Statistics 312/612, fall 2016
Homework # 1
Due: Monday 12 September

[1] (5 points) Suppose $A = (a_1, \ldots, a_n)$ is an $n \times n$ matrix for which $\|At\| = \|t\|$ for all $t \in \mathbb{R}^n$. (That is, A is an isometry.) Show that the columns $a_1, \ldots, a_n$ are orthogonal unit vectors, that is, A is an orthogonal matrix. Hint: Let e_i denote the ith column of the identity matrix I_n. Consider $\|Ae_1\|$ and $\|A(e_1 + e_2)\|^2$.

[2] (10 points) If y and w are vectors in $\mathbb{R}^n$ the correlation between them is defined as

$$\text{corr}(y, w) = \frac{\langle y - \bar{y}\mathbb{1}, w - \overline{w}\mathbb{1}\rangle}{\|y - \bar{y}\mathbb{1}\| \, \|w - \overline{w}\mathbb{1}\|}$$

It equals the cosine of the angle between the vectors $y - \bar{y}\mathbb{1}$ and $w - \overline{w}\mathbb{1}$, which always lies between -1 and $+1$.

Now suppose that $q_1 = n^{-1/2}\mathbb{1}, q_2, \ldots, q_n$ is an orthonormal basis for $\mathbb{R}^n$ and that $\mathcal{X}$ is the m-dimensional subspace spanned by $q_1, \ldots, q_m$. Let w equal $\hat{y}$, the orthogonal projection of y onto $\mathcal{X}$. The following facts could be shown using the representation $y = \sum_{i \le n} s_i q_i$ where $s_i = \langle q_i, y\rangle$, although that is not the quickest method.

(i) Show that $\overline{w} = \bar{y}$. Hint: The residual vector $r = y - \hat{y}$ is orthogonal to $\mathcal{X}$.

(ii) Show that $\langle y - \bar{y}\mathbb{1}, \hat{y} - \bar{y}\mathbb{1}\rangle = \|\hat{y} - \bar{y}\mathbb{1}\|^2$.

(iii) Deduce that $\text{corr}(y, \hat{y})^2 = \|\hat{y} - \bar{y}\mathbb{1}\|^2 / \|y - \bar{y}\mathbb{1}\|^2$.

Remark. The quantity $\text{corr}(y, \hat{y})^2$ is usually denoted by R^2, sometimes called the "multiple R-squared". (Not to be confused with the matrix R from the QR decomposition nor the program **R**). It takes values between 0 and 1. If $R^2 = 1$ then $\hat{y} = y$, which means the model has done a good job at approximating y. Some users of regression get very happy when R^2 is close enough to 1, where 'close enough' can sometimes mean $R^2 > 0.1$. In my experience, $R^2 \approx 1$ is usually a sign that something is very wrong with the model.

[3] (10 points) The handout `RMDdemo.Rmd` considered a toy least squares example where the model space $\mathcal{X}$ was spanned by the columns of a matrix not of full rank. Using only the vector y and *out$qr* from `out <- lm(y ~ ., data=mydata)`, I showed how the fitted vector $\hat{y}$ could be calculated using matrix calculations.

For this problem I want you to show how some of the other parts of `summary(out)` (see next page) could be calculated using only `y` and `out$qr`. Display your calculations using R Markdown. *No cheating by cutting and pasting from the summary.*

Show how to get the two lines following: "Residuals:"; the coefficients (but ignore the stuff about standard errors and t-values); the residual standard error, which is defined as

$$\sqrt{\sum\nolimits_{i \le n} r_i^2 / (\text{degrees of freedom})} \quad ;$$

and the multiple R-squared.

```
> set.seed(10)   # for reproducibility
> mydata <- data.frame(y=rnorm(10),
+  x1=1:10,x2= 11:20, x3= 0.5*(1:10)-3*(11:20))
> out <- lm(y ~ ., data=mydata)
> summary(out)

Call:
lm(formula = y ~ ., data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0211 -0.5231  0.1832  0.4320  0.9085

Coefficients: (2 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.18175    0.49193  -0.369    0.721
x1          -0.05616    0.07928  -0.708    0.499
x2                NA         NA      NA       NA
x3                NA         NA      NA       NA

Residual standard error: 0.7201 on 8 degrees of freedom
Multiple R-squared:  0.05903,Adjusted R-squared:  -0.05859
F-statistic: 0.5019 on 1 and 8 DF,  p-value: 0.4988
```