Homework 1

Samuel Rosenberg

Due September 13, 2021 at 11:59pm

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 $\boldsymbol{X'} = \boldsymbol{XA}$ for some nonsingular $p \times p$ matrix $\boldsymbol{A}$. In plain language, express what the operation $\boldsymbol{X} \mapsto \boldsymbol{XA}$ 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 $\boldsymbol{X'} \equiv \boldsymbol{XA}$, 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 $\boldsymbol{X'} = \boldsymbol{XA}$ ($\boldsymbol{A}$ 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 $\boldsymbol{X} \equiv [\boldsymbol{1}, \boldsymbol{x}_{*1}, \boldsymbol{x}_{*2}]$ and $\boldsymbol{X'} \equiv [\boldsymbol{1}, \boldsymbol{x}'_{*1}, \boldsymbol{x}'_{*2}]$, find the matrix $\boldsymbol{A}$ such that $\boldsymbol{X'} = \boldsymbol{XA}$. 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) Because each column space is full rank (rank p), the columns of each model matrix form a basis; we write each of these bases as follows: $\{\boldsymbol{x_1}, \ldots, \boldsymbol{x_p}\}$ and $\{\boldsymbol{x'_1}, \ldots, \boldsymbol{x'_p}\}$ respectively.
$\Rightarrow$: Suppose that $C(\boldsymbol{X}) = C(\boldsymbol{X'})$. Since $C(\boldsymbol{X}) = C(\boldsymbol{X'})$, $\boldsymbol{x'_j} \in C(\boldsymbol{X})$. This means that $\boldsymbol{x'_j} = \sum_{k=1}^{p} a_{kj} \boldsymbol{x_k} = \sum_{k=1}^{p} a_{kj} (x_{1k}, \ldots, x_{nk})^T$ for some a_{kj}. So, $x'_{ij} = \sum_{k=1}^{p} a_{kj} x_{ik} = (x_{i1}, \ldots, x_{ip})(a_{1j}, \ldots, a_{pj})^T$. This is equivalent to saying that $\boldsymbol{X'} = \boldsymbol{XA}$, for the $p \times p$ matrix $\boldsymbol{A}$ with entry a_{ij} in the i-th row and j-th column.
Suppose for contradiction that $\boldsymbol{A}$ is singular, then $\text{rank}(\boldsymbol{A}) < p$. So, there is some column $\boldsymbol{a_\ell} = (a_{1\ell}, \ldots, a_{p\ell})^T$ that can be written as a linear combination of the other $\boldsymbol{a_j}$; i.e. $\boldsymbol{a_\ell} = \beta_1 \boldsymbol{a_1} + \cdots + \beta_{\ell-1} \boldsymbol{a_{\ell-1}} + \beta_{\ell+1} \boldsymbol{a_{\ell+1}} + \cdots + \beta_p \boldsymbol{a_p}$ and $a_{i\ell} = \beta_1 a_{i1} + \cdots + \beta_{\ell-1} a_{i\ \ell-1} + \beta_{\ell+1} a_{i\ \ell+1} + \cdots + \beta_p a_{ip}$.

Then

$$\begin{aligned}\boldsymbol{x'_\ell} &= \sum_{k=1}^{p} a_{k\ell}\boldsymbol{x_k} \\ &= a_{1\ell}\boldsymbol{x_1} + \sum_{k=2}^{p} a_{k\ell}\boldsymbol{x_k} \\ &= (\sum_{k=2}^{p} \beta_k a_{1k}\boldsymbol{x_1}) + \sum_{k=2}^{p} a_{k\ell}\boldsymbol{x_k} \\ &= \sum_{k=2}^{p} (\beta_k a_{1k} + a_{k\ell})\boldsymbol{x_k}.\end{aligned}$$

But this means that $\boldsymbol{x'_\ell}$ is a non-unique linear combination of the $\boldsymbol{x_k}$, so $\text{rank}(C(\boldsymbol{X})) < p$. This is a contradiction since we assumed $\boldsymbol{X}$ was full rank, so $\boldsymbol{A}$ must be nonsingular.
$\Leftarrow$: Now suppose that $\boldsymbol{X'} = \boldsymbol{XA}$ for some nonsingular $p \times p$ matrix $\boldsymbol{A}$. Then each entry x'_{ij} of $\boldsymbol{X'}$ has the following form: $x'_{ij} = (x_{i1}, \ldots, x_{ip})(a_{1j}, \ldots, a_{pj})^T = \sum_{k=1}^{p} a_{kj}x_{ik}$. Consequently, the j-th column of $\boldsymbol{X'}$ is $\boldsymbol{x'_j} = \sum_{k=1}^{p} a_{kj}\boldsymbol{x_k}$. Thus, $\text{span}\{\boldsymbol{x'_1}, \ldots, \boldsymbol{x'_p}\} = C(\boldsymbol{X'}) \subseteq C(\boldsymbol{X})$. But $\dim(C(\boldsymbol{X'})) = \dim(C(\boldsymbol{X}))$, so we must have that $C(\boldsymbol{X}) = C(\boldsymbol{X'})$.
Thus, $C(\boldsymbol{X}) = C(\boldsymbol{X}')$ if and only if $\boldsymbol{X'} = \boldsymbol{XA}$ for some nonsingular $p \times p$ matrix $\boldsymbol{A}$.
The operation $\boldsymbol{X} \mapsto \boldsymbol{XA}$ converts the columns of $\boldsymbol{X}$ to use the same units as $\boldsymbol{X'}$ (i.e. it is a change of bases).

