HG
Revised Sept. 2019

ECON 4130 H19

Extra exercises for no-seminar week 41

(Solutions will be put on the net on the Thursday, 10 Oct., of the week)

On modelling relationships that show how one rv depends on another rv

As illustration we will use data taken from Aitchison, J. (1986), *The statistical analysis of composite data*, London: Chapman and Hall. The data set can be downloaded as a STATA data set from my personal webpage, http://folk.uio.no/haraldg , by clicking on "*Hong Kong Consumer Data*" under Econ 4130. (After having downloaded the data use the command `use` to read the data into STATA – or, simply use the open-file icon in STATA and browse to the data-file.)

Note. *Access to Stata: Log in at the web page,* kiosk.uio.no, *and find Stata under Analyse.*

The following exercise can be seen as an introduction to basic regression analysis from a theoretical point of view with main focus on model concepts and less focus on formal inference procedures. It is illustrated by a concrete data set. The exercise is self contained in the sense that, even for students who have never been in touch with regression analysis, all questions should be possible to answer based on what is said in the lectures - and some common sense in addition. Although not necessary, those who have no experience with regression from before, may find it useful to consult the lecture note "*Regresjon I*" (in Norwegian) on the web-page of Econ 2130 for 2017, where similar concepts are discussed with main focus on the data and less on models (i.e., a descriptive approach).

The data come from a Hong Kong survey of household expenditure and give the expenditure of 20 single men (M) and 20 single women (W) on four commodity groups. The units of expenditure are Hong Kong dollars (HKD), and the commodity groups are as follows:

1. Housing, including fuel and light
2. Foodstuffs, including alcohol and tobacco
3. Other goods, including clothing, footwear and durable goods
4. Services, including transport and vehicles

A. In the beginning of the exercise we will focus on commodity group 4 (services) for women. We want to study how the expenditure (which we call Y) depends on income (here, for simplicity, expressed by X = total expenditure = the sum of expenditures in the four groups).
We want to model the relationship between X and Y, looking at Y as a response variable and X as an explanatory variable.

The observations are given in table 1.

Table 1. Expenditure on services (Y) and total expenditure (X), women

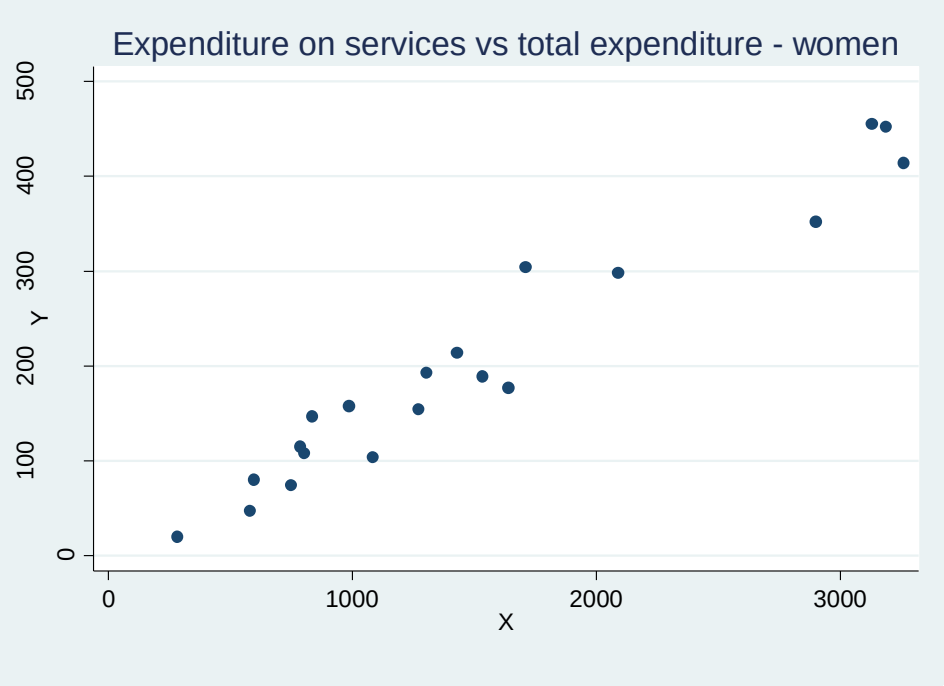
Exp. Y	154	20	455	115	104	193	214	80	352	414
Total exp X	1271	284	3128	786	1084	1303	1428	596	2899	3258
Exp. Y	47	452	108	189	298	158	304	74	147	177
Total exp X	581	3186	804	1533	2088	986	1709	748	836	1639

