide for guidance on compilation, creation of figures and tables, and presentation quality. Show all the code you wrote to produce your numerical results, and include complete derivations typeset in LaTeX for the mathematical questions.

Programming. The `tidyverse` paradigm for data manipulation (`dplyr`) and plotting (`ggplot2`) are strongly encouraged, but points will not be deducted for using base `R`.

```
library(tidyverse)
```

Grading. Each sub-part of each problem will be worth 3 points: 0 points for no solution or completely wrong solution; 1 point for some progress; 2 points for a mostly correct solution; 3 points for a complete and correct solution modulo small flaws. The presentation quality of the solution for each problem (as exemplified by the guidelines in Section 3 of the preparing reports guide) will be evaluated out of an additional 3 points.

Submission. Compile your writeup to PDF and submit to Gradescope.

Problem 1. Change of basis. (Adapted from Agresti Ex. 1.17)

Let $\boldsymbol{X}$ and $\boldsymbol{X}'$ be full-rank $n \times p$ model matrices.

(a) Show that $C(\boldsymbol{X}) = C(\boldsymbol{X}')$ if and only if $\boldsymbol{X}' = \boldsymbol{X}\boldsymbol{A}$ for some nonsingular $p \times p$ matrix $\boldsymbol{A}$. In plain language, express what the operation $\boldsymbol{X} \mapsto \boldsymbol{X}\boldsymbol{A}$ does to the columns of $\boldsymbol{X}$ (one sentence is sufficient).

(b) Let $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}'$ be the least squares solutions obtained from regressing a response vector $\boldsymbol{y}$ on $\boldsymbol{X}$ and $\boldsymbol{X}' \equiv \boldsymbol{X}\boldsymbol{A}$, respectively, where $\boldsymbol{A}$ is a nonsingular $p \times p$ matrix. What is the relationship between $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}'$ (express the latter in terms of the former)? Justify your answer.

(c) Consider the linear model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon; \quad \epsilon \sim (0, \sigma^2), \tag{1}$$

so that $\boldsymbol{X} = [\boldsymbol{1}, \boldsymbol{x}_{*1}, \boldsymbol{x}_{*2}]$ for columns $\boldsymbol{x}_{*j} \equiv (x_{1j}, \ldots, x_{nj})^T$, $j \in \{1, 2\}$. Sometimes it is useful to center the predictors by subtracting their means:

$$\boldsymbol{x}'_{*j} \equiv \boldsymbol{x}_{*j} - \bar{x}_j \boldsymbol{1}; \quad \bar{x}_j \equiv \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \quad j \in \{1, 2\}.$$

Defining $\boldsymbol{X}' \equiv [\boldsymbol{1}, \boldsymbol{x}'_{*1}, \boldsymbol{x}'_{*2}]$, find the matrix $\boldsymbol{A}$ such that $\boldsymbol{X}' = \boldsymbol{X}\boldsymbol{A}$ ($\boldsymbol{A}$ may itself be expressed in terms of $\boldsymbol{X}$). Express the coefficient estimates from the centered regression ($\widehat{\beta}'_0$, $\widehat{\beta}'_1$, $\widehat{\beta}'_2$) in terms of those from the original regression ($\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{\beta}_2$).

(d) Let $w \in \{a, b, c\}$ be a categorical variable with three levels. Define $x_1 \equiv \mathbb{1}(w = b)$ and $x_2 \equiv \mathbb{1}(w = c)$, and consider the linear regression (1). This corresponds to regressing y on the categorical variable w, with baseline category a. Sometimes a different baseline category may make more sense, e.g. category b. In this case, we would define $x'_1 \equiv \mathbb{1}(w = a)$ and $x'_2 \equiv \mathbb{1}(w = c)$. Defining $\boldsymbol{X}$ and $\boldsymbol{X}'$ as in part (c), find the matrix $\boldsymbol{A}$ such that $\boldsymbol{X}' = \boldsymbol{X}\boldsymbol{A}$. Express the coefficient estimates from the transformed regression ($\widehat{\beta}'_0$, $\widehat{\beta}'_1$, $\widehat{\beta}'_2$) in terms of those from the original regression ($\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{\beta}_2$). What are the interpretations of the original and transformed coefficients, and why do the relationships between these coefficients derived above make sense in terms of these interpretations?

Solution 1.

