Statistics 312a/612a, fall 2016
Homework #4 solutions

[1] *The catheter data set is taken from a well known text book. If you happen to know the book please do not just repeat the analysis it presents.*

```
cath <- read.table("catheter.txt",header=T)
outH <- lm(distance ~ height, cath)
outW <-lm(distance ~ weight, cath)
outHW <- lm(distance ~ height + weight,cath)
```

The summary information for each fit (`outHW`*,* `outH`*,* `outW`*) seems to suggest that height by itself is a good predictor of distance, that weight by itself is a good predictor of distance, but when both predictors are used then neither is particularly useful. (The stars in the summary table suggest 'significance'.)*

Solution: *Here are the shorter summaries, without the stars:*

```
look(outH)

## lm(formula = distance ~ height, data = cath)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   12.124      4.247   2.855    0.017
## height         0.597      0.101   5.894    0.000

look(outW)

## lm(formula = distance ~ weight, data = cath)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   25.637      2.004  12.792        0
## weight         0.277      0.044   6.303        0

look(outHW)

## lm(formula = distance ~ height + weight, data = cath)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   21.008      8.751   2.401    0.040
## height         0.196      0.361   0.545    0.599
## weight         0.191      0.165   1.155    0.278
```

The stars would appear next to the rows where `Pr(>|t|)` *is very small.*

(i) (15 points) Add more variables to the cath data.frame: height and weight centered to zero means (call them hcen and wcen); and

`cath$wres = <-lm(wcen~ hcen, cath)$res`.

Explain why some of the coeffs and std. errors are the same and some are different for the models:

```
distance ~ height       # outH
distance ~ hcen
distance ~ hcen + wres
distance ~ height+wres
distance ~ hcen+wcen
distance ~ height+weight        # outHW
```

Solution: *The data came from Section 14.5 of the textbook "Mathematical Statistics and Data Analysis" by John Rice.*

```
mH <- mean(cath$height); cath$hcen <- cath$height - mH
mW <- mean(cath$weight); cath$wcen <- cath$weight - mW
cath$wres <- lm(wcen ~ hcen,cath)$res
cWres <- lm(wcen ~ hcen,cath)$coeff[2]      # constant c_{wh}

out.hcen <- lm(distance ~ hcen, cath)
out.hcen.wres <- lm(distance ~ hcen + wres, cath)
out.h.wres <- lm(distance ~ height + wres, cath)
out.hcen.wcen <- lm(distance ~ hcen + wcen, cath)
```

First observe that `outH` *and* `out.hcen` *represent the same least squares fit* $\widehat{y}_h$*, and* `outHW` *gives the same* $\widehat{y}$ *as* `out.hcen.wres`*,* `out.h.wres`*, and* `out.hcen.wcen`*.*

```
diffs <- cbind(out.hcen$fit - outH$fit, out.hcen.wres$fit- outHW$fit ,
        out.h.wres$fit- outHW$fit , out.hcen.wcen$fit- outHW$fit  )
round(  max (abs(diffs)) ,digits=5)

## [1] 0
```

Thus you had only to explain why various pairs of estimated coefficients are the same. (Whenever two coefficients for the same least squares fit match so do their standard errors.)

Define

$$
\begin{aligned}
y &= \text{distance} \\
\boldsymbol{h} &= \text{the vector of heights} \\
\boldsymbol{w} &= \text{the vector of weights} \\
m_h &= \text{average of the heights } = 40.36 \\
m_w &= \text{average of the weights } = 38.12 \\
\boldsymbol{h}_c &= \boldsymbol{h} - m_h\mathbb{1} \\
\boldsymbol{w}_c &= \boldsymbol{w} - m_w\mathbb{1} \\
\boldsymbol{w}_r &= \boldsymbol{w}_c - c_{wh}\boldsymbol{h}_c = \text{part of } \boldsymbol{w}_c \text{ orthogonal to } \boldsymbol{h}_c \qquad \text{where } c_{wh} = 2.098.
\end{aligned}
$$

