HOMEWORK 4

Problem 1. Let $\boldsymbol{x}_1 = (1,1,1,1,1,1)'$, $\boldsymbol{x}_2 = (3,-1,4,6,3,3)'$, $\boldsymbol{x}_3 = (7,3,2,0,3,3)'$, $\boldsymbol{x}_4 = (8,4,9,-5,4,4)'$, $\boldsymbol{Y} = (4,36,44,12,16,8)'$, $V = \mathscr{L}(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3,\boldsymbol{x}_4)$. Suppose we wish to test $H_0 : \beta_4 = 0$, $\beta_2 = \beta_3$.

(a) Find two matrix $\boldsymbol{A}$ such that H_0 is equivalent to $\boldsymbol{A}\boldsymbol{\beta} = 0$.
(b) Find $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$, and $\boldsymbol{Z} = \boldsymbol{A}\hat{\boldsymbol{\beta}}$ for one of your choices of $\boldsymbol{A}$.
(c) Define V_0 so that $\boldsymbol{\theta} := \mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{X}] \in V_0$ if and only if $\boldsymbol{A}\boldsymbol{\beta} = 0$. Find $\hat{\boldsymbol{Y}}_0 = \Pi_{V_0}\boldsymbol{Y}$, $\boldsymbol{Y} - \hat{\boldsymbol{Y}}$ and $\hat{\boldsymbol{Y}}_1 = \hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_0$.
(d) Determine $\mathrm{SS}_{\mathrm{Res}} = \|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2$, $\mathrm{SS}_{\mathrm{Res}}(V_0) = \|\boldsymbol{Y} - \hat{\boldsymbol{Y}}_0\|$, and the F-statistic.
(e) Verify that $\|\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_0\|^2 = \boldsymbol{Z}'[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}']^{-1}\boldsymbol{Z}$.
(f) Find $\boldsymbol{a}$ so that $\|\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_0\|^2 = \langle a, \boldsymbol{Y}\rangle^2 / \|\boldsymbol{a}\|^2]$

Solution to Problem 1. We have that

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

Another choice is

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

The model matrix is

$$\boldsymbol{X} = \begin{pmatrix} 1 & 3 & 7 & 8 \\ 1 & -1 & 3 & 4 \\ 1 & 4 & 2 & 9 \\ 1 & 6 & 0 & -5 \\ 1 & 3 & 3 & 4 \\ 1 & 3 & 3 & 4 \end{pmatrix}.$$

To find the projection onto V, we find

$$\Pi_V = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = \begin{pmatrix} \frac{11}{12} & -\frac{1}{12} & -\frac{1}{12} & -\frac{1}{12} & \frac{1}{6} & \frac{1}{6} \\ -\frac{1}{12} & \frac{11}{12} & -\frac{1}{12} & -\frac{1}{12} & \frac{1}{6} & \frac{1}{6} \\ -\frac{1}{12} & -\frac{1}{12} & \frac{11}{12} & -\frac{1}{12} & \frac{1}{6} & \frac{1}{6} \\ -\frac{1}{12} & -\frac{1}{12} & -\frac{1}{12} & \frac{11}{12} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix}.$$

We find that the fitted values are then

$$\hat{\boldsymbol{Y}} = \begin{pmatrix} \frac{11}{12} & -\frac{1}{12} & -\frac{1}{12} & -\frac{1}{12} & \frac{1}{6} & \frac{1}{6} \\ -\frac{1}{12} & \frac{11}{12} & -\frac{1}{12} & -\frac{1}{12} & \frac{1}{6} & \frac{1}{6} \\ -\frac{1}{12} & -\frac{1}{12} & \frac{11}{12} & -\frac{1}{12} & \frac{1}{6} & \frac{1}{6} \\ -\frac{1}{12} & -\frac{1}{12} & -\frac{1}{12} & \frac{11}{12} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix} \begin{pmatrix} 4 \\ 36 \\ 44 \\ 12 \\ 16 \\ 8 \end{pmatrix} = \begin{pmatrix} 0 \\ 32 \\ 40 \\ 8 \\ 20 \\ 20 \end{pmatrix}.$$

The fitted values and residuals when projecting onto V is

$$\hat{\boldsymbol{Y}} = \begin{pmatrix} 0 \\ 32 \\ 40 \\ 8 \\ 20 \\ 20 \end{pmatrix}, \quad \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \begin{pmatrix} 4 \\ 4 \\ 4 \\ 4 \\ -4 \\ -12 \end{pmatrix}.$$

Also,

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} 41 \\ -3 \\ -8 \\ 3 \end{pmatrix}, \quad \boldsymbol{A}\hat{\boldsymbol{\beta}} = \begin{pmatrix} 5 \\ 3 \end{pmatrix}.$$

We have $\mathrm{SS}_{\mathrm{Res}}(V) = 224$.