(a) Suppose $C(\boldsymbol{X}) = C(\boldsymbol{X}')$. This means that for each $j \in \{1, \ldots, p\}$, we have $\boldsymbol{x}'_{*j} \in C(\boldsymbol{X}') = C(\boldsymbol{X})$, so by definition of $C(\boldsymbol{X})$ there exists a vector $\boldsymbol{a}_{*j} \in \mathbb{R}^p$ such that $\boldsymbol{x}'_{*j} = \boldsymbol{X}\boldsymbol{a}_{*j}$. This implies that $\boldsymbol{X}' = \boldsymbol{X}\boldsymbol{A}$, where $\boldsymbol{A} \equiv [\boldsymbol{a}_{*1}, \ldots, \boldsymbol{a}_{*p}]$. The matrix $\boldsymbol{A}$ must be nonsingular because $p = \text{rank}(\boldsymbol{X}') = \text{rank}(\boldsymbol{X}\boldsymbol{A}) \leq \text{rank}(\boldsymbol{A})$, so $\text{rank}(\boldsymbol{A}) = p$.

Conversely, suppose $\boldsymbol{X}' = \boldsymbol{X}\boldsymbol{A}$ for some nonsingular $p \times p$ matrix $\boldsymbol{A}$. Then, we have

$$C(\boldsymbol{X}') \equiv \{\boldsymbol{X}'\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^p\} = \{\boldsymbol{X}\boldsymbol{A}\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^p\} \subseteq C(\boldsymbol{X}).$$

On the other hand, $\boldsymbol{X} = \boldsymbol{X}'\boldsymbol{A}^{-1}$, so by the same reasoning we also have $C(\boldsymbol{X}) \subseteq C(\boldsymbol{X}')$. Therefore, $C(\boldsymbol{X}) = C(\boldsymbol{X}')$, which completes the proof.

The columns of $\boldsymbol{X}\boldsymbol{A}$ are linear combinations of the columns of $\boldsymbol{X}$.

(b) The fitted vectors $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{X}'\widehat{\boldsymbol{\beta}}'$ are the projections of the response vector $\boldsymbol{y}$ onto $C(\boldsymbol{X})$ and $C(\boldsymbol{X}')$, respectively. Since these two model speces are the same, this implies that the fitted vectors are the same as well, i.e. $\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'\widehat{\boldsymbol{\beta}}'$. Therefore, $\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'\widehat{\boldsymbol{\beta}}' = \boldsymbol{X}\boldsymbol{A}\widehat{\boldsymbol{\beta}}'$. Since $\boldsymbol{X}$ is full rank, it follows that $\widehat{\boldsymbol{\beta}} = \boldsymbol{A}\widehat{\boldsymbol{\beta}}'$, so we conclude that $\widehat{\boldsymbol{\beta}}' = \boldsymbol{A}^{-1}\widehat{\boldsymbol{\beta}}$.

(c) We have

$$\boldsymbol{X}' \equiv (\mathbf{1}, \boldsymbol{x}'_{*1}, \boldsymbol{x}'_{*2}) = (\mathbf{1}, \boldsymbol{x}_{*1} - \bar{x}_1\mathbf{1}, \boldsymbol{x}_{*2} - \bar{x}_2\mathbf{1}) = (\mathbf{1}, \boldsymbol{x}_{*1}, \boldsymbol{x}_{*2}) \begin{pmatrix} 1 & -\bar{x}_1 & -\bar{x}_2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \equiv \boldsymbol{X}\boldsymbol{A},$$

where

$$\boldsymbol{A} \equiv \begin{pmatrix} 1 & -\bar{x}_1 & -\bar{x}_2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

The inverse of this matrix is easily seen to be

$$\boldsymbol{A}^{-1} = \begin{pmatrix} 1 & \bar{x}_1 & \bar{x}_2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Therefore, by part (b), we have

$$\begin{pmatrix} \widehat{\beta}'_0 \\ \widehat{\beta}'_1 \\ \widehat{\beta}'_2 \end{pmatrix} = \boldsymbol{A}^{-1}\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 1 & \bar{x}_1 & \bar{x}_2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \widehat{\beta}_0 + \bar{x}_1\widehat{\beta}_1 + \bar{x}_2\widehat{\beta}_2 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix}. \tag{2}$$

(d) We have

$$\begin{aligned} (1, x'_1, x'_2) &= (1, \mathbb{1}(w = a), \mathbb{1}(w = c)) \\ &= (1, 1 - \mathbb{1}(w = b) - \mathbb{1}(w = c), \mathbb{1}(w = c)) \\ &= (1, 1 - x_1 - x_2, x_2) \\ &= (1, x_1, x_2) \begin{pmatrix} 1 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & -1 & 1 \end{pmatrix}. \end{aligned}$$