(b) Recall that the least square estimators for the linear regression of $\boldsymbol{y}$ on $\boldsymbol{X}$ and $\boldsymbol{X'}$ are $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ and $\widehat{\boldsymbol{\beta}}' = ((\boldsymbol{X'})^T\boldsymbol{X'})^{-1}(\boldsymbol{X'})^T\boldsymbol{y}$ respectively (e.g. by equation 2.3 in Agresti). Taking $\boldsymbol{X'} \equiv \boldsymbol{XA}$ and using properties of the matrix transpose and matrix inverse (since $\boldsymbol{X}$, $\boldsymbol{X'}$, $\boldsymbol{A}$ are full rank), we have that

$$\begin{aligned}\widehat{\boldsymbol{\beta}}' &= ((\boldsymbol{X'})^T\boldsymbol{X'})^{-1}(\boldsymbol{X'})^T\boldsymbol{y} \\ &= ((\boldsymbol{XA})^T\boldsymbol{XA})^{-1}(\boldsymbol{XA})^T\boldsymbol{y} \\ &= (\boldsymbol{A}^T\boldsymbol{X}^T\boldsymbol{XA})^{-1}\boldsymbol{A}^T\boldsymbol{X}^T\boldsymbol{y} \\ &= (\boldsymbol{X^TXA})^{-1}(\boldsymbol{A}^T)^{-1}\boldsymbol{A}^T\boldsymbol{X}^T\boldsymbol{y} \\ &= (\boldsymbol{X^TXA})^{-1}\boldsymbol{X}^T\boldsymbol{y} \\ &= \boldsymbol{A}^{-1}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \\ &= \boldsymbol{A}^{-1}\widehat{\boldsymbol{\beta}}.\end{aligned}$$

(c) We take

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

Then

$$\begin{aligned}\boldsymbol{XA} &= \begin{pmatrix}1 & x_{11} & x_{12}\\ \vdots & \vdots & \vdots\\ 1 & x_{n1} & x_{n2}\end{pmatrix}\begin{pmatrix}1 & -\overline{x}_1 & -\overline{x}_2\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix}\\ &= \begin{pmatrix}1 & x_{11}-\overline{x}_1 & x_{12}-\overline{x}_2\\ \vdots & \vdots & \vdots\\ 1 & x_{n1}-\overline{x}_1 & x_{n2}-\overline{x}_2\end{pmatrix}\\ &= \boldsymbol{X'}\end{aligned}$$

as desired.
Note that

$$\boldsymbol{A}\begin{pmatrix}1 & \overline{x}_1 & \overline{x}_2\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix} = \begin{pmatrix}1 & -\overline{x}_1 & -\overline{x}_2\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix}\begin{pmatrix}1 & \overline{x}_1 & \overline{x}_2\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix} = \boldsymbol{I}_3,$$

so

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

As we saw in (b), $\widehat{\boldsymbol{\beta}}' = \boldsymbol{A}^{-1}\widehat{\boldsymbol{\beta}}$, so

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

(d) We take

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

Note that $x_2 = x_2'$ implies that $x_{i2} = x_{i2}'$ for all i. Also, x_1' is 1 if $w = a$ (and $x_1 = x_2 = 0$) and 0 otherwise, so $x_1' = 1 - x_1 - x_2$ and $x_{i1}' = 1 - x_{i1} - x_{i2}$.

Then

$$\begin{aligned}
\boldsymbol{XA} &= \begin{pmatrix} 1 & x_{11} & x_{12} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \\
&= \begin{pmatrix} 1 & 1 - x_{11} - x_{12} & x_{12} \\ \vdots & \vdots & \vdots \\ 1 & 1 - x_{n1} - x_{n2} & x_{n2} \end{pmatrix} \\
&= \begin{pmatrix} 1 & x'_{11} & x'_{12} \\ \vdots & \vdots & \vdots \\ 1 & x'_{n1} & x'_{n2} \end{pmatrix} \\
&= \boldsymbol{X'}
\end{aligned}$$

as desired.
Note that $\boldsymbol{A}^2 = \boldsymbol{I}_3$, so $\boldsymbol{A} = \boldsymbol{A}^{-1}$. As we saw in (b), $\widehat{\boldsymbol{\beta}}' = \boldsymbol{A}^{-1}\widehat{\boldsymbol{\beta}}$, so

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

We interpret the coefficients as follows: $\widehat{\beta}_0$ is the mean for the subgroup with $w = a$, while $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are the differences in the means of the subgroups where $w = b$ and $w = a$ or $w = c$ and $w = a$ respectively. Likewise, $\widehat{\beta}'_1$ is the mean for the subgroup with $w = b$, while $\widehat{\beta}'_1$ and $\widehat{\beta}'_2$ are the differences in the means of the subgroups where $w = a$ and $w = b$ or $w = c$ and $w = a$ respectively.
Intuitively, the relationships between the coefficients make sense: $\widehat{\beta}'_0 = \widehat{\beta}_0 + \widehat{\beta}_1$ is the mean of the subgroup with $w = a$ plus the difference in the subgroups with $w = b$ and $w = a$; i.e. it is $\widehat{\beta}'_0$, the mean of the subgroup with $w = b$. Likewise, $\widehat{\beta}'_1 = -\widehat{\beta}_1 = \widehat{\beta}_0 - (\widehat{\beta}_0 + \widehat{\beta}_1)$ is the difference between the means of the subgroups where $w = a$ and $w = b$. Finally, $\widehat{\beta}'_2 = -\widehat{\beta}_1 + \widehat{\beta}_2 = (\widehat{\beta}_0 + \widehat{\beta}_2) - (\widehat{\beta}_0 + \widehat{\beta}_1)$ is the difference between the means of the subgroups where $w = c$ and $w = b$.