The total expenditure data are not in the data set, so I calculated them with the STATA command *generate*. *Do this yourself (both for women and men)*.

In situations with only one explanatory variable it is always a good idea to plot the data to help formulate a model. In fig. 1 I have plotted *Y* against *X* for the data.

(Use e.g. the `Graphics -> Twoway graph -> create -> scatter` menu to reproduce fig 1)

Figure 1



The plot indicates among other things a linear type of relationship with a strong positive correlation.

The first step in the modelling process is to establish the sampling procedure in model terms, and assume representativity of the data. Representativity will be ensured if we can assume that the data come from a *simple random sample* (i.e., where all units in the population have the same chance of being chosen), drawn from the population of single female households in

Hong Kong. Assuming this, and considering that the population is relatively large, we may assume that the data are observations of n = 20 independent and identically distributed (*iid*) random pairs, $(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)$, all with a common distribution with pdf, $f(x,y)$, which we may call *the population distribution*, and which represents the distribution of (X,Y) in the total population. (Note that when we talk about independent pairs, we mean that only observations from different pairs (i.e., different consumers) are independent, while observations from the same pair may be dependent[1].) In short[2]:

(1) (X_i,Y_i) are *iid* pairs, and $(X_i,Y_i) \sim f(x,y) \qquad i=1,2,\ldots,20$

If $(X,Y) \sim f(x,y)$, we may now estimate certain characteristics of *f*. In particular the following 5 parameters will be needed below:

$$\mu_X = E(X), \quad \mu_Y = E(Y), \quad \sigma_X^2 = \text{var}(X), \quad \sigma_Y^2 = \text{var}(Y), \quad \sigma_{XY} = \text{cov}(X,Y)$$

which, as usual (also described in the basic statistic course), are estimated by

$$\hat{\mu}_X = \bar{X} = 1507.35, \qquad \hat{\mu}_Y = \bar{Y} = 202.75$$

$$\hat{\sigma}_X = S_X = \left[\frac{1}{n-1}\sum_i (X_i - \bar{X})^2\right]^{\frac{1}{2}} = 935.3827, \qquad \hat{\sigma}_Y = S_Y = \left[\frac{1}{n-1}\sum_i (Y_i - \bar{Y})^2\right]^{\frac{1}{2}} = 133.0358$$

$$\hat{\sigma}_{XY} = S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = 121017$$

(i) *Check these results yourself by* STATA [use *Statistics* -> …..-> *Summary statistics*].

We are also interested in the correlation coefficient, $\rho = \text{corre}(X,Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$, that is usually estimated in the natural way

$$\hat{\rho} = \frac{S_{XY}}{S_X S_Y} = 0.9725$$

showing a strong positive relationship in this case.

---

[1] More precisely: That eg. two pairs, (X_1,Y_1), (X_2,Y_2) are *iid* pairs, means that, if $g(x_1,y_1,x_2,y_2)$ is the joint pdf for all 4 rv's, X_1,Y_1,X_2,Y_2, then $g(x_1,y_1,x_2,y_2) = f(x_1,y_1)\cdot f(x_2,y_2)$, where $f(x_i,y_i)$ is the marginal joint pdf for (X_i,Y_i), i = 1, 2 (see, e.g., Rice page 77 for joint pdf's for more than two rv's). Thus, the two different pairs are independent of each other (and have the same joint distribution) while X_i,Y_i within the same pair may be dependent.