Therefore, $\boldsymbol{X}' = \boldsymbol{X}\boldsymbol{A}$ for

$$\boldsymbol{A} \equiv \begin{pmatrix} 1 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & -1 & 1 \end{pmatrix}.$$

The inverse of this matrix is easily seen to be

$$\boldsymbol{A}^{-1} = \boldsymbol{A} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & -1 & 1 \end{pmatrix}.$$

Therefore, by part (b), we have

$$\begin{pmatrix} \widehat{\beta}'_0 \\ \widehat{\beta}'_1 \\ \widehat{\beta}'_2 \end{pmatrix} = \boldsymbol{A}^{-1}\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \widehat{\beta}_0 + \widehat{\beta}_1 \\ -\widehat{\beta}_1 \\ -\widehat{\beta}_1 + \widehat{\beta}_2 \end{pmatrix}.$$

For the original regression, the coefficient β_0 is the mean response value in category a, the coefficient β_1 is the difference in mean response values between categories b and a, and the coefficient β_2 is the difference in mean response values between categories c and a. For the transformed regression, the coefficient β_0' is the mean response value in category b, the coefficient β_1' is the difference in mean response values between categories a and b, and the coefficient β_2 is the difference in mean response values between categories c and b. The relationships among the corresponding fitted coefficients derived above are consistent with these interpretations; for example, the estimated mean response value in category b ($\widehat{\beta}_0'$) is the sum of the estimated mean response value in category a ($\widehat{\beta}_0$) and the estimated difference in mean response values between categories b and a ($\widehat{\beta}_1$).

Problem 2. Predictor correlation. (Adapted from Agresti Ex. 2.9)

Consider the linear regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon; \quad \epsilon \sim (0, \sigma^2),$$

with observed predictor vectors denoted $\boldsymbol{x}_{*1} \equiv (x_{11}, \ldots, x_{n1})^T$ and $\boldsymbol{x}_{*2} \equiv (x_{12}, \ldots, x_{n2})^T$. (This is the same setup as in Problem 1(c).)

(a) Suppose $\boldsymbol{x}_{*1}$ and $\boldsymbol{x}_{*2}$ have sample correlation $\rho \in (-1, 1)$. In terms of ρ, what is the correlation between the estimates $\widehat{\beta}_1$ and $\widehat{\beta}_2$ (which are random variables due to the randomness in ϵ)?

(b) To build intuition for the preceding result, consider the extreme case when $\boldsymbol{x}_{*1} = \boldsymbol{x}_{*2}$. In this case, $\rho = 1$ and the regression is not identifiable. For a fixed parameter vector $(\beta_0^0, \beta_1^0, \beta_2^0)$, write down the set $\mathcal{S}$ of parameter vectors $(\beta_0, \beta_1, \beta_2)$ giving the same value of $\mathbb{E}[\boldsymbol{y}]$ as $(\beta_0, \beta_1, \beta_2) = (\beta_0^0, \beta_1^0, \beta_2^0)$. In what sense does the result in part (a) reflect the relationship between β_1 and β_2 for $(\beta_0, \beta_1, \beta_2) \in \mathcal{S}$? (Ignore the fact that the case $\rho = 1$ is not covered in part (a).)

(c) Suppose $z_1, z_2 \overset{\text{i.i.d.}}{\sim} N(0, 1)$, and $x_1 \equiv z_1 + 0.5 z_2$ and $x_2 \equiv z_1 - 0.5 z_2$. What is the correlation between the random variables x_1 and x_2? Suppose the predictors in each row $\{(x_{i1}, x_{i2})\}_{i=1}^n$ are a sample from this joint distribution. Roughly what do we expect to be the sample correlation between $\boldsymbol{x}_{*1}$ and $\boldsymbol{x}_{*2}$? Fixing $\boldsymbol{x}_{*1}$ and $\boldsymbol{x}_{*2}$ at their realizations, roughly what do we expect to be the correlation between $\widehat{\beta}_1$ and $\widehat{\beta}_2$?

(d) To check the conclusions in part (b), run a numerical simulation with $n = 100$, $\sigma^2 = 1$, $(\beta_0, \beta_1, \beta_2) = (0, 1, 2)$, and $\epsilon \sim N(0, \sigma^2)$. Sample one realization of $\boldsymbol{x}_{*1}$ and $\boldsymbol{x}_{*2}$, generate 250 realizations of the response $\boldsymbol{y}$, and for each realization calculate least squares estimates $\widehat{\boldsymbol{\beta}}$. Summarize the results of your simulation by creating scatter plots of $\boldsymbol{x}_{*2}$ versus $\boldsymbol{x}_{*1}$ and $\widehat{\beta}_2$ versus $\widehat{\beta}_1$, with the title of each plot containing the sample correlations of the data it displays. On the scatter plot of $\widehat{\beta}_2$ versus $\widehat{\beta}_1$, indicate the theoretical expected value of $(\widehat{\beta}_1, \widehat{\beta}_2)$ with a red point. Display these two scatter plots side by side using `cowplot::plot_grid`. Do the sample correlations match what you predicted in part (c)?

