36-401 HW #3 Solutions

1. If $\mathbf{A}$ and $\mathbf{B}$ are 8×3 matrices,

   (a) $\mathbf{A}_{8\times3}\mathbf{B}_{8\times3}$ is not possible since the "inner dimensions" (3 columns for $\mathbf{A}$ and 8 rows for $\mathbf{B}$) disagree.

   (b) $\mathbf{B}_{8\times3}\mathbf{A}_{8\times3}$ is not possible since the "inner dimensions" disagree (3 columns for $\mathbf{B}$ and 8 rows for $\mathbf{A}$) disagree.

   (c) $\mathbf{A}_{8\times3}\mathbf{B}'_{3\times8}$ is possible, and the result is 8×8 (the "inner dimensions" cancel).

   (d) $\mathbf{A}^{-1}_{8\times3}$ is not possible since it is not possible to compute the inverse of a non-square matrix.

   (e) $(\mathbf{A}'_{3\times8}\mathbf{A}_{8\times3})^{-1}$ is possible, since $\mathbf{A}'_{3\times8}\mathbf{A}_{8\times3}$ is a square 3×3 matrix.

2. Want to compute $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ for

$$X = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

```
> X _ matrix(scan(),
+ ncol=4,byrow=T)
1: 1 1 0 0
5: 1 0 1 0
9: 1 0 0 1
13: 1 0 0 0
17:
> X
     [,1] [,2] [,3] [,4]
[1,]    1    1    0    0
[2,]    1    0    1    0
[3,]    1    0    0    1
[4,]    1    0    0    0
> XTX _ t(X) %*% X
> XTX
     [,1] [,2] [,3] [,4]
[1,]    4    1    1    1
[2,]    1    1    0    0
[3,]    1    0    1    0
[4,]    1    0    0    1
> XTX.INV _ solve(XTX)
> XTX.INV
     [,1] [,2] [,3] [,4]
[1,]    1   -1   -1   -1
[2,]   -1    2    1    1
[3,]   -1    1    2    1
[4,]   -1    1    1    2
> H _ X %*% XTX.INV %*% t(X)
> round(H,digits=6)
     [,1] [,2] [,3] [,4]
[1,]    1    0    0    0
[2,]    0    1    0    0
[3,]    0    0    1    0
[4,]    0    0    0    1
> #
> # Or, in a single line:
> #
> H _ X %*% solve(t(X)%*%X) %*% t(X)
> round(H,digits=6)
     [,1] [,2] [,3] [,4]
[1,]    1    0    0    0
[2,]    0    1    0    0
[3,]    0    0    1    0
[4,]    0    0    0    1
```

3. Regression exploration of the quantitative variables in the `prestige` data set; begins below...

(a) After extracting the quantitative variables with `p.sub <- prestige[,-c(5:6)]`, the correlation matrix looks like this:

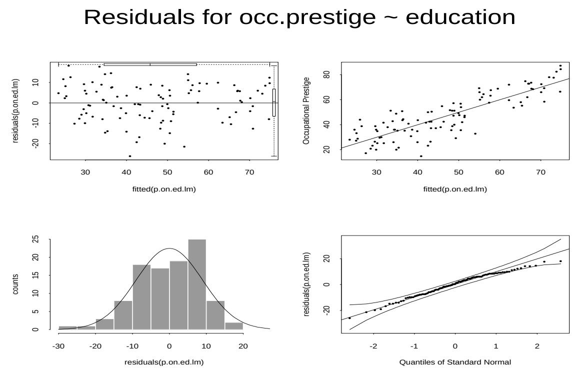
```
> cor(p.sub)
              education     income  pct.female occ.prestige
   education 1.00000000  0.5775802  0.06185286    0.8501769
      income 0.57758023  1.0000000 -0.44105927    0.7149057
  pct.female 0.06185286 -0.4410593  1.00000000   -0.1183342
occ.prestige 0.85017689  0.7149057 -0.11833419    1.0000000
```