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) Note that the model

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

is equivalent to the model

$$y = \beta_0' + \beta_1 (x_1 - \overline{x}_1) + \beta_2 (x_2 - \overline{x}_2), \ \epsilon \sim (0, \sigma^2), \ \text{with } \beta_0' = \beta_0 + \beta_1 \overline{x}_1 + \beta_2 \overline{x}_2.$$

So, the coefficients $\widehat{\beta}_1$, $\widehat{\beta}_2$ are unaffected by mean centering. Note that the regression $y = \beta_1(x_1 - \overline{x}_1) + \beta_2(x_2 - \overline{x}_2)$ has the same coefficients of this second model by orthogonality; $\overline{x}_1$ is the projection of x_1 onto 1 and $\overline{x}_2$ is the projection of x_2 onto 1, so $x_1 - \overline{x}_1, x_2 - \overline{x}_2 \perp 1$.

Thus, the covariance matrix $\boldsymbol{\Sigma}$ for the initial model has the following form:

$$\boldsymbol{\Sigma} = \begin{pmatrix} s_{11} & s_{12} & s_{13} \\ s_{21} & & \\ s_{31} & & \boldsymbol{\Sigma}' \end{pmatrix},$$

where $\boldsymbol{\Sigma}'$ is the covariance matrix for $\widehat{\beta}_1$ and $\widehat{\beta}_2$ under the model $y = \beta_1(x_1 - \overline{x}_1) + \beta_2(x_2 - \overline{x}_2)$.

Recall that for an identifiable regression with model matrix $\boldsymbol{X}$, the covariance matrix is given by $\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Note that our model matrix for the third (reduced) regression is

$$\boldsymbol{X} = \begin{pmatrix} x_{11} - \overline{x}_1 & x_{12} - \overline{x}_2 \\ \vdots & \vdots \\ x_{n1} - \overline{x}_1 & x_{n2} - \overline{x}_2 \end{pmatrix}.$$

Then

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} ||\boldsymbol{x}_{*1} - \overline{x}_1\mathbf{1}||^2 & (\boldsymbol{x}_{*1} - \overline{x}_1\mathbf{1})^T(\boldsymbol{x}_{*2} - \overline{x}_2\mathbf{1}) \\ (\boldsymbol{x}_{*1} - \overline{x}_1\mathbf{1})^T(\boldsymbol{x}_{*2} - \overline{x}_2\mathbf{1}) & ||\boldsymbol{x}_{*2} - \overline{x}_2\mathbf{1}||^2 \end{pmatrix}$$

and

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \frac{1}{\det(\boldsymbol{X}^T\boldsymbol{X})} \begin{pmatrix} ||\boldsymbol{x}_{*2} - \overline{x}_2\mathbf{1}||^2 & -(\boldsymbol{x}_{*1} - \overline{x}_1\mathbf{1})^T(\boldsymbol{x}_{*2} - \overline{x}_2\mathbf{1}) \\ -(\boldsymbol{x}_{*1} - \overline{x}_1\mathbf{1})^T(\boldsymbol{x}_{*2} - \overline{x}_2\mathbf{1}) & ||\boldsymbol{x}_{*1} - \overline{x}_1\mathbf{1}||^2 \end{pmatrix}.$$

We have that

$$\operatorname{cov}(\widehat{\beta}_1, \widehat{\beta}_2) = \sigma^2\boldsymbol{\Sigma}_{23} = \sigma^2\boldsymbol{\Sigma}'_{12} = -\frac{\sigma^2}{\det(\boldsymbol{X}^T\boldsymbol{X})}(\boldsymbol{x}_{*1} - \overline{x}_1\mathbf{1})^T(\boldsymbol{x}_{*2} - \overline{x}_2\mathbf{1}),$$

$$\sigma^2_{\widehat{\beta}_1} = \sigma^2\boldsymbol{\Sigma}_{22} = \sigma^2\boldsymbol{\Sigma}'_{11} = \frac{\sigma^2}{\det(\boldsymbol{X}^T\boldsymbol{X})}||\boldsymbol{x}_{*2} - \overline{x}_2\mathbf{1}||^2,$$

and

$$\sigma^2_{\widehat{\beta}_2} = \sigma^2\boldsymbol{\Sigma}_{33} = \sigma^2\boldsymbol{\Sigma}'_{22} = \frac{\sigma^2}{\det(\boldsymbol{X}^T\boldsymbol{X})}||\boldsymbol{x}_{*1} - \overline{x}_1\mathbf{1}||^2.$$