Solution 2.

(a) Recall that $\widehat{\boldsymbol{\beta}} \sim (\beta, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$. From conclusion (2) in the solution of Problem 1, we may assume without loss of generality that the predictors $\boldsymbol{x}_{*1}$ and $\boldsymbol{x}_{*2}$ are centered. Due to the resulting orthogonality between $\mathbf{1}$ and $(\boldsymbol{x}_{*1}, \boldsymbol{x}_{*2})$, it follows that

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} n & 0 & 0 \\ 0 & \boldsymbol{x}_{*1}^T\boldsymbol{x}_{*1} & \boldsymbol{x}_{*1}^T\boldsymbol{x}_{*2} \\ 0 & \boldsymbol{x}_{*2}^T\boldsymbol{x}_{*1} & \boldsymbol{x}_{*2}^T\boldsymbol{x}_{*2} \end{pmatrix} = \begin{pmatrix} n & 0 & 0 \\ 0 & \|x_{*1}\|^2 & \|x_{*1}\|\|x_{*2}\|\rho \\ 0 & \|x_{*1}\|\|x_{*2}\|\rho & \|x_{*2}\|^2 \end{pmatrix}$$

Therefore,

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{pmatrix} n^{-1} & 0 & 0 \\ 0 & \frac{\|\boldsymbol{x}_{*1}\|}{\sqrt{1-\rho^2}\|\boldsymbol{x}_{*2}\|} & \frac{-\rho}{\sqrt{1-\rho^2}} \\ 0 & \frac{-\rho}{\sqrt{1-\rho^2}} & \frac{\|\boldsymbol{x}_{*2}\|}{\sqrt{1-\rho^2}\|\boldsymbol{x}_{*1}\|} \end{pmatrix}.$$

Therefore,

$$\begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} \sim \left(\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \sigma^2 \begin{pmatrix} \frac{\|\boldsymbol{x}_{*1}\|}{\sqrt{1-\rho^2}\|\boldsymbol{x}_{*2}\|} & \frac{-\rho}{\sqrt{1-\rho^2}} \\ \frac{-\rho}{\sqrt{1-\rho^2}} & \frac{\|\boldsymbol{x}_{*2}\|}{\sqrt{1-\rho^2}\|\boldsymbol{x}_{*1}\|} \end{pmatrix} \right).$$

From this expression, we can read off that

$$\text{Cor}[\widehat{\beta}_1, \widehat{\beta}_2] = \frac{\text{Cov}[\widehat{\beta}_1, \widehat{\beta}_2]}{\text{sd}[\widehat{\beta}_1]\text{sd}[\widehat{\beta}_2]} = \frac{\frac{-\rho}{\sqrt{1-\rho^2}}}{\left(\frac{\|\boldsymbol{x}_{*1}\|}{\sqrt{1-\rho^2}\|\boldsymbol{x}_{*2}\|}\right)^{1/2}\left(\frac{\|\boldsymbol{x}_{*2}\|}{\sqrt{1-\rho^2}\|\boldsymbol{x}_{*1}\|}\right)^{1/2}} = -\rho.$$

Therefore, if the sample correlation between $\boldsymbol{x}_{*1}$ and $\boldsymbol{x}_{*2}$ is ρ, then the correlation between $\widehat{\beta}_1$ and $\widehat{\beta}_2$ is $-\rho$.

(b) We have $\mathbb{E}[\boldsymbol{y}] = \beta_0\mathbf{1} + \beta_1\boldsymbol{x}_{*1} + \beta_2\boldsymbol{x}_{*2} = \beta_0\mathbf{1} + (\beta_1+\beta_2)\boldsymbol{x}_{*1}$. Therefore,

$$\mathcal{S} = \{(\beta_0, \beta_1, \beta_2) : \beta_0 = \beta_0^0, \beta_1 + \beta_2 = \beta_1^0 + \beta_2^0\} = \{(\beta_0^0, \beta_1^0 - C, \beta_2^0 + C) : C \in \mathbb{R}\}.$$

Since $\rho = 1$, part (a) suggests that $\text{Cor}[\widehat{\beta}_1, \widehat{\beta}_2] = -1$. This is consistent with the fact that $(\beta_1, \beta_2) = (\beta_1^0 - C, \beta_2^0 + C)$ have a perfect inverse relationship.

