Solutions to HW 6

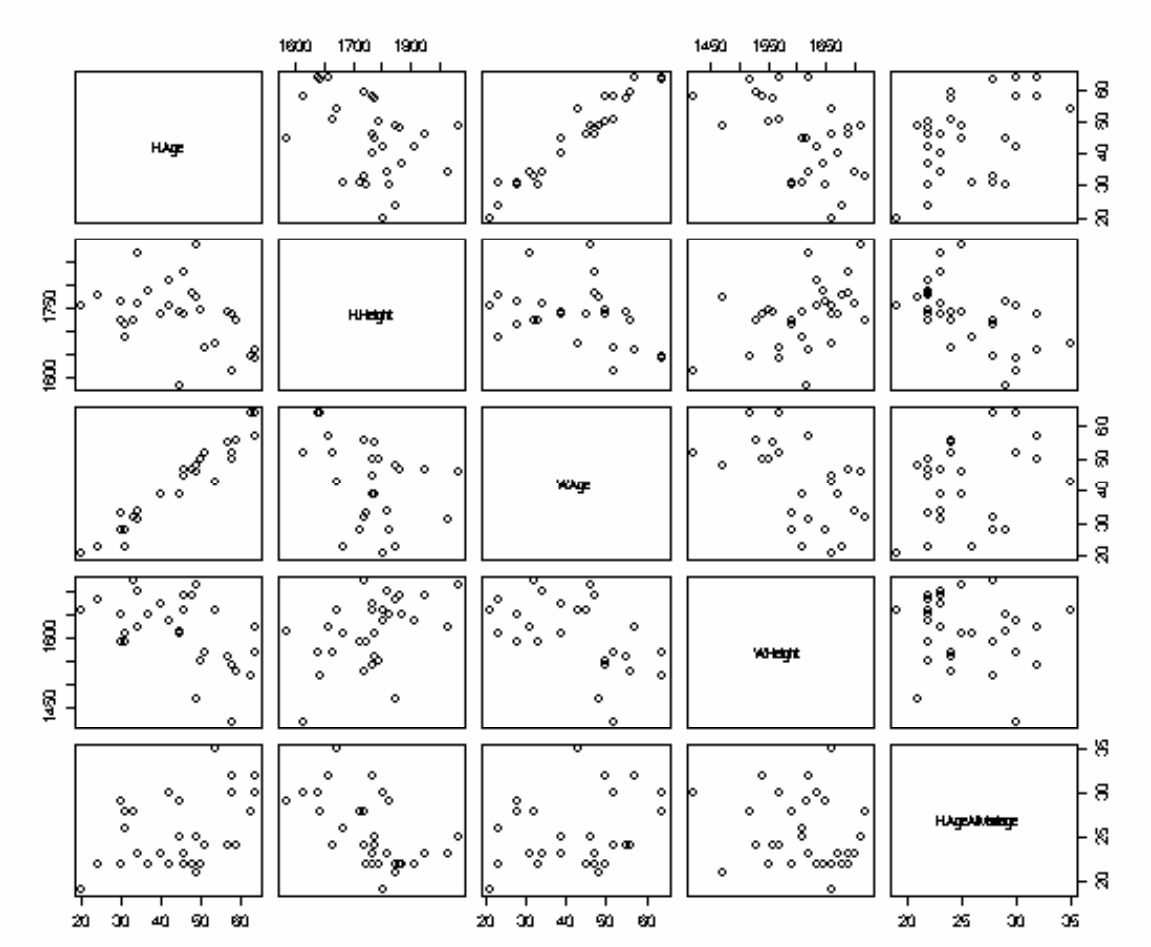
#1. See attached.

#2. There are two ways. Either you can algebraically start with X'X and take the inverse (you should know how to invert a 2 by 2 matrix; if not, look it up). Or you can show that (X'X)^(-1) times (X'X) is equal to the identity matrix (the matrix with 1's on the diagonal and 0's off the diagonals.)