[2] Note that this is an assumption of *representativity* of the sample. A non representative sample would, e.g., occur if the data were sampled from a single district in Hong Kong leading to iid pairs, (X_i,Y_i) $i = 1,2,...,n$ drawn from a distribution $g(x,y)$ which is not necessarily equal to the population distribution $f(x,y)$.

(ii) *Find the corresponding estimates for men and make a plot corresponding to figure 1 for men. Any striking differences between the results for men and women?*

[**Note** that you can copy a STATA graph for pasting into e.g. a Word document by right-clicking on the graph and then Copy or using one of the icons on top.]

B. Modelling the relationship: A disadvantage of ρ as a measure of relationship in this context is that it treats *X* and *Y* in the same symmetrical way without considering which of them is the response variable and which is the explanatory variable. It turns out that the difference in meaning between response and explanation can be expressed more naturally when we use the conditional distribution, $f(y \mid x)$, as a basis for studying the relationship. Actually, it appears more natural to look at $f(y \mid x)$ considering *X* to be explanatory since it represents the distribution of *Y* when *X* is kept fixed at a chosen value *x*. In particular, it is common to focus interest on the expectation and variance of $f(y \mid x)$, which are both functions of *x*, i.e.,

$$\mu(x) = E(Y \mid x) = \int_{-\infty}^{\infty} y f(y \mid x) dy \qquad \text{and} \qquad \tau^2(x) = \text{var}(Y \mid x) = \int_{-\infty}^{\infty} (y - \mu(x))^2 f(y \mid x) dy$$

Since the main interest is on these functions, it is usually more natural to model these directly, instead of, as we did in the lecture, model first $f(x, y)$ and then derive $E(Y \mid x)$ and $\text{var}(Y \mid x)$[3].

For example, in the present situation, looking at figure 1, it seems reasonable to postulate that $E(Y \mid x)$ is a linear function of *x*, and that $\text{var}(Y \mid x)$ is constant for all *x*. Our model is therefore

(2) $E(Y \mid x) = \alpha + \beta x$

(3) $\text{var}(Y \mid x) = \tau^2$

where α, β, τ^2 are unknown parameters (to be estimated by the data). This is an example of a *regression model*, in fact the simplest case of such a model when (1) holds as well. A regression model where the conditional variance is assumed to be constant (as here) is called *homoscedastic* in econometrics, or *heteroscedastic* if the conditional variance is allowed to vary with *x*.

Now it turns out that α, β, τ^2 are determined by the five parameters in **A**, $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \sigma_{XY}$ as follows:

(4) $\beta = \dfrac{\sigma_{XY}}{\sigma_X^2}, \qquad \alpha = \mu_Y - \beta \mu_X$

[3] An important exception to this approach is when we are in a situation where it is reasonable to assume that *X* and *Y* are *jointly* normally distributed. Then the model (2)-(3) is automatically fulfilled. See example C in section 3.5.2 in Rice (without bothering too much about the messy algebra).

(5) $\tau^2 = \text{var}(Y \mid x) = \sigma_Y^2(1-\rho^2)$, where $\rho = \dfrac{\sigma_{XY}}{\sigma_X \sigma_Y}$

(i) *Prove (4) and (5), using Theorem A and B in section 4.4.1, (i.e.,* $EY = E\big[E(Y \mid X)\big]$ *and* $\text{var}(Y) = E\big[\text{var}(Y \mid X)\big] + \text{var}\big[E(Y \mid X)\big]$ *respectively).*

[**Hint:** Replacing *x* by the rv *X* in the function $\mu(x) = E(Y \mid x)$, gives the rv $\mu(X) = \alpha + \beta X$. Taking the expectation of this, we get from Th. A that $\mu_Y = E(Y) = E\mu(X) = E(\alpha + \beta X) = \alpha + \beta\mu_X$, proving the second relationship in (4).