(b) For each simple regression I executed a set of commands like this:

```
> p.on.ed.lm _ lm(occ.prestige ~ education,data=prestige)
> summary(p.on.ed.lm)
> par(oma=c(0,0,3,0),mfrow=c(2,2))
> scatbox(fitted(p.on.ed.lm), residuals(p.on.ed.lm))
> abline(h=0)
> plot(fitted(p.on.ed.lm),prestige$occ.prestige,
+ ylab="Occupational Prestige")
> abline(0,1)
> ndhist(residuals(p.on.ed.lm))
> qqn(residuals(p.on.ed.lm))
> mtext("Residuals for occ.prestige ~ education",outer=T,cex=3)
```

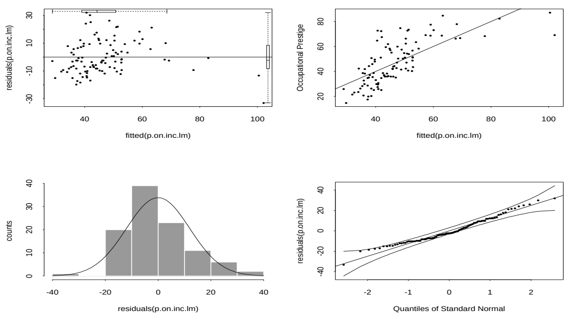
Extracts of the results follow:

```
Call: lm(formula = occ.prestige ~ education, data = prestige)
Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept) -10.7320   3.6771    -2.9186   0.0043
  education   5.3609   0.3320    16.1478   0.0000
Residual standard error: 9.103 on 100 degrees of freedom
Multiple R-Squared: 0.7228
F-statistic: 260.8 on 1, 100 degrees of freedom, p-value is 0
```



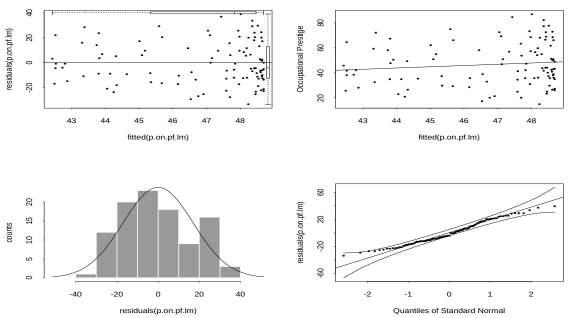
```
Call: lm(formula = occ.prestige ~ income, data = prestige)
Coefficients:
               Value Std. Error t value Pr(>|t|)
(Intercept) 27.1412  2.2677    11.9686  0.0000
     income  0.0029  0.0003    10.2243  0.0000
Residual standard error: 12.09 on 100 degrees of freedom
Multiple R-Squared: 0.5111
F-statistic: 104.5 on 1, 100 degrees of freedom, p-value is 0
```

Residuals for occ.prestige ~ income




```
Call: lm(formula = occ.prestige ~ pct.female, data = prestige)
Coefficients:
                Value Std. Error  t value Pr(>|t|)
(Intercept)  48.6930   2.3076    21.1011   0.0000
 pct.female  -0.0642   0.0538    -1.1917   0.2362
Residual standard error: 17.17 on 100 degrees of freedom
Multiple R-Squared: 0.014
F-statistic: 1.42 on 1, 100 degrees of freedom, p-value is 0.2362
```

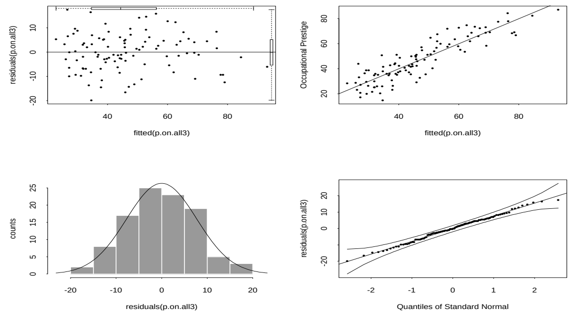
Residuals for occ.prestige ~ pct.female



(c) Using a similar set of Splus commands, beginning with `p.on.all3 <- lm(occ.prestige ~ education + income + pct.female, data=prestige)`, the results are as follows:

```
Call: lm(formula = occ.prestige ~ education + income + pct.female,
      data = prestige)
Coefficients:
                 Value  Std. Error  t value Pr(>|t|)
(Intercept)  -6.7943   3.2391    -2.0976   0.0385
  education   4.1866   0.3887    10.7708   0.0000
     income   0.0013   0.0003     4.7288   0.0000
 pct.female  -0.0089   0.0304    -0.2929   0.7702
Residual standard error: 7.846 on 98 degrees of freedom
Multiple R-Squared: 0.7982
F-statistic: 129.2 on 3, 98 degrees of freedom,  p-value is 0
```