(c) We have

$$\begin{aligned}\text{Cor}[x_1, x_2] &= \frac{\text{Cov}[x_1, x_2]}{(\text{Var}[x_1])^{1/2}(\text{Var}[x_2])^{1/2}} \\ &= \frac{\mathbb{E}[(z_1 + 0.5z_2)(z_1 - 0.5z_2)]}{(\text{Var}[z_1 + 0.5z_2])^{1/2}(\text{Var}[z_1 - 0.5z_2])^{1/2}} = \frac{0.75}{1.25^{1/2}1.25^{1/2}} = 0.6.\end{aligned}$$

We therefore expect the sample correlation between $\boldsymbol{x}_{*1}$ and $\boldsymbol{x}_{*2}$ to be approximately 0.6. By part (a), we expect the correlation between $\widehat{\beta}_1$ and $\widehat{\beta}_2$ to be approximately -0.6.

(d) First we run the numerical simulation:

```
# simulation parameters
n = 100      # samples size
beta_0 = 0   # true regression coefficients
beta_1 = 1
beta_2 = 2
sigma = 1    # noise standard deviation
reps = 250   # number of random realizations

# set seed for reproducibility
set.seed(961)

# generate predictors
z1 = rnorm(n)
z2 = rnorm(n)
x1 = z1 + 0.5*z2
x2 = z1 - 0.5*z2

# run simulation
beta_1_hat = numeric(reps)        # initialize vectors for fitted coefficients
beta_2_hat = numeric(reps)
for(rep in 1:reps){               # loop over random realizations
  eps = rnorm(n)                  # generate data for this realization
  data = tibble(x1,
```

```
                x2,
                y = beta_0 + beta_1*x1 + beta_2*x2 + eps)
  lm_fit = lm(y ~ ., data = data) # run linear regression
  coefs = coef(lm_fit)            # extract coefficients
  beta_1_hat[rep] = coefs["x1"]   # record first coefficient
  beta_2_hat[rep] = coefs["x2"]   # record second coefficient
}
```

Next we summarize the results of the simulation:

```
# create scatter plot of x2 versus x1
p1 = tibble(x1, x2) %>%
  ggplot(aes(x = x1, y = x2)) +
  geom_point() +
  labs(x = quote(x[1]),
       y = quote(x[2]),
       title = sprintf("Correlation = %0.2f", cor(x1, x2))) +
  theme_bw() + theme(plot.title = element_text(hjust = 0.5))

# create scatter plot of beta_2_hat versus beta_1_hat
p2 = tibble(beta_1_hat, beta_2_hat) %>%
  ggplot(aes(x = beta_1_hat, y = beta_2_hat)) +
  geom_point() +
  geom_point(x = beta_1, y = beta_2, colour = "red") +
  labs(x = quote(widehat(beta)[1]),
       y = quote(widehat(beta)[2]),
       title = sprintf("Correlation = %0.2f", cor(beta_1_hat, beta_2_hat)))+
  theme_bw() + theme(plot.title = element_text(hjust = 0.5))

# concatenate the plots side by side
p = cowplot::plot_grid(p1, p2, labels = "auto", align = "hv")

# save final plot
ggsave(filename = "figures/predictor-correlation.png", plot = p, device = "png",
       width = 5, height = 2.75)
```

Figure 1 displays the results of the numerical simulation. The correlations of 0.61 and -0.53 in these scatter plots are roughly what was predicted in part (c) ($\rho \approx 0.6$ and $-\rho \approx -0.6$, respectively).