The vectors $\boldsymbol{c}_1 = (0,1,-1,0)'$ and $\boldsymbol{c}_2 = (1,0,0,0)'$ are in the kernel of $\boldsymbol{A}$ and are linearly independent (and so span the kernel), so OLS onto V_0 is equivalent to OLS onto

$$\boldsymbol{W} = \boldsymbol{X}\boldsymbol{C} = \begin{pmatrix} 1 & 3 & 7 & 8 \\ 1 & -1 & 3 & 4 \\ 1 & 4 & 2 & 9 \\ 1 & 6 & 0 & -5 \\ 1 & 3 & 3 & 4 \\ 1 & 3 & 3 & 4 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 10 & 1 \\ 2 & 1 \\ 6 & 1 \\ 6 & 1 \\ 6 & 1 \\ 6 & 1 \end{pmatrix}.$$

This can also be seen by making the substitution $\beta_3 = \beta_2$ and $\beta_4 = 0$ in the model equation

$$\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{X}] = \beta_1 \boldsymbol{x}_1 + \beta_2 \boldsymbol{x}_2 + \beta_2 \boldsymbol{x}_3 = \beta_1 \boldsymbol{x}_1 + \beta_2 (\boldsymbol{x}_2 + \boldsymbol{x}_3).$$

Then

$$\hat{\boldsymbol{Y}}_0 = \begin{pmatrix} 4 \\ 36 \\ 20 \\ 20 \\ 20 \\ 20 \end{pmatrix}, \quad \boldsymbol{Y} - \hat{\boldsymbol{Y}}_0 = \begin{pmatrix} 0 \\ 0 \\ 24 \\ -8 \\ -4 \\ -12 \end{pmatrix}.$$

We conclude that $\|\hat{\boldsymbol{Y}}_0 - \boldsymbol{Y}\|^2 = 800$. Thus

$$F = \frac{(800-224)/2}{224/2} = 2.5714.$$

We have

$$\boldsymbol{Z}'(\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}')^{-1}\boldsymbol{Z} = \begin{pmatrix} 5 & 3 \end{pmatrix} \begin{pmatrix} 18 & -30 \\ -30 & 114 \end{pmatrix} \begin{pmatrix} 5 \\ 3 \end{pmatrix} = 576 = 800 - 224.$$

The analysis of variance table is displayed in Table 1. □

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	4	800.00				
2	2	224.00	2	576.00	2.57	0.2800

TABLE 1. F statistic comparing V_0 and V

Problem 2. Consider the data in the dataset `teengamb` in the package `faraway`:

```
install.packages("faraway")
library(faraway)
data(teengamb)
```

The last line should bring up a description of the variables.

Is there a difference between males and females as relates to gambling behavior? Fit any appropriate model(s) and carry out any appropriate test(s).

Solution to Problem 2. First, let us investigate the linear model giving `gamble` as the response and the other variables as covariates. The estimated coefficients and the associated t-tests are given in Table 2. The t-tests are only significant on sex and income.

We might test whether the coefficients of status and verbal are *both* zero. The F-test of this hypothesis is given in Table 3. Since it is not significant, we will assume that status and verbal have zero coefficients, i.e. are not included in the model.

Finally, we have not considered whether the coefficient of income should depend on sex. We introduce the produce `sex:income` to test if there should be a different coefficient of income depending on sex, given that verb and status are not included. The F-test is shown in 4. The test is significant, so we conclude that the coefficient should depend on sex.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	22.56	17.20	1.31	0.20
status	0.05	0.28	0.19	0.85
income	4.96	1.03	4.84	0.00
verbal	-2.96	2.17	-1.36	0.18
sex1	-22.12	8.21	-2.69	0.01

TABLE 2. All variables (no interaction)

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	44	22781.32				
2	42	21623.77	2	1157.55	1.12	0.3345

TABLE 3. F test on verbal and status

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	44	22781.32				
2	43	18929.91	1	3851.41	8.75	0.0050

TABLE 4. F test on sex:income

In summary, we conclude that the model should fit both separate intercepts and slopes (coefficient of income), depending on sex. Thus if $\boldsymbol{x}$ is income and $\boldsymbol{\delta}$ is the indicator of female, we fit

$$\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{x}, \boldsymbol{\delta}] = \beta_0 + \beta_1 \boldsymbol{x} + \beta_2 \boldsymbol{\delta} + \beta_3 \boldsymbol{\delta} \boldsymbol{x}.$$

