STOR 664 CWE Summer 2015

(1) Consider a one-way ANOVA model. n patients participating in a clinical trial for treating hypertension are equally divided in I groups with J patients in each group (i.e. $n = IJ$). Let y_{ij} represent the change in blood pressure of patient j in group i. Assume y_{ij}'s are independent random variables with $y_{ij} \sim N(\theta_i, \sigma^2)$, $j = 1, ..., J$, $i = 1, ..., I$; and $\theta_i, i = 1, ..., I$ and σ^2 are unknown parameters.

(1a) Express this ANOVA model in a matrix form $Y = X\theta + \epsilon$ where Y and ϵ are n-dimensional vectors, X is a $n \times p$ matrix and θ is a p-dimensional vector. Specify p.

(1b) Consider testing the contrasts H_0: $\theta_1 - \theta_2 = \cdots = \theta_{I-1} - \theta_I$ (reduced model) versus the full model H_1. Complete the following ANOVA table by filling each empty cell with the correct degree of freedom and corresponding mean square.

Source	*Sum of Squares*	*D.F.*	*Mean Square*
Full Model	SSR_1		
Reduced Model	SSR_0		
Difference	$SSE_0 - SSE_1$		
Residual	SSE_1		
Total	$SSTO$		

Table 1: ANOVA table for two nested models

(1c) Express the null hypothesis H_0 in a matrix form $C\theta = h$ where C is a $q \times p$ matrix and h is a q-dimensional vector. Specify q and h.

(1d) Spell out the following sums of squares in terms of y_{ij}'s:

$$SSTO =$$

$$SSR_1 =$$

$$SSE_1 =$$

$$SSR_0 =$$

$$SSE_0 =$$

Hint: If you have difficulty to obtain SSE_0 or SSR_0, just note that in this case X^tX enjoys a simple form that makes the formula

$$\widehat{\beta}_0 = \widehat{\beta}_1 - (X^tX)^{-1}C^t[C(X^tX)^{-1}C^t]^{-1}(C\widehat{\beta}_1 - h)$$

manageable. $\widehat{\beta}_0$ and $\widehat{\beta}_1$ in the current problem are least square estimates for the parameter vector θ in the reduced and full models respectively.

(1e) Consider the special case $I = 3$ with two different alternatives

$$H_1': \quad \theta_1 - \theta_2 = \theta_2 - \theta_3 - 1,$$
$$H_1'': \quad \theta_1 - \theta_2 = \theta_2 - \theta_3 + 2.$$

Which one would give a greater power against H_0? Justify your answer.

(1f) Does H_0 given in (1b) imply that responses in all I groups are indistinguishable? If not, how should we formulate a null hypothesis to answer that question? Write down the expression of the test statistic, and specify its probability distribution under your new H_0 and the related degrees of freedom.

(2) **Read the following example taken from Faraway's book; answer the inserted questions and make your comments.**
Car drivers like to adjust the seat position for their own comfort. Car designers would find it helpful to know where different drivers will position the seat depending on their size and age. Researchers at the HuMoSim laboratory at the University of Michigan collected data on 38 drivers. They measured age in years, weight in pounds, height with shoes and without shoes in cm, seated height arm length, thigh length, lower leg length and hipcenter the horizontal distance of the midpoint of the hips from a fixed location in the car in mm. A model is fitted with all the predictors:

R output-1

```
   > data (seatpos)
   > g < - lm (hipcenter ~. , seatpos)
   > summary (g)
Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept)   436.4321   166.5716    2.62    0.014
Age             0.7757     0.5703    1.36    0.184
Weight          0.0263     0.3310    0.08    0.937
HtShoes        -2.6924     9.7530   -0.28    0.784
Ht              0.6013    10.1299    0.06    0.953
Seated          0.5338     3.7619    0.14    0.888
```

```
Arm            -1.3281     3.9002   -0.34    0.736
Thigh          -1.1431     2.6600   -0.43    0.671
Leg            -6.4390     4.7139   -1.37    0.182
Residual standard error: 37.7 on 29 degrees of freedom
Multiple R-Squared: 0.687, Adjusted R-squared: 0.6
F-statistic: 7.94 on 8 and 29 DF, p-value: 1.31e-05
```

(2a) Noticing the small p-value for the F-test, the fairly significant R^2-value, and that none of the individual predictors is significant. Do they send conflicting signals for the model fitting? Share your thoughts.

R output-2: pairwise correlations

```
> round(cor(seatpos), 3)
             Age Weight HtShoes     Ht  Seated    Arm  Thigh    Leg   hipcenter
Age        1.000  0.081  -0.079 -0.090  -0.170  0.360  0.091 -0.042       0.205
Weight     0.081  1.000   0.828  0.829   0.776  0.698  0.573  0.784      -0.640
HtShoes   -0.079  0.828   1.000  0.998   0.930  0.752  0.725  0.908      -0.797
Ht        -0.090  0.829   0.998  1.000   0.928  0.752  0.735  0.910      -0.799
Seated    -0.170  0.776   0.930  0.928   1.000  0.625  0.607  0.812      -0.731
Arm        0.360  0.698   0.752  0.752   0.625  1.000  0.671  0.754      -0.585
Thigh      0.091  0.573   0.725  0.735   0.607  0.671  1.000  0.650      -0.591
Leg       -0.042  0.784   0.908  0.910   0.812  0.754  0.650  1.000      -0.787
hipcenter  0.205 -0.640  -0.797 -0.799  -0.731 -0.585 -0.591 -0.787       1.000
```

R output-3: variance inflation factors (VIFs)

```
> vif (x)
   Age Weight  HtShoes       Ht Seated    Arm  Thigh    Leg
1.9979 3.6470 307.4294 333.1378 8.9511 4.4964 2.7629 6.6943
```

(2b) Comment on evidences provided by R output-2 and output-3.

The 6 length variables (from "HtShoes" to "Leg") are strongly correlated with each other — any one of them might do a good job of representing the others. Pick "Ht" as the simplest to measure. We are not claiming that the other predictors are not associated with the response, just that no need for including all of them to predict the response. Refitting the reduced model yields

R output-4

```
> g2 < - lm (hipcenter ~ Age + Weight + Ht, seatpos)
> summary (g2)
Coefficients:
             Estimate Std. Error t value Pr (>|t|)
(Intercept) 528.29773  135.31295    3.90   0.00043
Age           0.51950    0.40804    1.27   0.21159
Weight        0.00427    0.31172    0.01   0.98915
Ht           -4.21190    0.99906   -4.22   0.00017
Residual standard error: 36.5 on 34 degrees of freedom
Multiple R-Squared: 0.656, Adjusted R-squared: 0.626
F-statistic: 21.6 on 3 and 34 DF, p-value: 5.13e-08
```

(2c) Interpret similarities and distinctions between R output-4 and R output-1, and make your recommendation.

(2d) If for some reason all predictors in R output-1 must be kept, what alternative estimation method would you consider to improve the fit? Describe it briefly.