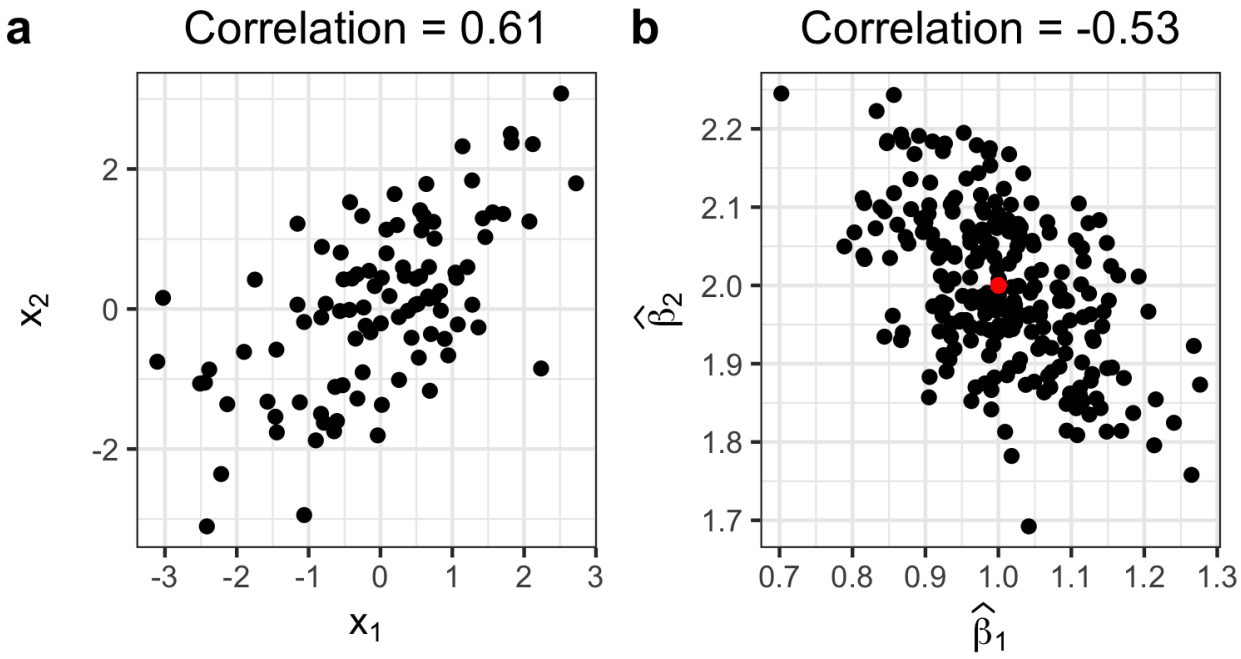

Figure 1: Simulation results: Positively correlated predictors (a) lead to negatively correlated coefficient estimates (b).

Problem 3. Data analysis: Anorexia treatment. (Adapted from Agresti Ex. 1.24)

For 72 young girls suffering from anorexia, the `Anorexia.dat` file under `stat-961-fall-2021/data` shows their weights before and after an experimental period:

```
anorexia_data = read_tsv("../../data/Anorexia.dat", col_types = "ifdd")
print(anorexia_data, n = 5)

## # A tibble: 72 x 4
##    subj therapy before after
##   <int> <fct>    <dbl> <dbl>
## 1     1 b         80.5  82.2
## 2     2 b         84.9  85.6
## 3     3 b         81.5  81.4
## 4     4 b         82.6  81.9
## 5     5 b         79.9  76.4
## # ... with 67 more rows
```

The girls were randomly assigned to receive one of three therapies during this period. A control group received the standard therapy, which was compared to family therapy and cognitive behavioral therapy. The goal of the study is to compare the effectiveness of the therapies in increasing the girls' weights.

(a) Prepare the data by (1) removing the `subj` variable, (2) re-coding the factor levels of `therapy` as `behavioral`, `family`, and `control`, (3) renaming `before` and `after` to `weight_before` and `weight_after`, respectively, and (4) adding a variable called `weight_gain` defined as the difference of `weight_after` and `weight_before`. Print the resulting tibble.

(b) Explore the data by (1) making box plots of `weight_gain` as a function of `therapy`, (2) making a scatter plot of `weight_gain` against `weight_before`, coloring points based on `therapy` and (3) creating a table displaying, for each `therapy` group, the mean weight gain, maximum weight gain, and fraction gained weight. Based on these summaries: What therapy appears overall the most successful and why? How effective does the standard therapy appear to be? What is the greatest weight gain observed in this study? Which girls tended to gain most weight, based on their weight before therapy? Why might this be the case?

(c) Run a linear regression of `weight_gain` on `therapy` and print the regression summary (print in `R`, without using `kable`). Identify the base category chosen by `R` and discuss the interpretations of the fitted coefficients. It makes more sense to choose `control` as the base category. Recode the factor levels so that `control` is the first (and therefore will be chosen as the base category), rerun the linear regression, and print the summary again. Do the relationships among the fitted coefficients in these two regressions match what was found in Problem 1d?

(d) Directly compute the between-groups, within-groups, and corrected total sums of squares (without appealing to the `aov` function or equivalent) and verify that the first two add up to the third. What is the ratio of the between-groups sum of squares and the corrected total sum of squares? What is the interpretation of this quantity, and what quantity in the regression summaries printed in part (c) is it equivalent to? Finally, compute an unbiased estimate for the error variance in the regression.

Solution 3.