Define $C := \frac{\sigma^2}{\det(\boldsymbol{X}^T\boldsymbol{X})}$. Then

$$\begin{aligned} \operatorname{corr}(\widehat{\beta}_1, \widehat{\beta}_1) &= \frac{\operatorname{cov}(\widehat{\beta}_1, \widehat{\beta}_2)}{\sigma_{\widehat{\beta}_1}\sigma_{\widehat{\beta}_2}} \\ &= -\frac{C(\boldsymbol{x}_{*1} - \overline{x}_1\mathbf{1})^T(\boldsymbol{x}_{*2} - \overline{x}_2\mathbf{1})}{\sqrt{C}||\boldsymbol{x}_{*2} - \overline{x}_2\mathbf{1}||\sqrt{C}||\boldsymbol{x}_{*1} - \overline{x}_1\mathbf{1}||} \\ &= -\rho. \end{aligned}$$

That is, the correlation between $\widehat{\beta}_1$ and $\widehat{\beta}_2$ is $-\rho$.

(b) Suppose that we have that $(\beta_0^0, \beta_1^0, \beta_2^0)$ is a parameter vector for the case when $\boldsymbol{x}_{*1} = \boldsymbol{x}_{*2}$. Consider $\mathcal{S} := \{(\beta_0, \beta_1, \beta_2) \in \mathbb{R}^3 : \beta_0 = \beta_0^0, \beta_1 + \beta_2 = \beta_1^0 + \beta_2^0\}$. Consider the models where y^0 and y have the same formula as (1c), with $(\beta_0^0, \beta_1^0, \beta_2^0)$ and $(\beta_0, \beta_1, \beta_2) \in \mathcal{S}$ as their respective parameter vectors. Then

$$\mathbb{E}[\boldsymbol{y}^0] = \mathbb{E}[\beta_0^0\mathbf{1} + \beta_1^0\boldsymbol{x}_{*1} + \beta_2^0\boldsymbol{x}_{*2} + \boldsymbol{\epsilon}] = \beta_0^0\mathbf{1} + \beta_1^0\boldsymbol{x}_{*1} + \beta_2^0\boldsymbol{x}_{*2} = \beta_0^0\mathbf{1} + (\beta_1^0 + \beta_2^0)\boldsymbol{x}_{*1},$$

where the last equality holds since $\boldsymbol{x}_{*1} = \boldsymbol{x}_{*2}$. But

$$\beta_0^0\mathbf{1} + (\beta_1^0 + \beta_2^0)\boldsymbol{x}_{*1} = \beta_0\mathbf{1} + (\beta_1 + \beta_2)\boldsymbol{x}_{*1} = \beta_0\mathbf{1} + \beta_1\boldsymbol{x}_{*1} + \beta_2\boldsymbol{x}_{*2} = \mathbb{E}[\boldsymbol{y}],$$

where the first equality holds by construction of $\mathcal{S}$. So, $\mathbb{E}[\boldsymbol{y}^0] = \mathbb{E}[\boldsymbol{y}]$; i.e. any parameter vector in $\mathcal{S}$ will yield the same expectation $\mathbb{E}[\boldsymbol{y}]$ as the original fixed parameter vector.