Remark. *The formula* `wres <- lm(wcen ~ hcen)$res` *suggests that* $\boldsymbol{w}_r = \boldsymbol{w}_c - c_{wh}\boldsymbol{h}_c - c_0\mathbb{1}$*. The constant* c_0 *must be zero, because all of* $\boldsymbol{w}_r, \boldsymbol{h}_c, \boldsymbol{w}_c$ *are orthogonal to* $\mathbb{1}$*.*

The six least squares fits involve the components $\widehat{y}$*,* $\widehat{y}_{\text{int}}$*,* $\widehat{y}_h$*,* $\widehat{y}_{w\perp h,1}$*, and* $\widehat{y}_{hw\perp 1}$ *of* y *in the subspaces*

$$
\begin{aligned}
\mathcal{X} &= \text{span}(\mathbb{1}, \boldsymbol{h}, \boldsymbol{w}) = \text{span}(\mathbb{1}, \boldsymbol{h}_c, \boldsymbol{w}_c) = \text{span}(\mathbb{1}, \boldsymbol{h}_c, \boldsymbol{w}_r) \\
\mathcal{X}_{\text{int}} &= \text{span}(\mathbb{1}) \\
\mathcal{X}_h &= \text{span}(\mathbb{1}, \boldsymbol{h}) = \text{span}(\mathbb{1}, \boldsymbol{h}_c) \\
\mathcal{X}_{w\perp h,1} &= \text{span}(\boldsymbol{w}_r) = \text{subspace of } \mathcal{X} \text{ orthogonal to } \mathcal{X}_h \\
\mathcal{X}_{hw\perp 1} &= \text{span}(\boldsymbol{h}_c, \boldsymbol{w}_c) = \text{span}(\boldsymbol{h}_c, \boldsymbol{w}_r) = \text{subspace of } \mathcal{X} \text{ orthogonal to } \mathbb{1}
\end{aligned}
$$

The fit outH expresses the component of y *in* $\mathcal{X}_h$ *as*

$$
\begin{aligned}
\widehat{y}_h &= \widehat{m}_H\mathbb{1} + \widehat{a}_H\boldsymbol{h} = 12.124\mathbb{1} + 0.597\boldsymbol{h} \\
&= (\widehat{m}_H + m_h\widehat{a}_H)\mathbb{1} + \widehat{a}_H\boldsymbol{h}_c = 36.208\mathbb{1} + 0.597\boldsymbol{h}_c.
\end{aligned}
$$

Those equalities explain why the fitted vectors for `outH` *and* `out.hcen` *are the same and their ahat coefficients are the same.*

For the comparisons between `outHW` *and* `out.[hcen.wres|h.wres|hcen.wcen]` *we have*

