Name: _____

1 Objective: Modeling, missing data values and data imputation

The data set `hus_wif_all.dat` contains 5 columns of tab separated numbers, with missing values donated with " * ". The columns are: husband's age [yrs], husband's height [mm], wife's age [yrs], wife's height [mm], husband's age at time of marriage.

2 Key Words

Least square regression, missing data, data imputation.

3 Modeling steps

TURN IN: Write a paragraph in which you answer the following questions.

(1) Fix the data so that SAS can read it.

(2) Create a data set `HuWi` with columns `ha hh wa wh ham`

(3) Find the mean and standard deviation and number of missing values for `ha` and for `wa`.

(4) Find the correlation between `ha` and `wa` , between `hh` and `wh`.

(5) Find the least squares regression line which uses response `wh` and predictor `hh`.

   (a) Compare the R^2 value for this with the correlation found in (4)

   (b) Would you say that short people only marry short people?

   (c) If you want to get a plot of predicted values, you could use code:

```
output out = lin p = plin;
```

   What columns are in lin?

(6) Get a scatter plot of the `wh` vs `hh`. Do you think a quadratic in `hh` would fit the `wh` data better?

(7) What is R^2 if `wh` is fitted using the most general quadratic in all the other 4 variables? Did this model fit the data much better than `hh` alone?

Finish any unfinished lab.