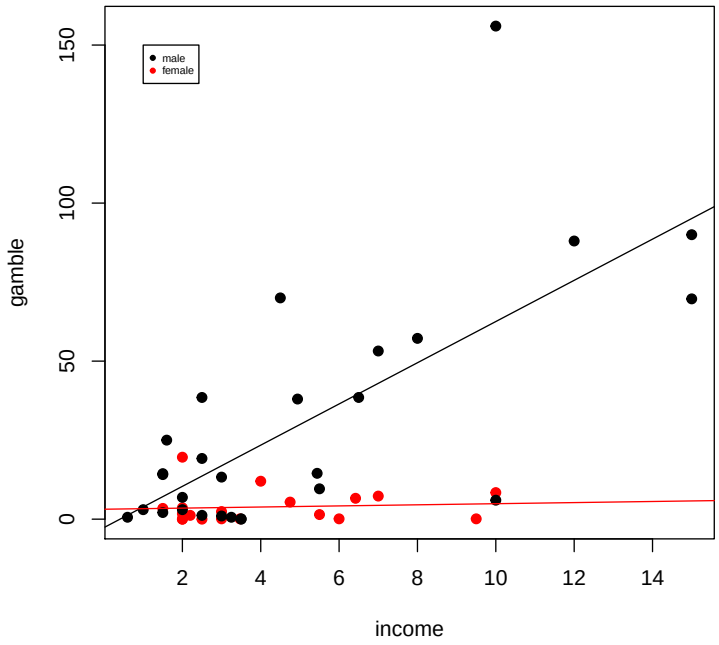

FIGURE 1. Gamble vs. income.

A plot of gamble vs. income is shown in Figure 1. Note that the slope for females looks close to zero. We want to test the hypothesis that $\beta_1 + \beta_3 = 0$. To do so, we fit the model

$$\begin{aligned} E[\boldsymbol{Y} \mid \boldsymbol{x}, \boldsymbol{\delta}] &= \beta_0 + \beta_1 \boldsymbol{x} + \beta_2 \boldsymbol{\delta} - \beta_1 \boldsymbol{\delta} \boldsymbol{x} \\ &= \beta_0 + \beta_1 \boldsymbol{x}(1 - \boldsymbol{\delta}) + \beta_2 \boldsymbol{\delta} \end{aligned}$$

and compare with the larger model via an F-test. (Table 5.) The test is not significant. We conclude that, for females, there is no relationship between income and gamble. In fact, females gamble very little on average, regardless of income. Males, however, tend to gamble, and an increasing amount depending on income.

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	44	18933.63				
2	43	18929.91	1	3.72	0.01	0.9272

TABLE 5. Test that coefficient of income for females is zero

Note that we formulated some of the hypotheses after looking at significance level of previous tests; this "data snooping" should make us skeptical of reported confidence levels.

□

Problem 3. Suppose that 11 plots of land are plotted with three varieties of corn. The following lists the yields for the three varieties:

I	II	III
52	64	53
56	57	55
60	62	58
56		50

(a) Test the hypothesis that the three varieties all have the same expected yield .
(b) Suppose that for the corn yield the true means were 70, 75, 95 and that $\sigma = 20$. Find the power of the $\alpha = 0.05$ level test for equal means.
(c) How large should n_0, the number of observations per treatment (number of plots per treatment) be in order to have power at least 0.90 for the parameters in (a)?

Solution to Problem 3. We fit the model

$$\mathbb{E}[Y \mid \boldsymbol{X}] = \beta_1 + \beta_2\delta_2 + \beta_3\delta_3,$$

where

$$\delta_{i,j} = \begin{cases} 1 & \text{if } i\text{-th plot is variety } j \\ 0 & \text{otherwise} \end{cases}.$$

Thus $\beta_1 = \mu_1$ and $\beta_j = (\mu_j - \mu_1)$ for $j = 2,3$, where μ_j is the expected yield of variety j. Testing $H_0 : \mu_1 = \mu_2 = \mu_3$ is given by the F statistic corresponding to $H_0 : \beta_2 = \beta_3 = 0$. The F-test is reported in Table 6, and is not significant.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	2	86.55	43.27	3.76	0.0705
Residuals	8	92.00	11.50		

TABLE 6. F test of equal means

Note that H_0 corresponds to

$$\boldsymbol{\theta} = \mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{X}] \in V_0 = \mathscr{L}(\mathbf{1}).$$