#3. R-squared is the proportion of variation due to the regression divided by the total variation: (SYY - RSS)/SYY = 3938/(3938+289.9) = .9313

The estimate of the standard error of the residuals is sqrt(RSS/(n-2)) = sqrt(289.9/24) = 3.475 (Note: although there are n=30 observations, there are missing values so that we have only 26 observations in the regression.)

#4)



Husbands and wifes age shows a strong positive linear association: older men tend to be married to older women, and vice versa. Husband's and wife's heights are positively associated, though not quite as strong.; taller people tend to marry each other, shorter

people tend to marry each other. There might be a negative association between the heights and ages variables, though these associations are too weak to see very clearly on these graphs.

b) Is there evidence of a historic trend in the age at which people marry? If we look at the plot of age at marriage against current age (for husbands), we see some evidence of a slight positive association. It apears that those who married later (in their thirties and later) tend to be older now (late fifties). This suggests that 20-30 years ago men in GB were marrying at a later age. However, this trend is very weak, and might not be real. This is one thing the regression analysis can help us with. And it is not clear the data support such a conclusion. The data consist only of married people, and so younger people (at the time the data were collected) might not have had enough time to get married. Some of you pointed out that any conclusion we reach could only be safely applied to men, although perhaps if we accept that husband's and wife's age have a strong correlation, we could extend this to women too.

c)

```
> naive <- lm(H.AgeAtMarriage ~ ., data=couple.table)
> summary(naive)

Call:
lm(formula = H.AgeAtMarriage ~ ., data = couple.table)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2853 -1.8032 -0.3124  1.4199  5.3207

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 45.004253  17.362308   2.592 0.017009 *
H.Age        0.730858   0.148619   4.918 7.28e-05 ***
H.Height    -0.026297   0.009168  -2.868 0.009202 **
W.Age       -0.645818   0.151304  -4.268 0.000342 ***
W.Height     0.012861   0.008077   1.592 0.126252
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 `
' 1

Residual standard error: 2.457 on 21 degrees of freedom
Multiple R-Squared: 0.6851,   Adjusted R-squared: 0.6251
F-statistic: 11.42 on 4 and 21 DF,  p-value: 4.415e-05
```

Assuming all of the usual assumptions are valid, the H.Age parameter tells us that among men of the same height whose wives were the same age and height, those husband's who were a year older tended to marry .73 years later, on average. Put less cryptically, all things being equal, those who are older now tended to marry older (which confirms the slight positive trend we saw in (b).

MANY of you gave what I would call a causal interpretation: for each additional year of a husband's age, the age at marriage goes up by .73 years. Interpreted literally this is a silly statement. The age at which you were married can't change (unless you divorce and remarry!) More generally, beware of suggesting, even hinting, that changes in x will cause a change in y unless the data come from a controlled (and well designed) study.

Finally, note that the regression equation is all about means, and trends. So we are not saying that husband's who differ in age by one year differ by exactly .73 years in the age they were married. We're saying that this is the trend -- on average this is how these two groups differ.

(d) One reason I asked about thisis that these variables show a strong positive association, and so one might expect them to have the same sign in this equation. However, keep in mind that the interpretation of the slopes is conditioned on the other variables being present in the equation. So the negative sign in front of W.Age means that, *all things being equal,* (in particular, among men who are the same age) men whose wives are older than average now tended to be younger when they married. The fact that H.AGe and W.Age have different signs suggests that to predict the age at which a randomly selected husband got married, the difference between his current age and his wife's current age might also be useful.