Residuals for
occ.prestige ~ education + income + pct.female



(d) The correlations we are most concerned about in part (a) are those in the last column (or last row) of the correlation matrix. This shows that `education` has a correlation of 0.85 with `occ.prestige`, `income` has a correlation of 0.71 with `occ.prestige`, and `pct.female` has a correlation of only –0.12 with `occ.prestige`. These correlations suggest that `education` and `income` may be strongly relatd to `occ.prestige`, but it will be surprising if there is any relationship between `pct.female` and `occ.prestige` in the regresions we fit.

The results of the simple regression analyses exactly parallel our conclusions from the correlation matrix. We see from the F statistics that both $\texttt{occ.prestige} = \beta_0 + \beta_1\texttt{education} + \epsilon$ and $\texttt{occ.prestige} = \beta_0 + \beta_1\texttt{income} + \epsilon$ are better models than the simple model $\texttt{occ.prestige} = \alpha + \epsilon$: $F = 260.8$ on 1 and 100 df for `education`, and $F = 104.5$ on 1 and 100 df for `income`. On the other hand, $F = 1.42$ on 1 and 100 df for `pct.female`, suggesting that $\texttt{occ.prestige} = \beta_0 + \beta_1\texttt{pct.female} + \epsilon$ is *not* a better model than $\texttt{occ.prestige} = \alpha + \epsilon$; i.e. `pct.female` is not a good predictor of occ.prestige.

No regression analysis is complete without looking at the data, and looking at the residuals in particular. The residuals for occ.prestige ~ education look fairly normal (quantile-normal plot). Note also that there is some separation of low-education and high-education subjects evident in the y-vs-fitted plot, that might be worth examining later. The residuals for occ.prestige ~ income do not look as good, with some right-skewing visible in the histogram plot and quantile-normal plot (in the bottom row), and some curvature and/or influential income outliers (which we explored in HW #02!) visible in the residuals-vs-fitted and y-vs-fitted plots (in the top row). The residuals for occ.prestige ~ pct.female look best of the three: the residuals are more nearly normal (quantile-normal plot) with some suggestion of a heavy upper tail (histogram); on the other hand, there appears to be some left skew in pct.female (resid-vs-fitted, and y-vs-fitted), which we might fix with a transformation of some kind. We conclude that test statistics (like the F we quoted above, or t statistics for individual β_j's) can be trusted for the regressions on education and pct.female, but are not as trustworthy for the regression on income.

The numerical results for the multiple regression tell a similar story. The F statistic ($F = 129.2$ on 3 and 98 df) suggest that $\texttt{occ.prestige} = \beta_0 + \beta_1\texttt{education} + \beta_2\texttt{income} + \beta_3\texttt{pct.female} + \epsilon$ is better than the simpler model $\texttt{occ.prestige} = \alpha + \epsilon$. The residual plots now look great, with little or no pattern in the residual-vs-fitted plot in the first line, and data conforming well to a normal distribution in the quantile-normal plot. Hence the F and t statistics in this regression seem trustworthy; we conclude from the t statistics in the summary() output that, as before, education and income matter, but pct.female does not, in predicting ratings of occupational prestige.

4. From http://www.maths.uq.edu.au/~gks/data/oz/medals.html:

The data give the number of medals won by each medal-winning country in the 1992 Summary Olympic Games in Barcelona, Spain, and the 1994 Winter Olympic Games in Lillehammer, Norway. Also given is the population and latitude of each country.

```
> medals _ read.table("medals.dat",header=T)
> dim(medals)
[1] 62  4
> medals[1:5,]
             Summer Winter Population Latitude
 UnifiedTeam    112     34      231.5       61
UnitesStates    108     13      260.7       38
     Germany     82     24       81.1       51
       China     54      3     1190.4       36
        Cuba     31      0       11.1       22
```