$$
\begin{aligned}
\widehat{y} &= \widehat{m}\mathbb{1} + \widehat{a}\boldsymbol{h} + \widehat{b}\boldsymbol{w} \qquad \text{with outHW\$coeff } = (21.008, 0.196, 0.191) \\
&= \widehat{m}\mathbb{1} + \widehat{a}(\boldsymbol{h}_c + m_h\mathbb{1}) + \widehat{b}(\boldsymbol{w}_c + m_w\mathbb{1}) \\
&= \widehat{m}\mathbb{1} + \widehat{a}(\boldsymbol{h}_c + m_h\mathbb{1}) + \widehat{b}(\boldsymbol{w}_r + c_{wr}\boldsymbol{h}_c + m_w\mathbb{1}) \\
&= \widehat{m}\mathbb{1} + \widehat{a}\boldsymbol{h} + \widehat{b}(\boldsymbol{w}_r + c_{wr}(\boldsymbol{h} - m_h\mathbb{1}) + m_w\mathbb{1}) \\
&= \begin{cases} (\widehat{m} + \widehat{a}m_h + \widehat{b}m_w)\mathbb{1} + \widehat{a}\boldsymbol{h}_c + \widehat{b}\boldsymbol{w}_c & \textit{for out.hcen.wcen} \\ (\widehat{m} + \widehat{a}m_h + \widehat{b}m_w)\mathbb{1} + (\widehat{a} + \widehat{b}c_{wr})\boldsymbol{h}_c + \widehat{b}\boldsymbol{w}_r & \textit{for out.hcen.wres} \\ (\widehat{m} - \widehat{b}m_h c_{wr} + \widehat{b}m_w)\mathbb{1} + (\widehat{a}m_h + \widehat{b}c_{wr})\boldsymbol{h} + \widehat{b}\boldsymbol{w}_r & \textit{for out.h.wres} \end{cases} \\
&= \begin{cases} 36.208\mathbb{1} + 0.196\boldsymbol{h}_c + 0.191\boldsymbol{w}_c & \textit{for hcen.wcen} \\ 36.208\mathbb{1} + 0.597\boldsymbol{h}_c + 0.191\boldsymbol{w}_r & \textit{for hcen.wres} \\ 12.124\mathbb{1} + 0.597\boldsymbol{h}_c + 0.191\boldsymbol{w}_r & \textit{for h.wres} \end{cases}
\end{aligned}
$$

In each of the last three lines the coefficients are uniquely determined because the three vectors are linearly independent.

```
## 6 x 3 sparse Matrix of class "dgCMatrix"
##                    int  ahat  bhat
## height          12.124 0.597 .
## hcen            36.208 0.597 .
## height+weight   21.008 0.196 0.191
## hcen+wcen       36.208 0.196 0.191
## hcen+wres       36.208 0.597 0.191
## height+wres     12.124 0.597 0.191
```

You should compare the coefficients in the table with the coefficients in the previous display.

In summary, y has components

$$
\begin{aligned}
\widehat{y}_{\text{int}} &= 36.208\mathbb{1} \qquad \text{in } \mathcal{X}_{\text{int}} \\
\widehat{y}_h &= \widehat{y}_{\text{int}} + \widehat{y}_{h\perp 1} = 12.124\mathbb{1} + 0.597\boldsymbol{h} = 36.208\mathbb{1} + 0.597\boldsymbol{h}_c \qquad \text{in } \mathcal{X}_h \\
\widehat{y}_{w\perp h,1} &= 0.191\boldsymbol{w}_r \qquad \text{in } \mathcal{X}_{w\perp h,1} \\
\widehat{y}_{hw\perp 1} &= 0.196\boldsymbol{h}_c + 0.191\boldsymbol{w}_c = 0.597\boldsymbol{h}_c + 0.191\boldsymbol{w}_r \qquad \text{in } \mathcal{X}_{hw\perp 1} \\
\widehat{y} &= 21.008\mathbb{1} + 0.196\boldsymbol{h} + 0.191\boldsymbol{w} = \widehat{y}_h + \widehat{y}_{w\perp h,1} \qquad \text{in } \mathcal{X}.
\end{aligned}
$$

(ii) (5 points) Explain why `summary(outHW)` *is misleading regarding the value of* `height` *and* `weight` *as predictors.*

```
look(outHW)

## lm(formula = distance ~ height + weight, data = cath)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    21.008      8.751   2.401    0.040
## height          0.196      0.361   0.545    0.599
## weight          0.191      0.165   1.155    0.278
```

Solution: *The height and weight are nearly linearly dependent. Loosely speaking, they are 'fighting' to represent one direction in the model space.*

More formally, the summary line for weight refers to the effect of adding the predictor $\boldsymbol{w}_r$ after the fits from the intercept and height have been removed:

$$\texttt{lm(outH\$res ~ -1 +wres,cath)\$coeff} = 0.191.$$

The `Pr(>|t|)` *is a comment on the coefficient of $\boldsymbol{w}_r$ in the* `outHW` *fit. Of course it seems less important because* `height` *has already taken out most of fit from* `height + weight`*.*

(iii) (5 points) Explain how the output from `cor(cath)` *is relevant to the problem.*

```
round(cor(cath)[1:2,],3)

##        height weight distance  hcen  wcen  wres
## height  1.000  0.961    0.881 1.000 0.961 0.000
## weight  0.961  1.000    0.894 0.961 1.000 0.276

# cor(height,weight) is close to 1
```

Height and weight really are highly correlated.