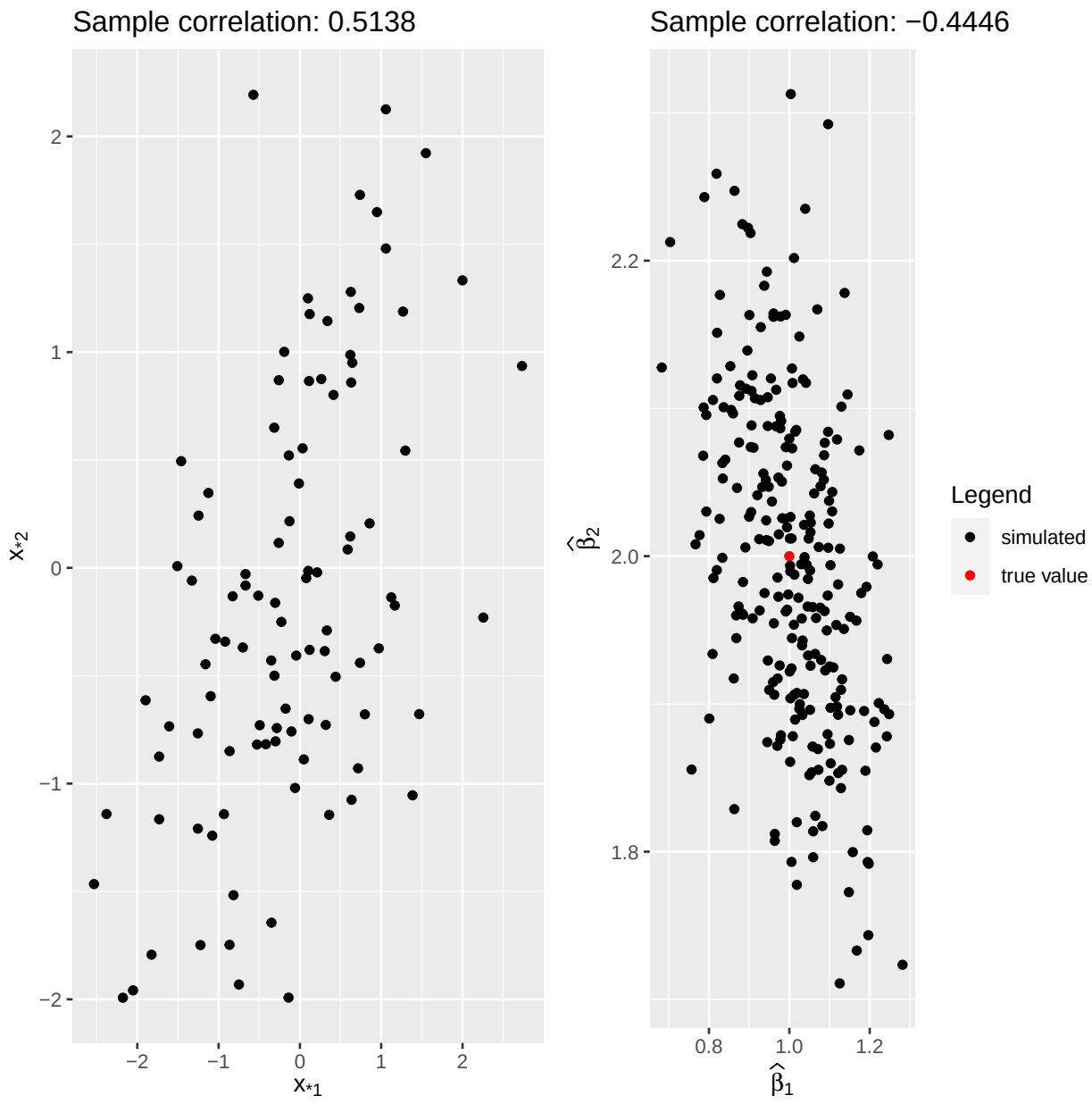
Take $c := \beta_1^0 + \beta_2^0$, which is one of the constraints that appears when defining $\mathcal{S}$ (i.e. $\beta_1 + \beta_2 = c$). When we rearrange this expression, we get $\beta_1 = c - \beta_2$. In other words, perfect correlation of the predictors x_1 and x_2 leads to a constraint of perfect anticorrelation of the coefficients β_1 and β_2, where the slope of the linear relationship between β_1 and β_2 is determined to be the negative of the slope of the linear relationship between x_1 and x_2. This corresponds to $\rho = 1$ implying that $\text{cor}(\hat{\beta}_1, \hat{\beta}_2) = -1$, if the result from (a) could be extended to $\rho = \pm 1$.

(c) Note that $\mathbb{E}[x_1] = \mathbb{E}[z_1 + 0.5z_2] = \mathbb{E}[z_1] + 0.5\mathbb{E}[z_2] = 0$; likewise, $\mathbb{E}[x_2] = 0$. Then $\text{cov}(x_1, x_2) = \mathbb{E}[(x_1 - \mathbb{E}[x_1])(x_2 - \mathbb{E}[x_2])] = \mathbb{E}[x_1 x_2] = \mathbb{E}[(z_1 + 0.5z_2)(z_1 - 0.5z_2)] = \mathbb{E}[z_1^2] - 0.25\mathbb{E}[z_2^2]$. But we have that $z_1^2, z_2^2 \sim \chi_1^2$, so $\mathbb{E}[z_1] = \mathbb{E}[z_2] = 1$ and $\text{cov}(x_1, x_2) = 3/4$. Also, $\sigma_{x1}^2 = \sigma_{x2}^2 = \text{var}(z_1) + 0.25\text{var}(z_2) = 5/4$. So, the correlation between x_1 and x_2 as random variables is $\text{corr}(x_1, x_2) = \frac{3/4}{\sqrt{5/4}^2} = 3/5$.

We expect the sample correlation between $\boldsymbol{x}_{*1}$ and $\boldsymbol{x}_{*2}$ to be roughly equal to the correlation between x_1 and x_2 as random variables; that is, $3/5$. This is because the sample correlation is an asymptotically unbiased estimator of the correlation between random variables. Using our result from (a), we thus expect the correlation between $\widehat{\beta}_1$ and $\widehat{\beta}_2$ to be approximately $-3/5$.