To prove the first one, we need to show that $\sigma_{XY} = \beta\sigma_X^2$. We have, replacing μ_Y by $\alpha + \beta\mu_X$, that $\sigma_{XY} = E(XY) - \mu_X\mu_Y = E(XY) - \alpha\mu_X - \beta\mu_X^2$. Now, using Th.A again, we get $E(XY) = E\big[E(XY \mid X)\big] = E\big[X \cdot E(Y \mid X)\big]$.

(**More details:** The last equality follows by remembering that two steps are involved. On step 1 we extract the underlying function $g(x) = E(XY \mid x) = E(XY \mid X = x) = E(xY \mid X = x) = E(xY \mid x) = x \cdot E(Y \mid x)$ remembering that the small *x* can be treated like a constant since it represents a fixed number. The value of the small *x* can be chosen in any way. In particular, the relation is valid if *x* is chosen as an observation of the rv *X*, which means that $g(X) = X \cdot E(Y \mid X)$ is also valid when we are replacing *x* by *X* – which is step 2.)

Hence we have
$E(XY) = E\big[X \cdot E(Y \mid X)\big] = E\big[X(\alpha + \beta X)\big] = \ldots$ and so on. *Complete the argument.*

For (5) use theorem B:
$\sigma_Y^2 = \text{var}(Y) = E\big[\text{var}(Y \mid X)\big] + \text{var}\big[E(Y \mid X)\big] = E\big[\tau^2\big] + \text{var}\big[\alpha + \beta X\big] = \ldots$ and so on.
Complete the argument using general properties of E and Var.]

(ii) *Show also that* $\rho = \text{corre}(X,Y) = 0 \;\Rightarrow\; \beta = 0$ (i.e., we then get a flat regression function).

(iii) *What happens if* $\rho = \pm 1$? [**Hint:** Look at $\text{var}(Y \mid x)$.]

C. A natural way to estimate (2) and (3) is now simply to utilize (4) and (5) and substitute the estimates we found for $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \sigma_{XY}$ in **A**. I.e.,

(6) $\hat{\beta} = \dfrac{S_{XY}}{S_X^2} = 0.1383, \qquad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = -5.7384$

(7) $\hat{\tau}^2 = S_Y^2(1-\hat{\rho}^2) = 960.118$ (or conditional standard deviation: $\hat{\tau} = 30.9858$)

It turns out that these estimates are identical with the so called OLS (ordinary least squares) estimates[4] you learn about in econometrics (and presently in Stat 1). It can be proven that the corresponding estimators have good statistical properties in the present model, like, e.g., that they are unbiased, approximately normally distributed, optimal in terms of variance etc. The estimates are also equal to the estimates you get when you run a simple regression in STATA (use e.g., the command: *regress* F4 XF (where F4 corresponds to Y in the data set and XF is total expenditure (women) - or whatever you wish to call that variable). *Confirm this!*