(e) The wife's height does not seem useful; the large p-value suggests that the slope is 0. We can try another model without it.

```
> naive1 <- lm(H.AgeAtMarriage ~ . -W.Height,
data=couple.table)
> summary(naive1)

Call:
lm(formula = H.AgeAtMarriage ~ . - W.Height, data =
couple.table)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4263 -1.9202 -0.1391  1.9722  5.2473

Coefficients:
```

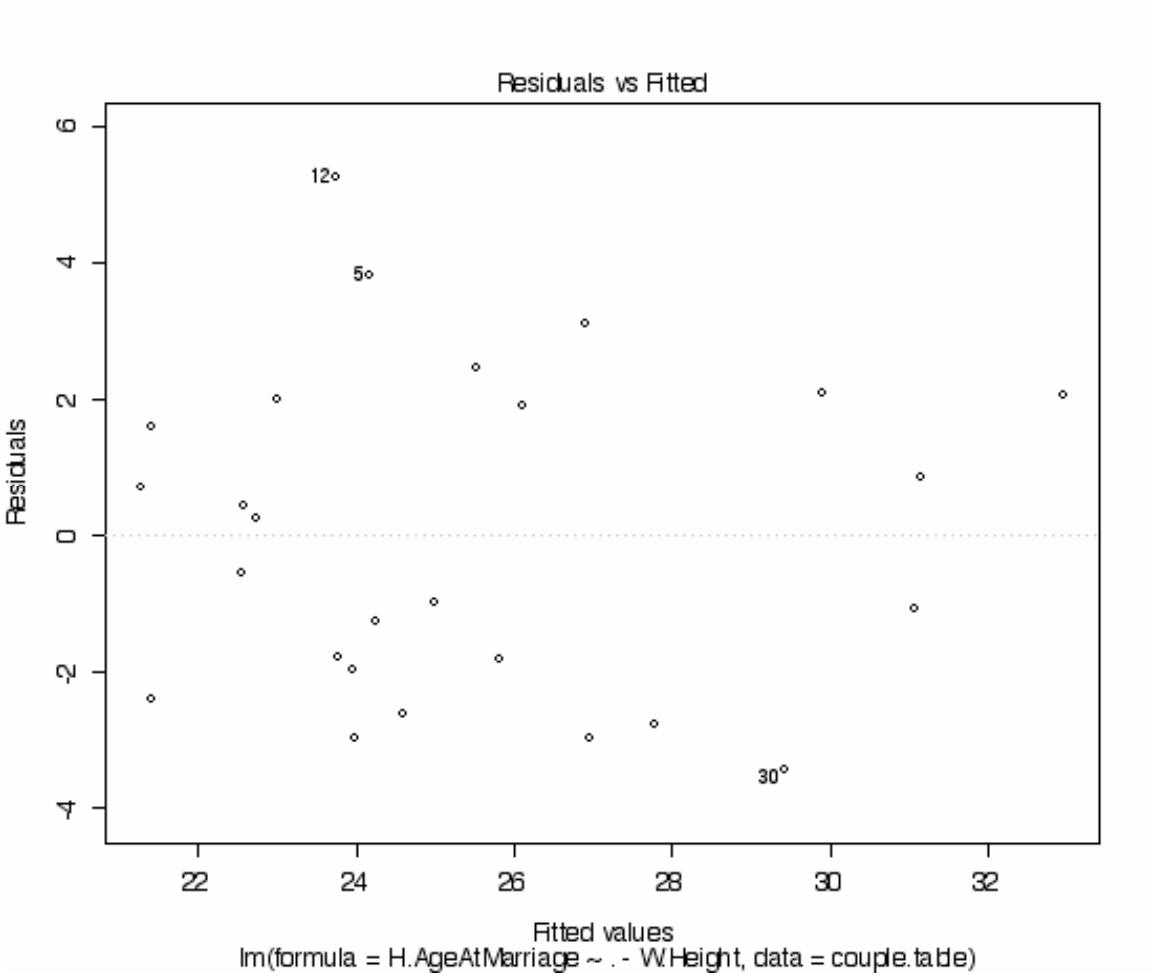
```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 57.305133  16.082530   3.563 0.001739 **
H.Age        0.719781   0.153549   4.688 0.000113 ***
H.Height    -0.020736   0.008768  -2.365 0.027268 *
W.Age       -0.663138   0.156090  -4.248 0.000329 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 `
' 1

Residual standard error: 2.541 on 22 degrees of freedom
Multiple R-Squared: 0.647,    Adjusted R-squared: 0.5989
F-statistic: 13.44 on 3 and 22 DF,  p-value: 3.371e-05
```