(d)
```r
library(cowplot)
library(latex2exp)

### Set simulation parameters
# Set values of model parameters
beta_0 <- 0
beta_1 <- 1
beta_2 <- 2

n <- 100
sigma <- 1

num_sim <- 250


### Fix one realization of x_{*1}, x_{*2} based on the distribution from (c)
# Generate n standard normal IID values for z_1, z_2
z_1 <- rnorm(n, mean = 0, sd = 1)
z_2 <- rnorm(n, mean = 0, sd = 1)

# Compute x_{*1}, x_{*2}
x_1 <- z_1 + 0.5 * z_2
x_2 <- z_1 - 0.5 * z_2


### Generate num_sim realizations of the response y and calculate \widehat{\beta}
# Vectors to store \widehat{\beta} estimates
beta_1_hat <- c()
beta_2_hat <- c()

# Run loop num_sim times
```

```r
for(i in 1:num_sim){
  # Get \epsilon and calculate y
  epsilon <- rnorm(n, mean = 0, sd = sigma)

  y <- beta_0 + beta_1 * x_1 + beta_2 * x_2 + epsilon

  # Combine data into data frame and run regression
  sim_data <- data.frame(y = y, x_1 = x_1, x_2 = x_2)
  sim_lm <- lm(y ~ x_1 + x_2, data = sim_data)

  # Extract and store \widehat{\beta} estimates
  beta_hat <- sim_lm$coefficients
  beta_1_hat <- c(beta_1_hat, beta_hat[2])
  beta_2_hat <- c(beta_2_hat, beta_hat[3])
}

x_plt <-
  ggplot(data = data.frame(x_1 = x_1, x_2 = x_2),
         mapping = aes(x = x_1, y = x_2)) +
    geom_point() +
    xlab(TeX("$x_{*1}$")) +
    ylab(TeX("$x_{*2}$")) +
    ggtitle(paste("Sample correlation:",
                  round(cor(x = x_1, y = x_2), digits = 4)))

beta_plt <-
  ggplot(data = data.frame(beta_1_hat = c(beta_1_hat, beta_1),
                           beta_2_hat = c(beta_2_hat, beta_2),
                           labels = c(rep("simulated", times = num_sim),
                                      "true value")),
         mapping = aes(x = beta_1_hat,
                       y = beta_2_hat,
                       color = labels)) +
    geom_point() +
    xlab(TeX("$\\widehat{\\beta}_1$")) +
    ylab(TeX("$\\widehat{\\beta}_2$")) +
    ggtitle(paste("Sample correlation:",
                  round(cor(x = beta_1_hat, y = beta_2_hat), digits = 4))) +
    scale_color_manual(name = "Legend", values = c("black", "red"))

plot_grid(plotlist = list(x_plt, beta_plt), nrow = 1, ncol = 2)
```



Note that we have that the sample correlation between $\boldsymbol{x}_{*1}$ and $\boldsymbol{x}_{*2}$ is roughly $3/5 = 0.60$, the correlation between x_1 and x_2 calculated in (c). Likewise, the sample correlation between $\widehat{\beta}_1$ and $\widehat{\beta}_2$ is roughly $-3/5 = -0.60$, which was predicted as the correlation between $\widehat{\beta}_1$ and $\widehat{\beta}_2$ in (c).