(a)
```
anorexia_data = anorexia_data %>%
  select(-subj) %>%                    # remove subj variable
  mutate(therapy =
           factor(therapy,             # recode therapy variable
                  labels = c("behavioral",
                             "family",
                             "control"))) %>%
  rename(weight_before = before,       # rename before and after variables
         weight_after = after) %>%
  mutate(weight_gain =                 # create weight_gain variable
           weight_after - weight_before)

print(anorexia_data, n = 5)

## # A tibble: 72 x 4
##   therapy    weight_before weight_after weight_gain
##   <fct>              <dbl>        <dbl>       <dbl>
## 1 behavioral          80.5         82.2       1.70
## 2 behavioral          84.9         85.6       0.700
## 3 behavioral          81.5         81.4      -0.100
## 4 behavioral          82.6         81.9      -0.700
## 5 behavioral          79.9         76.4      -3.5
## # ... with 67 more rows
```

(b)
```
# (1) box plots of weight_gain versus therapy
p1 = anorexia_data %>%
  ggplot(aes(x = therapy, y = weight_gain, fill = therapy)) +
  geom_boxplot() +
  labs(x = "Therapy",
       y = "Weight gain (lbs)") +
  theme_bw() + theme(legend.position = "none")

# (2) scatter plot of weight_gain versus weight_before
p2 =  anorexia_data %>%
  ggplot(aes(x = weight_before,
             y = weight_gain,
             colour = therapy)) +
  geom_point() +
  labs(x = "Weight before therapy (lbs)",
       y = "Weight gain (lbs)",
       colour = "Therapy") +
  theme_bw()

# concatenate the first two plots side by side
p = cowplot::plot_grid(p1, p2, labels = "auto", align = "h",
                       rel_widths = c(1,2))
ggsave(filename = "figures/summary-anorexia-fig.png", plot = p, device = "png",
```

```
       width = 6.5, height = 3)

# (3) table with summary statistics for each therapy group
summary_table = anorexia_data %>%
  group_by(therapy) %>%
  summarise(`Mean weight gain` = mean(weight_gain),
            `Max weight gain` = max(weight_gain),
            `Frac. gained weight` = mean(weight_gain > 0)) %>%
  rename(Therapy = therapy)

# save table
summary_table %>%
  kableExtra::kable(format = "latex", row.names = NA,
                    booktabs = TRUE, digits = 2) %>%
  kableExtra::save_kable("figures/summary-anorexia-tab.png")
```

Figure 2 and Table 1 explore how `weight_gain` depends on `therapy` and `weight_before`. Overall, the family therapy appears most successful; it has the greatest mean weight gain (7.26 pounds), maximum weight gain (21.5 pounds), and fraction gaining weight (76%) among the three therapies. The standard therapy does not appear effective at all, resulting in a loss of weight on average. The greatest weight gain observed in the study was 21.5 pounds, and based on Figure 2 it appears that girls who weighed the less before therapy tended to gain more weight. This might be the case because these girls were further from a normal weight to begin with and thus had more weight to gain.