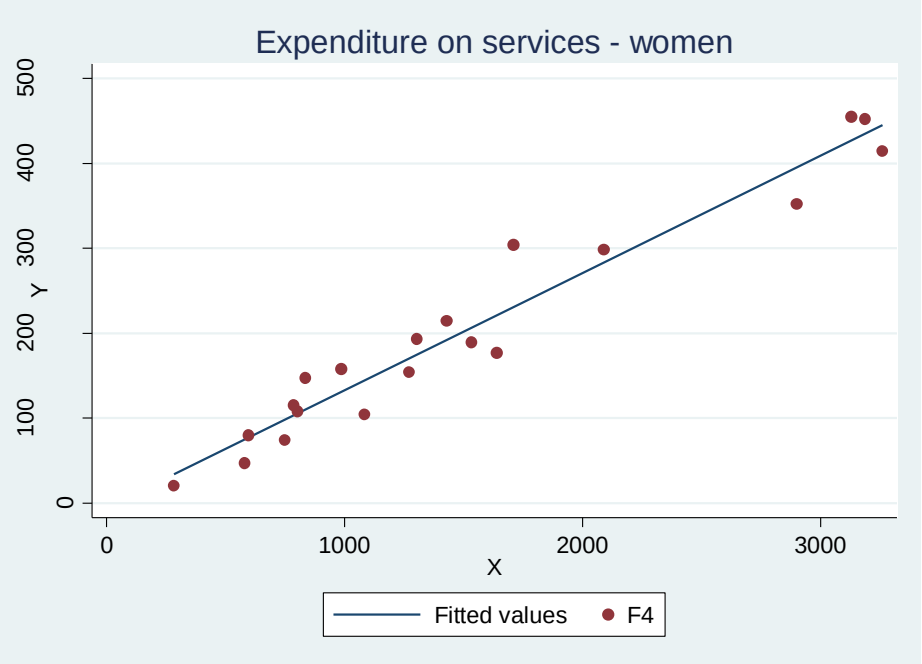
You find the estimates of α, β under the heading *Coef* in the output. You may get slightly different results due to rounding effects - STATA calculates with higher precision than most calculators…. The estimate of τ you find in the STATA output under heading "Root MSE". Notice a slight difference from the result above because of rounding errors.

The estimated regression line is

(8) $\hat{\mu}(x) = -5.7384 + 0.1383 \cdot x$

which is plotted in the scatter plot in figure 2. (I used the menu: `graphics -> twoway graph`. Then, inside the menu, use create twice, first `create -> scatter...` , and then `create -> fit plots -> linear predictions...`).

Figure 2



[4] Defined as those values of *a* and *b* that minimize the sum of squares, $Q(a,b) = \sum_{I=1}^{n} (Y_i - a - bX_i)^2$ leading to the same formulas as in (6) and (7).

(i) *Do the same for men using STATA output, i.e. estimate the regression line, the conditional standard deviation, and draw the fitted line in the scatterplot.*

D. We will now reformulate the model in (2)-(3) in a way that lends itself to some further interpretation: Define the rv ε (epsilon) by

(9) $\varepsilon = Y - E(Y \mid X) = Y - \alpha - \beta X$

The rv ε is called an error term ("restledd") and measures that remaining part of *Y* which is *not* explained by *X*.

(i) *Show that ε satisfies (10)-(12) below*

(10) $E(\varepsilon \mid x) = 0$ and $\text{var}(\varepsilon \mid x) = \text{var}(Y \mid x) = \tau^2$

which by Th. A and B of section 4.4.1 implies

(11) $E(\varepsilon) = 0$ and $\text{var}(\varepsilon) = \tau^2$

(12) $\text{cov}(\varepsilon, X) = 0$ [same argument as in the hint of **C**]

Hence we can decompose the response $Y = E(Y \mid X) + \varepsilon$ into an explained part, $\mu(X) = E(Y \mid X)$, and an unexplained part, ε, or, in other words, the model (2)-(3) can be reformulated as

(13) $Y = \alpha + \beta X + \varepsilon$

where the error term ε is uncorrelated with *X*, has expectation 0, conditional expectation 0 given $X = x$, and constant conditional variance given $X = x$.

Similarly, using the variance relationship in the hint of **C,** we can decompose the total variance of *Y*, σ_Y^2, into the variance of the explained part, $\text{var}\left[E(Y \mid X)\right]$, plus the variance of the unexplained part:

(14) $\text{var}(Y) = E\left[\text{var}(Y \mid X)\right] + \text{var}\left[E(Y \mid X)\right] = \tau^2 + \text{var}\left[E(Y \mid X)\right] = \text{var}(\varepsilon) + \text{var}\left[E(Y \mid X)\right]$

A natural measure of the strength of *X* as an explanatory variable of *Y* would then be the ratio $\text{var}\left[E(Y \mid X)\right] / \text{var}(Y)$, which shows how large part of the variance of *Y* is explained by *X* (through the variance of the explained part *Y*, i.e., $E(Y \mid X)$).