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:

```
setwd("C:/Users/Sam/Documents/School/2021-2022/First Semester/Stat 961/stat-961-fall-2021/homew
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 of girls who gained weight (i.e. `weight_gain > 0`). 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 (in the absolute sense), based on their weight before therapy? Why might this be the case?

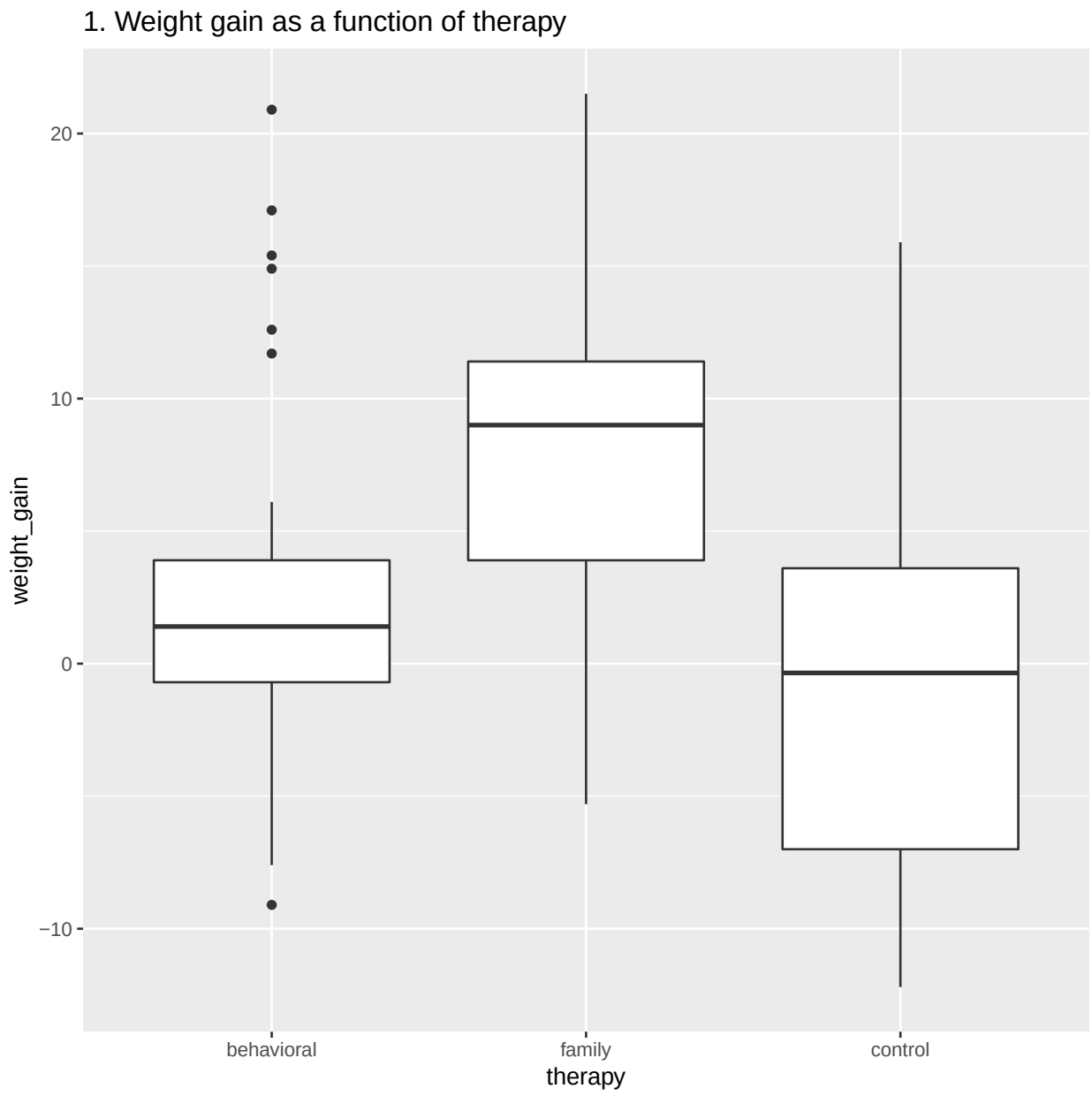
(c) Run a linear regression of `weight_gain` on `therapy` and print the regression summary (print in `R`, without using `kable`). Identify the base category chosen by `R` and discuss the interpretations of the fitted coefficients. It makes more sense to choose `control` as the base category. Recode the 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?

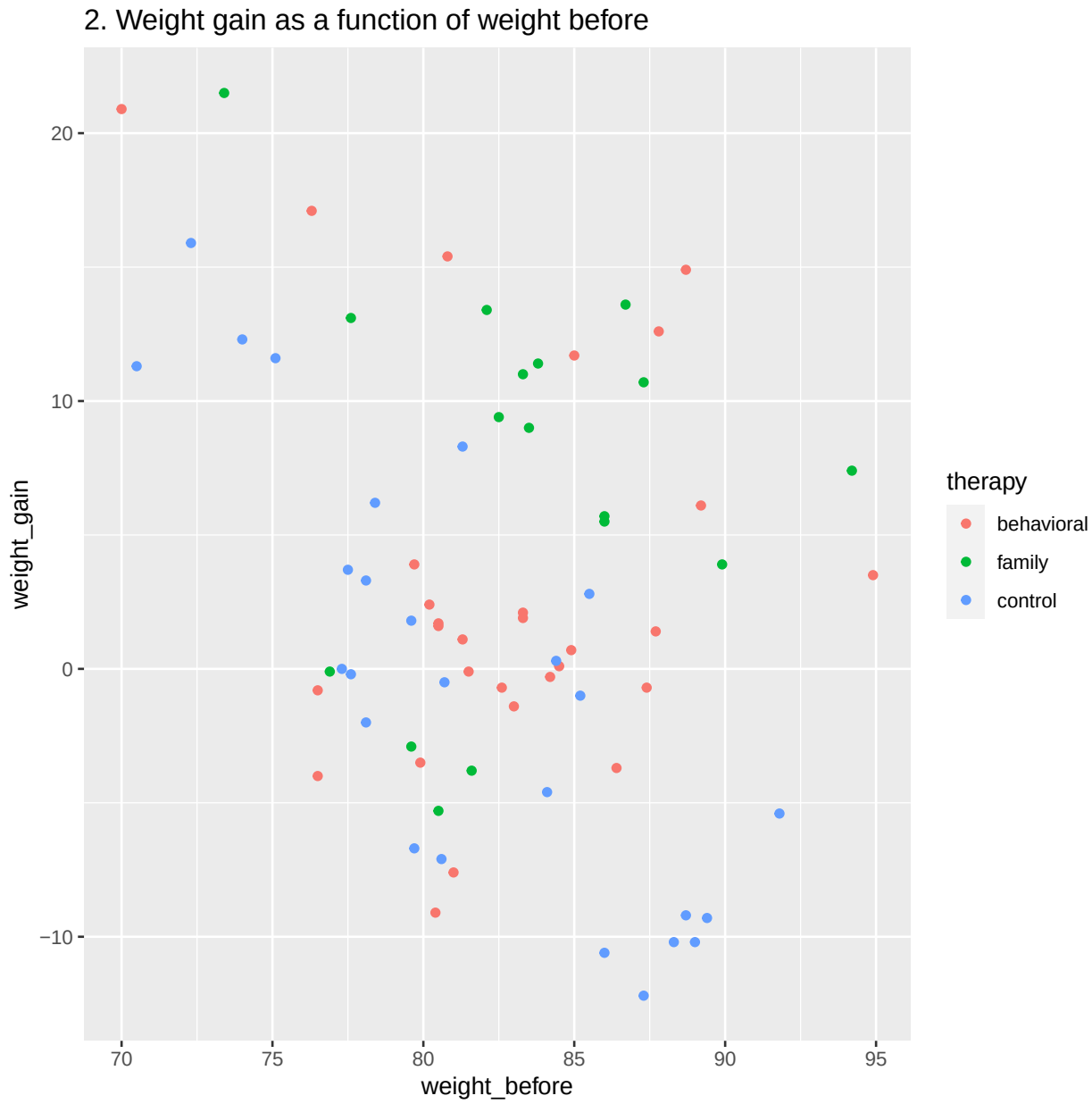
Solution 3.

(a)
```
anorexia_data_processed <-
  anorexia_data %>%
    select(-subj) %>%
    mutate(therapy =
          recode(therapy,
                 b = "behavioral",
                 f = "family",
                 c = "control")
          ) %>%
    rename(weight_before = "before", weight_after = "after") %>%
    mutate(weight_gain = weight_after - weight_before)