[2] The handout two_factors.pdf showed how to calculate several least squares fits using the Box-Cox data:

```
BC <- read.table("../Handouts/boxcox.data", header=T,sep="\t")
BC$rate <- 1/BC$time # transformation suggested by BHH page 235
BC$Htreatment <- C(BC$treatment,helmert)
out5 <- lm(rate ~ -1 + treatment,BC)
out7 <- lm(rate ~ Htreatment,BC)
```

In class I showed (page 9 of the handout) how to transform results from one parametrization into results for a different parametrization, using `out5` *and* `out6` *as an example. For this homework problem I want you to recreate the shortened summary*

```
## lm(formula = rate ~ -1 + treatment, data = BC)
##           tA    tB    tC    tD
## Est    3.519 1.862 2.947 2.161
## StdErr 0.292 0.292 0.292 0.292
```

using only the information contained in `out7`*, which is essentially the same as the* `out9` *generated by:*

```
C7 <- contrasts(BC$Htreatment)
dummyT <- outer(BC$treat,levels(BC$treat),"==")+0
X7 <- cbind(1, dummyT %*% C7)
out9 <- lm(BC$rate ~ -1+X7)
```

*Display all the **R** code that you use.*

(i) (5 points) Show that `X7` *is equal to* `dummyT %*% K7` *where* `K7 <- cbind(1,C7)`*.*

SOLUTION: *Not really hat I had in mind when I posed the question, but it would suffice:*

```r
K7 <- cbind(1,C7)
print( round (max(abs(X7- dummyT %*% K7))))

## [1] 0
```

(ii) *(10 points) If $\widehat{g}$ is the vector of coefficients from* `out7` *and $\widehat{b}$ is the vector of coefficients from* `out5`*, show that $\widehat{b} = K_7\widehat{g}$.*

SOLUTION: *Again I didn't really intend just a numerical check, but again the wording left open that possibility. I really wanted something like: Let $F = dummyT$. Then*

$$\widehat{y} = X_7\widehat{g} = F(K_7\widehat{g})$$

From the notes, $\widehat{y} = F\widehat{b}$ because $X_5 = F$. Linear independence of columns of F forces $K_7\widehat{g} = \widehat{b}$.

(iii) *(10 points) Use (ii) and* `out7` *to recreate the shortened summary for* `out5`*.*

```r
ghat <- out7$coefficients
bhat <- out5$coefficients
new.bhat <- as.vector(K7 %*% ghat)
V7 <- summary(out7)$cov
newV5 <- K7 %*% V7 %*% t(K7)
new.stderr <- sqrt(diag(newV5))
BC.coeff(out7,3)

## lm(formula = rate ~ Htreatment, data = BC)
##         (Int)    Ht1    Ht2     Ht3
## Est     2.622 -0.829 0.0855 -0.1538
## StdErr 0.146  0.207 0.1193  0.0844

BC.coeff(out5,3)

## lm(formula = rate ~ -1 + treatment, data = BC)
##            tA    tB    tC    tD
## Est     3.519 1.862 2.947 2.161
## StdErr 0.292 0.292 0.292 0.292

print(round(rbind(new.bhat,new.stderr),3))

##                  A     B     C     D
## new.bhat   3.519 1.862 2.947 2.161
## new.stderr 0.289 0.289 0.289 0.289
```