Statistics 312/612, fall 2016
Homework # 6
Due: Monday 24 October

[1] (extra credit) Suppose $y \sim N(X\theta, \sigma^2 I_n)$ where X is an $n \times p$ matrix of rank $m < p$ and θ is constrained to belong to the subspace of $\mathbb{R}^p$ generated by the columns of a $p \times m$ matrix $\mathbb{M}$ of rank m. That is, $\theta \in \Theta = \{\mathbb{M}t : t \in \mathbb{R}^m\}$. For a given g in $\mathbb{R}^p$, we seek an L in $\mathbb{R}^n$ for which $L^T y \sim N(g^T\theta, \sigma^2 \|L\|^2)$ under the θ model, where $\|L\|$ is as small as possible.

(i) Express the minimizing L in terms of the singular value decomposition $U_1\Lambda_1 V_1^T$ for the matrix $\widetilde{X} = X\mathbb{M}$ and the vector d for which $\mathbb{M}^T g = V_1 d$. (I know the solution is similar to something in the notes but I want to see you derive the whole result.)

(ii) Deduce from (i) that $L^T y = g^T\widehat{\theta}$.

[2] The handout Normal.pdf discussed once more the reparametrization via Helmert contrasts for the least squares fit `outBC <- lm(rate ~ Ht + Hp,BC)`. It asserted that there is a one-to-one correspondence between the vector of constrained parameters

$$\theta = [int, A, B, C, D, I, II, III]$$

and the vector of parameters for the reparametrized fit

$$\tau = [H.int, Ht1, Ht2, Ht3, Hp1, Hp2].$$

Remark. For typesetting convenience I am referring to the components of θ and τ by the names used in **R**, rather in subscripted notation: $\theta_{int}, \theta_A, \ldots$ and $\tau_{H.int}, \tau_{Ht1}, \ldots$.

The correspondence comes via the equality $z = X\theta = \widetilde{X}\tau$ for z in $\mathcal{X}$, the 6-dimensional subspace of $\mathbb{R}^{48}$ spanned by the columns of the matrix

$$X = (\mathbb{1}_{48}, F_1, F_2, F_3, F_4, G_1, G_2, G_3).$$

Here $F = (F_1, F_2, F_3, F_4)$ is the matrix of dummy variables for the factor `Ht` and $G = (G_1, G_2, G_3)$ is the matrix of dummy variables for the factor `Hp`. The space $\mathcal{X}$ is also spanned by the columns of $\widetilde{X} =$ `model.matrix(outBC)`.

(i) (20 points) Show how each of the θ's is represented as a linear combination of the τ's. Also show how each of the τ's is represented as a linear combination of the θ's.

You might find it to be more convenient to present your answers as matrices with dimnames, printed out by **R**:

```
dimnames(theta.from.tau);              # print(theta.from.tau)

## [[1]]
## [1] "int" "A"   "B"   "C"   "D"   "I"   "II"  "III"
##
## [[2]]
## [1] "H.int" "Ht1"   "Ht2"   "Ht3"   "Hp1"   "Hp2"

dimnames(tau.from.theta);        # print(fractions(tau.from.theta))

## [[1]]
## [1] "H.int" "Ht1"   "Ht2"   "Ht3"   "Hp1"   "Hp2"
##
## [[2]]
## [1] "int" "A"   "B"   "C"   "D"   "I"   "II"  "III"
```

To make your matrices look pretty use the fractions() function from the MASS library.

Hint: Two rows of X are they same if they correspond to the same combination of the two factors. The problem is the essentially unchanged if we discard all duplicate rows, replacing X by the 12×8 matrix $X_0 =$ `unique(X)` and $\widetilde{X}$ by $\widetilde{X}_0 =$ `unique(Xtilde)`.