print(anorexia_data_processed, 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)
```
anorexia_data_processed %>%
  ggplot() +
  geom_boxplot(mapping = aes(x = therapy, y = weight_gain)) +
  ggtitle("1. Weight gain as a function of therapy")
```




```
anorexia_data_processed %>%
  ggplot() +
  geom_point(mapping = aes(x = weight_before, y = weight_gain, color = therapy)) +
  ggtitle("2. Weight gain as a function of weight before")
```




```
anorexia_data_summary <-
  anorexia_data_processed %>%
    group_by(therapy) %>%
    summarise(mean_weight_gain = mean(weight_gain),
              max_weight_gain = max(weight_gain),
              frac_weight_gain = sum(weight_gain > 0)/length(weight_gain))

print(anorexia_data_summary)
```

```
## # A tibble: 3 x 4
##   therapy    mean_weight_gain max_weight_gain frac_weight_gain
##   <fct>                 <dbl>           <dbl>            <dbl>
## 1 behavioral             3.01            20.9            0.621
```

```
## 2 family                 7.26             21.5            0.765
## 3 control               -0.450            15.9            0.423
```

Note that the family therapy seems the most successful as it has the highest average weight gain, highest max weight gain, and highest proportion of individuals who gained weight as compared to the other two treatments. The standard (control) therapy seems quite ineffective, as less than half of the recipients of this therapy gained weight and the average weight gain for this group is negative (i.e. these participants had a tendency to *lose* weight). The greatest weight gain observed in the study was 21.5 pounds, for an individual in the family therapy group. Girls with lower weights prior to the study tended to have higher weight gains following the study than those who had higher initial weights. This is negative correlation is possibly a sort of regression to the mean, where it is easier for individuals who are far from their "equilibrium" body weight to move toward that weight than away (i.e. an increase in weight for those below and a decrease/smaller increase for those above).

(c)
```
lm1 <- lm(weight_gain ~ therapy, data = anorexia_data_processed)
print(summary(lm1))
```

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

The base category chosen by R was behavioral therapy. The intercept of 3.007 means that we expect the average weight gain of the behavioral therapy to be about 3.007 pounds, in agreement with the 3.01 number from (2). The coefficient of 4.258 for the family therapy indicator means that the mean weight gain of the family therapy group should be 4.258 pounds higher than that of the behavioral therapy group (7.265 pounds, again agreeing with the figure from (2)). Finally, the coefficient of -3.457 pounds for the control therapy indicator means that the weight gain of the control therapy group should be -3.457 pounds lower than that of the behavioral therapy group; i.e. the mean weight gain for the group is -0.450 pounds, in agreement with the figure found in (2).

```
anorexia_data_processed <-
  anorexia_data_processed %>%
```

```
    mutate(therapy = relevel(therapy, ref = "control"))

lm2 <- lm(weight_gain ~ therapy, data = anorexia_data_processed)
print(summary(lm2))

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

Taking $a :=$ behavioral, $b :=$ control, and $c :=$ family, we see that we are in a specific case of (1d) where we have changed the baseline value of a three level categorical variable.

Note that $\widehat{\beta}_0' = -0.450 = 3.007 + (-3.457) = \widehat{\beta}_0 + \widehat{\beta}_1$, $\widehat{\beta}_1' = 3.457 = -(-3.457) = -\widehat{\beta}_1$, and $\widehat{\beta}_2' = 7.715 = -(-3.457) + 4.258 = -\widehat{\beta}_1 + \widehat{\beta}_2$, so the coefficients have the same relationship as predicted by (1d).

(d)
```
y <- anorexia_data_processed$weight_gain
y_bar <- mean(y)
y_bar_by_gp <- list(
              behavioral = lm1$coefficients[1],
              family = lm1$coefficients[1] + lm1$coefficients[2],
              control = lm1$coefficients[1] + lm1$coefficients[3])
y_bar_ci <-
  anorexia_data_processed %>%
    select(therapy) %>%
    summarise(gp_mean = lm1$coefficients[1] +
                ifelse(therapy == "behavioral",
                       0,
                       ifelse(therapy == "family",
                              lm1$coefficients[2],
                              lm1$coefficients[3])))
```

```r
# SST = \sum_{y=1}^n(y_i - \overline{y})^2
sst <- sum((y - y_bar)^2)

# SSB = \sum_{i=1}^n (\overline{y}_{C(i)} - \overline{y})^2
ssb <- sum((y_bar_ci - y_bar)^2)

# SSW = \sum_{i=1}^n (y_i - \overline{y}_{C(i)})^2
ssw <- sum((y - y_bar_ci)^2)

print(paste("SST:", sst))

## [1] "SST: 4525.38611111111"

print(paste("SSB:", ssb))

## [1] "SSB: 614.643666892044"

print(paste("SSW:", ssw))

## [1] "SSW: 3910.74244421907"
```

Note that we indeed have that SST = SSB + SSW. The ratio of SSB to SST is $614.6437/4525.386 = 0.1358$. This corresponds to R^2 in both regression summaries and can be interpreted as saying that about 13.6% of the variation in the data can be captured just by accounting for the type of therapy each individual received.