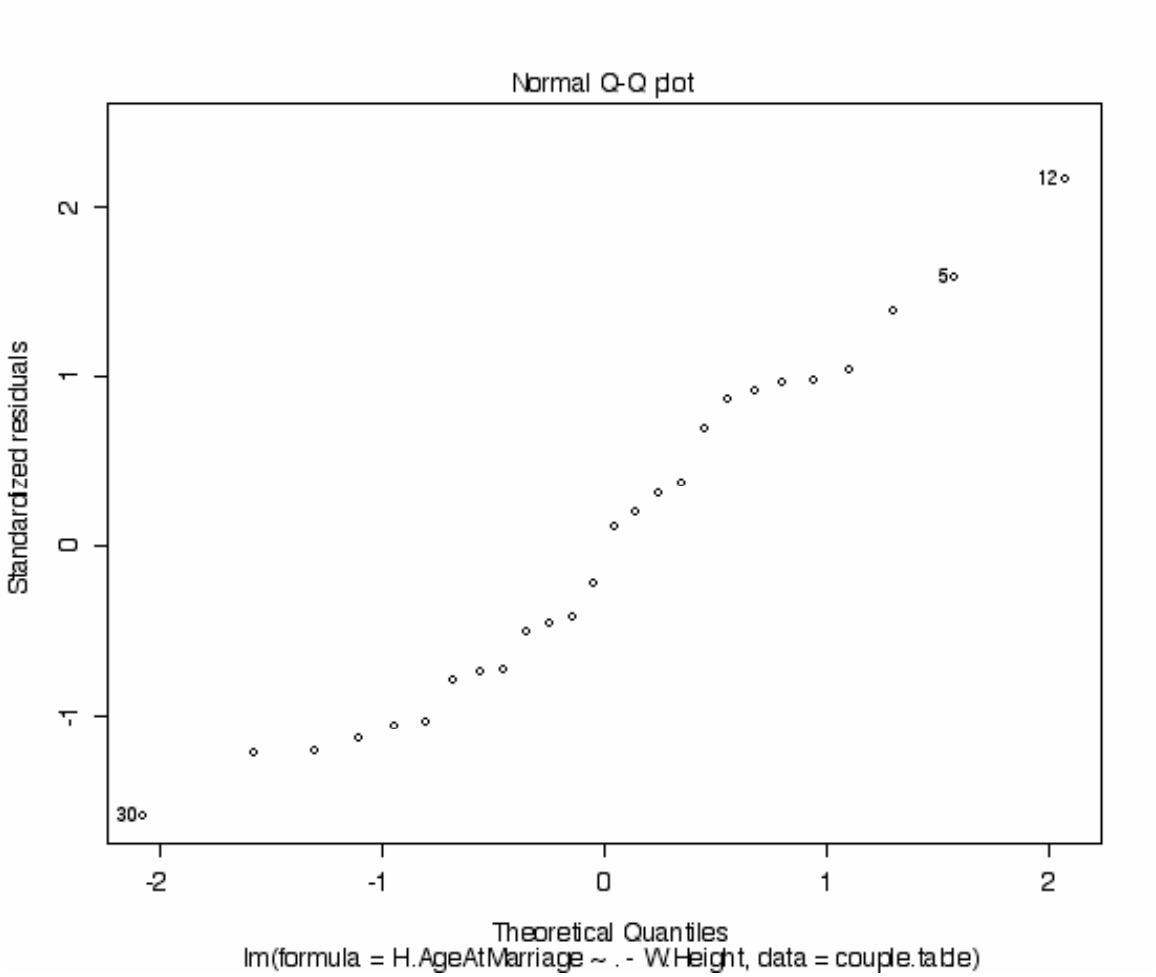
(f) H.ageatmarriage = 57 + .72*H.Age - .02 * H.Height + -.66 W.Age

Among all age groups, taller men tend to marry when younger. If we make comparisons among men of the same height whose wives are the same age , then those that are older now tended to marry later. Among men the same height and age, those whose wives are older tended to marry younger.