(ii) (20 points) With X replaced by X_0, the generic element of $\mathcal{X}_0 = \text{span}(X_0)$ is a 12×1 vector z, whose components can be labelled as

$$AI, BI, CI, DI, AII, BII, CII, DII, AIII, BIII, CIII, DIII$$

Produce displays showing how z is a linear function of τ and how τ is a linear funcion of z.

```
options(width=100)
dimnames(z.from.tau);    # print(z.from.tau)

## [[1]]
##  [1] "AI"   "BI"   "CI"   "DI"   "AII"  "BII"  "CII"  "DII"  "AIII" "BIII" "CIII" "DIII"
##
## [[2]]
## [1] "H.int" "Ht1"   "Ht2"   "Ht3"   "Hp1"   "Hp2"

cat("\n")
```

```
dimnames(tau.from.z);    # print(fractions(tau.from.z))

## [[1]]
## [1] "H.int" "Ht1"   "Ht2"   "Ht3"   "Hp1"   "Hp2"
##
## [[2]]
##  [1] "AI"   "BI"   "CI"   "DI"   "AII"  "BII"  "CII"  "DII"  "AIII" "BIII" "CIII" "DIII"
```

(iii) (5 points) As a check, show the **R** code confirming that (within round-off error) the 12 distinct values from `outBC$fit` are given by $X_0\widehat{\theta} = \widetilde{X}_0\widehat{\tau}$.

(iv) (10 points) Now suppose we are interested in some linear combination of the parameters, $g^T\theta$, for a specified g. Find the distribution of $g^T\widehat{\theta}$ under the θ-model. Explain your reasoning. Express your answer as an expression involving the unknown σ^2, the unknown θ, quantities available from `outBC$qr`, and the matrix $\mathbb{M}$ for which $\widetilde{X} = X\mathbb{M}$.

(v) (10 points) Find the constant c for which $cg^T\widehat{\theta}/\widehat{\sigma}$ has a t-distribution if $g^T\theta = 0$.

(vi) (10 points) Calculate the t-statistic and the two-sided p-value for that statistic for testing the hypothesis that $\theta_A = \theta_B$, assuming the truth of the model.

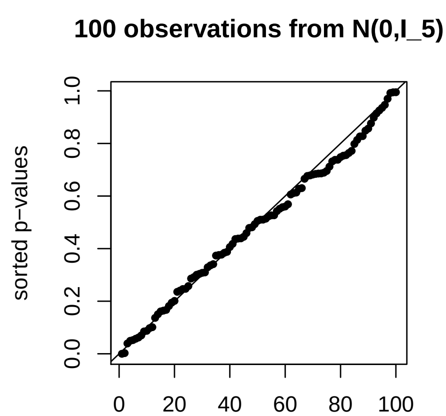
[3] In class you learned that if $Z = [Z_1, \ldots, Z_5] \sim N(0, I_5)$ then the statistic

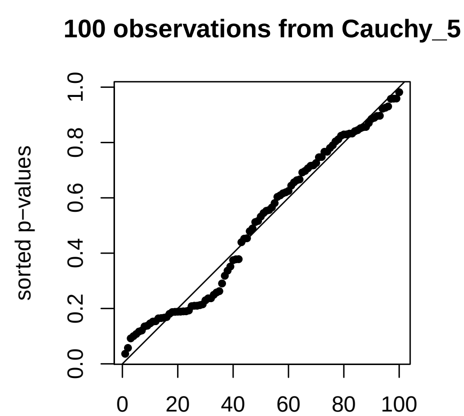
$$T = \sqrt{5}\overline{Z}/\sqrt{\sum\nolimits_{i\leq 5}(Z_i - \overline{Z})^2/4}$$

has a t_4-distribution. Equivalently $pt(T, 4)$ has a uniform distribution, where `pt(t,4)` is the **R** command to calculate $\mathbb{P}\{T_4 \leq t\}$ for $T_4 \sim t_4$.

(i) (5 points) Write an **R** function that calculates T from a 5×1 vector Z. Show your code.

(ii) (5 points) Write an **R** function that takes a 5×100 matrix ZZ, calculates the p-value for each column, then plots the sorted p-values for each column against `1:100`. If you want to be fancy, add a straight line as in the following pictures. Show your code.





(iii) (5 points) Show **R** code that would generate the two pictures: first when ZZ is filled with independent $N(0, 1)$'s and then when ZZ is filled with independent standard Cauchy random variables. (I used `set.seed(0)`. You might want to do the same if you hope to convince us by picture that your code works.)

(iv) (extra points) Draw some wise conclusion from a comparison of the two pictures. For example, why don't the long tails of the Cauchy have a stronger effect on the second picture?