If $W = V \cap V_0^{\perp}$, then the distribution of the F-statistic has non-centrality parameter $\|\Pi_W\boldsymbol{\theta}\|^2/\sigma^2$. We write $\boldsymbol{x}^{\perp} = \boldsymbol{x} - \Pi_{V_0}\boldsymbol{x}$. Then

$$\|\Pi_W\boldsymbol{\theta}\|^2 = \beta_2^2\|\delta_2^{\perp}\|^2 + \beta_3^2\|\delta_3^{\perp}\|^2 + 2\beta_2\beta_3\langle\delta_2^{\perp}, \delta_3^{\perp}\rangle.$$

We determine the projection of $\boldsymbol{\delta}_j$ on $\mathbf{1}$, the corresponding residual vector $\boldsymbol{\delta}_j^{\perp}$, and the norm:

$$\begin{aligned}
\Pi_{V_0}\boldsymbol{\delta}_j &= \frac{\langle\boldsymbol{\delta}_j, \mathbf{1}\rangle}{\|\mathbf{1}\|^2}\mathbf{1} = \frac{n_j}{n}\mathbf{1} \\
\boldsymbol{\delta}_j^{\perp} &= \boldsymbol{\delta}_j - \frac{n_j}{n}\mathbf{1} \\
\delta_{i,j}^{\perp} &= \begin{cases} 1 - \frac{n_j}{n} & \text{if } i\text{-th plot is variety } j \\ -\frac{n_j}{n} & \text{otherwise} \end{cases} \\
\|\boldsymbol{\delta}_j^{\perp}\|^2 &= n_j\left(1 - \frac{n_j}{n}\right)^2 + \left(\frac{n_j}{n}\right)^2(n - n_j) \\
&= n_j\left(1 - \frac{n_j}{n}\right).
\end{aligned}$$

Also, since $\boldsymbol{\delta}_j \perp \boldsymbol{\delta}_k$, for $j \neq k$,

$$\begin{aligned}\langle \boldsymbol{\delta}_j^{\perp}, \boldsymbol{\delta}_k^{\perp} \rangle &= \langle \boldsymbol{\delta}_j - \frac{n_j}{n}\mathbf{1}, \boldsymbol{\delta}_k - \frac{n_k}{n}\mathbf{1} \rangle \\ &= \langle \boldsymbol{\delta}_j, \boldsymbol{\delta}_k \rangle - \langle \frac{n_j}{n}\mathbf{1}, \boldsymbol{\delta}_k \rangle - \langle \boldsymbol{\delta}_j, \frac{n_k}{n}\mathbf{1} \rangle + \langle \frac{n_k}{n}\mathbf{1}, \frac{n_k}{n}\mathbf{1} \rangle \\ &= -\langle \boldsymbol{\delta}_k, \frac{n_j}{n}\mathbf{1} \rangle - \langle \boldsymbol{\delta}_j, \frac{n_k}{n}\mathbf{1} \rangle + \langle \frac{n_j}{n}\mathbf{1}, \frac{n_k}{n}\mathbf{1} \rangle \\ &= -\frac{n_j n_k}{n} - \frac{n_j n_k}{n} + \frac{n_j n_k n}{n^2} = -\frac{n_j n_k}{n}.\end{aligned}$$

Thus,

$$\|\Pi_W \boldsymbol{\theta}\|^2 = \beta_2^2 n_2 \left(1 - \frac{n_2}{n}\right) + \beta_3^2 n_3 \left(1 - \frac{n_3}{n}\right) - 2\beta_2\beta_3 \frac{n_j n_k}{n}.$$

When $\mu_1 = 70, \mu_2 = 75, \mu_3 = 95$, we have $\beta_2 = 75 - 70 = 5$ and $\beta_3 = 95 - 70 = 25$, and thus the noncentrality parameter is

$$\gamma = \frac{\|\Pi_W \boldsymbol{\theta}\|^2}{\sigma^2} = \frac{1}{20^2}\left(5^2 3(1 - \frac{3}{11}) + 25^2 4(1 - \frac{4}{11}) - 2 \cdot 5 \cdot 25 \frac{3 \cdot 4}{11}\right) = 3.43$$

Since the 95-th percentile for a $F_{2,8}$ distribution is 4.459, the power of the test is

$$\mathbb{P}(F_{2,8,3.43} > 4.459) = 0.261.$$

If $n_1 = n_2 = n_3 = n_0$, then $n_j/n = 1/3$ and the non-centrality parameter γ is

$$\gamma(n_0) = \frac{\|\Pi_W \boldsymbol{\theta}\|^2}{\sigma^2} = \frac{2n_0}{3\sigma^2}(\beta_2^2 + \beta_3^2 - \beta_2\beta_3) = 0.875 n_0.$$

By trial and error, we see that we need $n_0 = 20$ to get the power at least 0.9. □