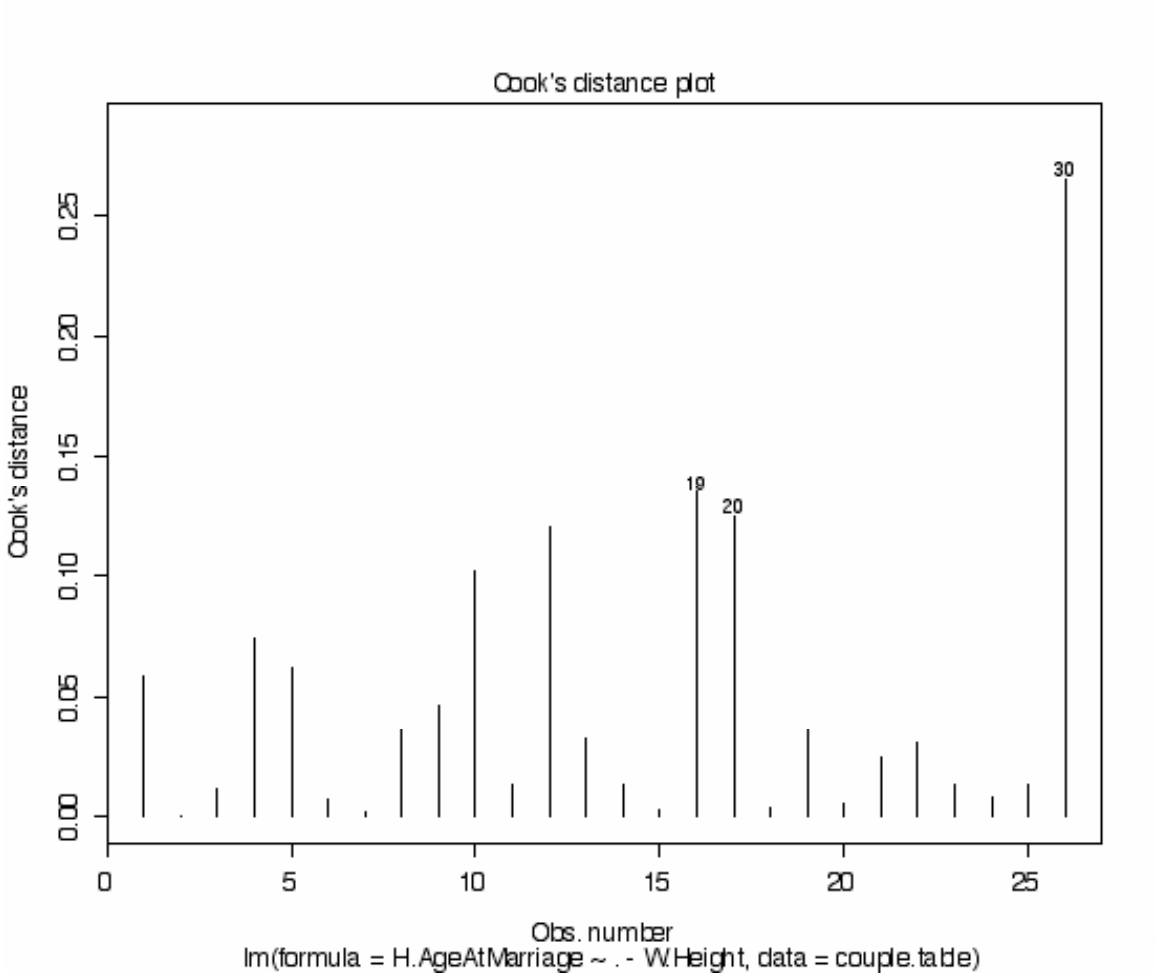
The residual plot does not look bad. The assumption of a linear relation does not seem to be bad.