We want to regress Summer on the other variables. So we set up a Y vector and an X matrix, as follows:

```
> X _ cbind(1,as.matrix(medals)[,-1])
> X[1:5,]
               Winter Population Latitude
 UnifiedTeam 1     34      231.5       61
UnitesStates 1     13      260.7       38
     Germany 1     24       81.1       51
       China 1      3     1190.4       36
        Cuba 1      0       11.1       22
> Y _ medals[,1]
> Y[1:5]
[1] 112 108  82  54  31
```

Now we calculate:

(a) The LS estimates $\mathbf{B}$:

```
> B _ solve(t(X)%*% X) %*% t(X) %*% Y
> B
                 [,1]
           1.77185097
    Winter 1.90565067
Population 0.05168137
  Latitude 0.07221700
```

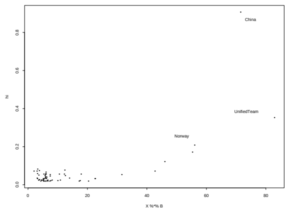
(b) The "hat matrix" $\mathbf{H}$, and the leverage measures:

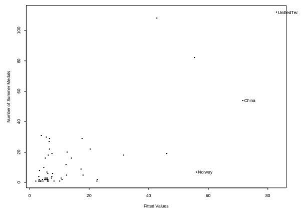
```
> H _ X %*% solve(t(X)%*% X) %*% t(X)
> hi _ diag(H)
> hi
 [1] 0.35165252 0.07234810 0.17124900 0.90656206 0.02715360
 [6] 0.03240560 0.02253288 0.02031955 0.02094471 0.01993264
[11] 0.02346633 0.04795795 0.12296841 0.04521443 0.05384466
[16] 0.03208943 0.03529650 0.02643239 0.05628479 0.02354276
[21] 0.02073511 0.08206316 0.20820534 0.03828541 0.05662789
[26] 0.02187625 0.05629750 0.07822145 0.03615005 0.05272941
[31] 0.04286126 0.05708507 0.02935695 0.05563708 0.01939255
[36] 0.05982262 0.01954302 0.03327066 0.02524568 0.02018805
[41] 0.06646035 0.02318865 0.04812498 0.02063846 0.05369644
[46] 0.03271342 0.03251135 0.02023042 0.03280477 0.01997010
[51] 0.02362964 0.07375199 0.05793705 0.07131237 0.02486767
[56] 0.02325892 0.05143341 0.03946175 0.02451128 0.07321805
[61] 0.02525482 0.03523123
>
> # Here are some ways to display the hi's
> # (not required for the problem):
>
```

```
> hi _ t(hi); dimnames(hi) _ list(NULL,dimnames(X)[[1]])
> hi

     UnifiedTeam UnitesStates  Germany     China      Cuba
[1,]   0.3516525    0.0723481 0.171249 0.9065621 0.0271536

       Hungary     France SouthKorea  Australia      Japan
[1,] 0.0324056 0.02253288 0.02031955 0.02094471 0.01993264

          Spain    Britain      Italy     Poland     Canada
[1,] 0.02346633 0.04795795 0.1229684 0.04521443 0.05384466

        Romania Netherlands  Bulgaria     Sweden
[1,] 0.03208943   0.0352965 0.02643239 0.05628479

     NewZealand NorthKorea      Kenya    Norway
[1,] 0.02354276 0.02073511 0.08206316 0.2082053

     Czechoslovakia    Denmark     Turkey   Finland
[1,]     0.03828541 0.05662789 0.02187625 0.0562975

      Indonesia    Jamaica    Nigeria    Belgium
[1,] 0.07822145 0.03615005 0.05272941 0.04286126

         Brazil    Croatia    Ethiopa       Iran
[1,] 0.05708507 0.02935695 0.05563708 0.01939255

         Latvia    Morocca    Austria   Slovenia
[1,] 0.05982262 0.01954302 0.03327066 0.02524568

        Algeria    Estonia     Greece    Ireland
[1,] 0.02018805 0.06646035 0.02318865 0.04812498

         Israel  Lithuania   Mongolia    Namibia
[1,] 0.02063846 0.05369644 0.03271342 0.03251135

     SouthAfrica Switzerland Argentina    Bahamas
[1,]  0.02023042  0.03280477 0.0199701 0.02362964

       Colombia      Ghana   Malaysia     Mexico
[1,] 0.07375199 0.05793705 0.07131237 0.02486767

       Pakistan       Peru Philippines      Qatar
[1,] 0.02325892 0.05143341  0.03946175 0.02451128

       Suriname     Taiwan   Thailand
[1,] 0.07321805 0.02525482 0.03523123
```



```
> plot( X%*%B,hi) ; identify(c(X%*%B),c(hi),labels=dimnames(X)[[1]])
[1] 4 1 23
> plot(X%*%B,y); identify(c(X%*%B),c(hi),labels=dimnames(X)[[1]],n=3,
+ pts=cbind((X%*%B)[,1],c(hi))[c(4,1,23),])
```