(ii) *Show that*

(15) $$\frac{\text{var}\left[E(Y \mid X)\right]}{\text{var}(Y)} = \rho^2 \quad \textit{for the model in } (2)\text{-}(3)$$

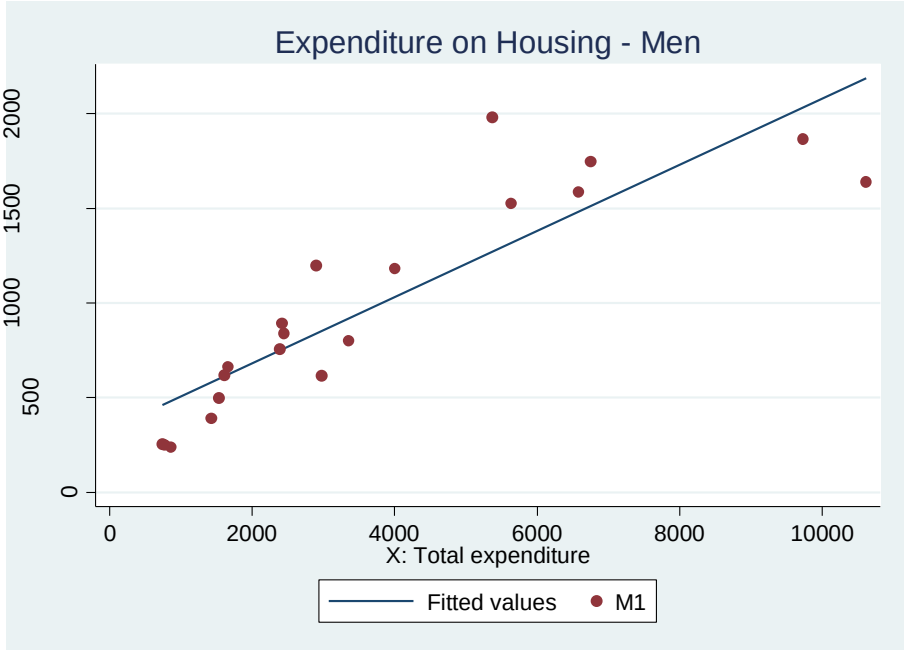
Thus (15) provides an alternative interpretation of ρ. Since in our example $\hat{\rho}^2 = 0.946$, we have reason to say that *X* explains 94,6% of the variation of *Y* in the data, indicating a strong relationship. In STATA this quantity appears under the heading "R-squared". Some programs call it "Coefficient of determination".

(iii) *Find the corresponding R-squared measure for men .*

(iv) *Why do you have to use capital X in the numerator of (15)? What happens if you use small x instead, i.e., what is the value of* $\text{var}\left[E(Y \mid x)\right]$ *?*

E. We will also look at a case where the model (2)-(3) appears less credible. In figure 3 I have plotted expenditures on commodity group 1 (housing) for men (i.e., our *Y*, which is called M1 in the data) versus total expenditure for men (*X*). In addition, I have fitted the OLS regression line as in figure 2 above. We see that, in the central area where roughly 4000 < *X* < 9000, the line lies below all observations. This does not seem reasonable considering that $E(Y \mid x)$ should capture (estimate) the average value of *Y* in the population when *X* is fixed to a value *x*. In other words, the *linear* regression line appears to underestimate the mean expenditure for given values of *X* in the central area.