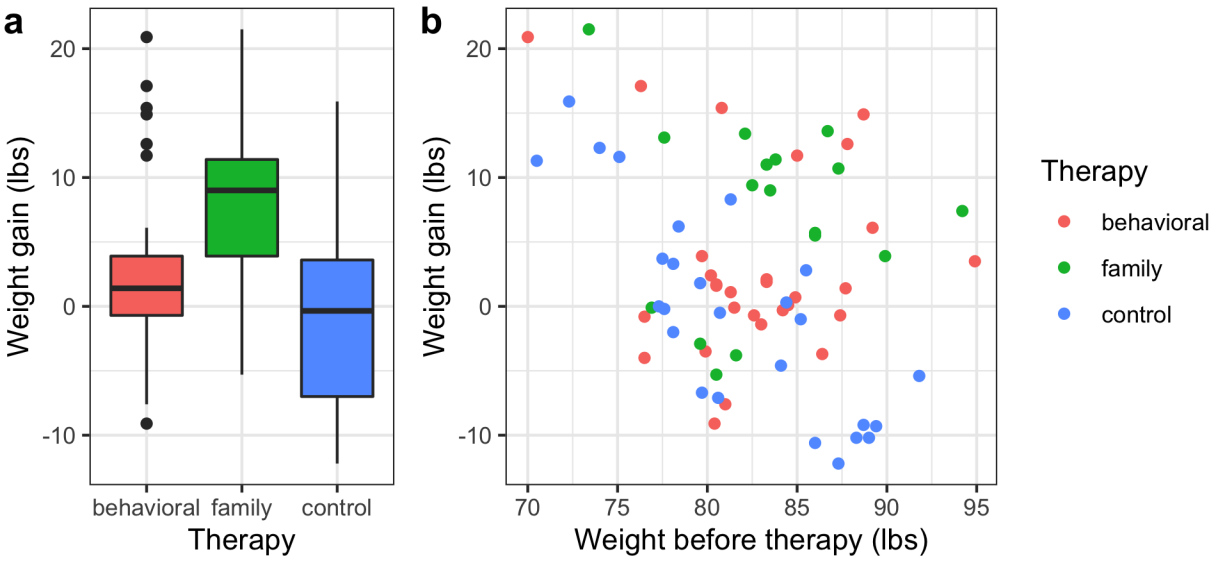

Figure 2: Summary plots for the anorexia data: Weight gain by therapy (a) and relationship between weight gain and weight before therapy (b). It appears that family therapy was most successful, and girls who weighed the less before therapy tended to gain more weight.

Therapy	Mean weight gain	Max weight gain	Frac. gained weight
behavioral	3.01	20.9	0.62
family	7.26	21.5	0.76
control	-0.45	15.9	0.42

Table 1: Summaries of weight gain by therapy group. Family therapy was most successful across all three metrics.

(c)
```
lm_fit = lm(weight_gain ~ therapy, data = anorexia_data)
summary(lm_fit)

##
## Call:
## lm(formula = weight_gain ~ therapy, data = anorexia_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.565  -4.543  -1.007   3.846  17.893
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.007      1.398   2.151   0.0350 *
## therapyfamily     4.258      2.300   1.852   0.0684 .
## therapycontrol   -3.457      2.033  -1.700   0.0936 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.528 on 69 degrees of freedom
## Multiple R-squared:  0.1358,Adjusted R-squared:  0.1108
## F-statistic: 5.422 on 2 and 69 DF,  p-value: 0.006499
```

We see that the base category chosen by `R` is `behavioral`. The intercept is therefore the mean weight gain for the `behavioral` category, and the other two coefficients are the increases in mean weight gain in the `family` and `control` categories, relative to the `behavioral` category. These categories have about 4.3 lbs more and 3.5 lbs less weight gain than the `behavioral` category, respectively.

```
anorexia_data = anorexia_data %>%
  mutate(therapy =
           factor(therapy, levels = c("control", "behavioral", "family")))
lm_fit = lm(weight_gain ~ therapy, data = anorexia_data)
summary(lm_fit)

##
## Call:
## lm(formula = weight_gain ~ therapy, data = anorexia_data)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.565  -4.543  -1.007   3.846  17.893
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -0.450      1.476  -0.305   0.7614
## therapybehavioral    3.457      2.033   1.700   0.0936 .
## therapyfamily        7.715      2.348   3.285   0.0016 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.528 on 69 degrees of freedom
## Multiple R-squared:  0.1358,Adjusted R-squared:  0.1108
## F-statistic: 5.422 on 2 and 69 DF,  p-value: 0.006499
```

Re-running the regression using `control` as a base category, we get the coefficient estimates above. These are indeed related to the coefficients in the first regression by the rules derived in Problem 1(d). For example, the coefficient of `family` in the second regression (7.7, the improvement of this treatment over `control`) is the difference between the coefficient of `family` (4.3) and the coefficient of `control` (-3.5) in the first regression.

(d)
```
# compute sum of squared quantities
SS_table = anorexia_data %>%
  group_by(therapy) %>%
  mutate(weight_gain_therapy = mean(weight_gain)) %>%
  ungroup() %>%
  mutate(weight_gain_mean = mean(weight_gain)) %>%
  summarise(`Between-groups` = sum((weight_gain_therapy-weight_gain_mean)^2),
            `Within-groups` = sum((weight_gain - weight_gain_therapy)^2),
            `Total` = sum((weight_gain - weight_gain_mean)^2))

# save table
SS_table %>%
  kableExtra::kable(format = "latex", row.names = NA,
                    booktabs = TRUE, digits = 2) %>%
  kableExtra::save_kable("figures/ss-anorexia.png")
```

Between-groups	Within-groups	Total
614.64	3910.74	4525.39

Table 2: Sum-of-squared quantities for the anorexia data.

Table 2 displays the three sum-of-squared quantities, and indeed it is the case that the sum of the first two equals the third: 614.64 + 3910.74 = 4525.39 (up to rounding). The ratio of the between-groups and total sums of squares is 614.64/4525.39 = 0.136; this quantity is the

fraction of the variance in weight gain explained by the therapy variable. This is the same as the R^2 quantity from the linear regressions run in part (c), as can be seen by inspecting the regression summaries.