(c) The variance-covariance matrix of the LS estimates in $\mathbf{B}$:

```
> resids _ y - X%*%B
> RSS _ sum(resids^2)
> dim(X)
[1] 62  4
> s2 _ RSS/(62-4)
> varcovB _ solve(t(X)%*%X) * s2
> varcovB
                             Winter    Population      Latitude
           21.9622493  0.3380840138 -0.0149133988 -0.5173611707
    Winter   0.3380840  0.1051080821 -0.0007226226 -0.0174382177
Population  -0.0149134 -0.0007226226  0.0001714728  0.0001960267
  Latitude  -0.5173612 -0.0174382177  0.0001960267  0.0159426286
> #
> # As a check, make the table of coefficients and Std Errors, and
> # check against lm() output:
> #
> cbind(Coeff=B[,1],StdErr=sqrt(diag(varcovB)))
                Coeff     StdErr
           1.77185097 4.68638979
    Winter 1.90565067 0.32420377
Population 0.05168137 0.01309476
  Latitude 0.07221700 0.12626412
> summary(lm(Summer ~ Winter + Population + Latitude,
+  data=medals))$coefficients
                 Value Std. Error  t value     Pr(>|t|)
(Intercept) 1.77185097 4.68638979 0.3780844 7.067481e-01
     Winter 1.90565067 0.32420377 5.8779412 2.164530e-07
 Population 0.05168137 0.01309476 3.9467203 2.166115e-04
   Latitude 0.07221700 0.12626412 0.5719519 5.695648e-01
> #
> # So it works!
```

We see from the coefficients and their standard errors (or from the t-tests in the Splus output) that number of medals won in the winter olympics is a good predictor of number won in the summer; that population is a good predictor, with approximately 0.05 additional medals per millon additional people (that is, 1 medal per 20 million additional people), and that latitude isn't much of a predictor in the context of the other two.

5. In simple regression, show that $E[b_1] = \beta_1$.

$$
\begin{aligned}
E[b_1] &= E\left[\frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_i (X_i - \overline{X})^2}\right] \\
&\stackrel{*}{=} E\left[\frac{\sum_i (X_i - \overline{X})Y_i}{\sum_i (X_i - \overline{X})^2}\right] \\
&= \frac{\sum_i (X_i - \overline{X})E[Y_i]}{\sum_i (X_i - \overline{X})^2} \\
&= \frac{\sum_i (X_i - \overline{X})(\beta_0 + \beta_1 X_i)}{\sum_i (X_i - \overline{X})^2} \\
&= \beta_0 \cdot \frac{\sum_i (X_i - \overline{X})}{\sum_i (X_i - \overline{X})^2} + \beta_1 \cdot \frac{\sum_i (X_i - \overline{X})X_i}{\sum_i (X_i - \overline{X})^2} \\
&\stackrel{*}{=} \beta_0 \cdot 0 + \beta_1 \cdot \frac{\sum_i (X_i - \overline{X})X_i}{\sum_i (X_i - \overline{X})^2} \\
&\stackrel{*}{=} \beta_1 \cdot \frac{\sum_i (X_i - \overline{X})(X_i - \overline{X})}{\sum_i (X_i - \overline{X})^2} \\
&= \beta_1 \cdot 1 \\
&= \beta_1.
\end{aligned}
$$

where the equalities marked $\stackrel{*}{=}$ are obtained from the fact that

$$\sum_i (X_i - \overline{X}) \cdot c = 0$$

for any c (and in particular, for $c = \overline{Y}$, for $c = 1$ and for $c = \overline{X}$).