The normality ssumption is also not too far off



Couple 30 seems to be influential to the model, but not that much more influential than other couples:



The assumption of independence is satisfied as long as this was truly a random sample from the population. It would fail if, for example, couples were sampled within the same families.

Here's another model that fits the difference in age between husband and wife. I'll show you all the R commands so you can see what I did. First, I created a new variable that subtracted the wife's age from the husband's.

```
> diff.age <- H_Age - W_Age
> hist(diff.age)
> summary(diff.age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 -3.000   0.000   1.000   2.346   3.000  11.000   4.000
```

Typically, husband's are about 1 to 2 years older than their wife, but there are wife's older than their husband. (But note that the distribution is right-skewed.)

```
> fit1 <- lm(H_AgeAtMarriage~diff.age + H_Age + W_Height + H_Height)
> summary(fit1)
```

```
Call:
lm(formula = H_AgeAtMarriage ~ diff.age + H_Age + W_Height +
    H_Height)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2853 -1.8032 -0.3124  1.4199  5.3207

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 45.004253  17.362308   2.592 0.017009 *
diff.age     0.645818   0.151304   4.268 0.000342 ***
H_Age        0.085040   0.045867   1.854 0.077832 .
W_Height     0.012861   0.008077   1.592 0.126252
H_Height    -0.026297   0.009168  -2.868 0.009202 **
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 2.457 on 21 degrees of freedom
Multiple R-Squared: 0.6851,  Adjusted R-squared: 0.6251
F-statistic: 11.42 on 4 and 21 DF,  p-value: 4.415e-05
```

Diagnostic plots show that this model fits pretty well. Note that the wife's height doesn't seem important. And note that we learn that all things being equal, each year older the husband is than the wife is associated with an average .6 years increase in age at marriage.

Let's take out wife's height, since it is not statistically significant.

```
> summary(fit2)

Call:
lm(formula = H_AgeAtMarriage ~ diff.age + H_Age + H_Height)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4263 -1.9202 -0.1391  1.9722  5.2473

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 57.305133  16.082530   3.563 0.001739 **
diff.age     0.663138   0.156090   4.248 0.000329 ***
H_Age        0.056643   0.043708   1.296 0.208430
H_Height    -0.020736   0.008768  -2.365 0.027268 *
```

Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 2.541 on 22 degrees of freedom
Multiple R-Squared: 0.647,   Adjusted R-squared: 0.5989
F-statistic: 13.44 on 3 and 22 DF, p-value: 3.371e-05

This has affected the husband's age variable. It now looks as if knowing the difference in age is sufficient for predicting age at marriage. Again, the diagnostic plots show the assumptions involving normality of errors, constant variance, and linearity seem fairly safe.

Let's remove Age.

lm(formula = H_AgeAtMarriage ~ diff.age + H_Height)

Residuals:
    Min   1Q Median   3Q   Max
-4.6465 -1.5174 -0.4705 2.1265 4.5457

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 67.113698 14.398289   4.661 0.000108 ***
diff.age   0.703122   0.155254   4.529 0.000151 ***
H_Height   -0.024981   0.008252 -3.027 0.005996 **

Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 2.578 on 23 degrees of freedom
Multiple R-Squared: 0.6201, Adjusted R-squared: 0.587
F-statistic: 18.77 on 2 and 23 DF, p-value: 1.467e-05

This model fits the best, at least as far as the diagnostic plots are concerned. It tells us that the age at which a man was married depends on how much younger his wife is: men who married younger wifes tended to get married at a later age. Also, assuming that the age gap is the same, taller husband's tended to marry at an earlier age.

This interpretation above sounds like it has more meaning than it really does. Suppose a wife was, say, 10 years younger than the husband. The average marrying age is 25, and so unless he married his wife when she was 15, he is likely to be older than average!