Figure 3



Since the shape of the plot looks more like a section of a parabola, a simple reformulation of the model might improve the situation, i.e., replace (2)-(3) with the specification

(16) $E(Y \mid x) = \alpha + \beta_1 x + \beta_2 x^2$

(17) $\text{var}(Y \mid x) = \tau^2$

where $\alpha, \beta_1, \beta_2, \tau$ are parameters. Again it is possible to derive relationships between these parameters and moments in $f(x, y)$ as above. This is, however, somewhat more complicated now (you will need moments up to order 4), so we will skip that. Luckily it is quite unnecessary since it turns out that we can estimate (16)-(17) directly with linear regression methods as implemented in STATA. This we can always do when the regression function is *linear in the parameters* as here. Note that, although $E(Y \mid x)$ in (16) is non-linear in *x* (a parabola), it is actually linear in the unknown parameters, α, β_1, β_2 ! The trick in STATA (and other regression programs as well) is to replace the single variable, *X*, with two new variables, $X_1 = X$ and $X_2 = X^2$ (the last one you can make with the *generate* command). The STATA then interprets (16) as

$$E(Y \mid x_1, x_2) = \alpha + \beta_1 x_1 + \beta_2 x_2$$

which is simply estimated by the command: regress Y X1 X2,
or, using my variable names, M1 for *Y*, XM for *X*, and XM2 for X^2,
the command becomes: regress M1 XM XM2

This produces the following output:

```
      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  2,    17) =   88.16
       Model |  5684365.51     2  2842182.75           Prob > F      =  0.0000
    Residual |  548080.693    17  32240.0408           R-squared     =  0.9121
-------------+------------------------------           Adj R-squared =  0.9017
       Total |   6232446.2    19  328023.484           Root MSE      =  179.56

------------------------------------------------------------------------------
          M1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          XM |   .4430856   .0544811     8.13   0.000     .3281404    .5580307
         XM2 |  -.0000255   5.00e-06    -5.10   0.000     -.000036   -.0000149
       _cons |  -109.4593   108.8407    -1.01   0.329    -339.0931    120.1745
------------------------------------------------------------------------------
```

The estimates of the regression coefficients are found under "Coef", and "Root MSE" estimates τ . Hence the estimated model becomes

(18) $\hat{E}(Y \mid x) = -109.4593 + 0.4430856 \cdot x - 0.0000255 \cdot x^2$

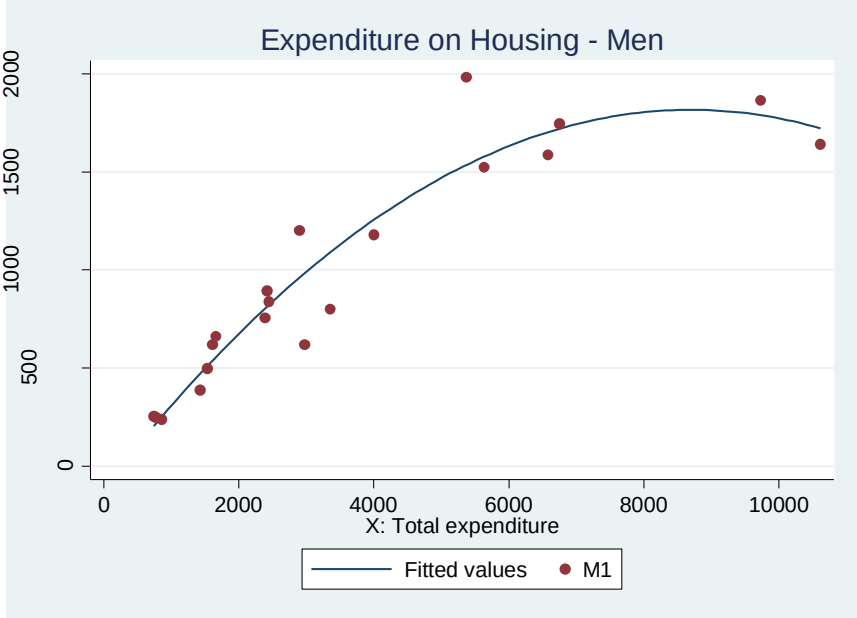
(19) $\sqrt{\hat{\text{var}}(\text{Y}|\text{x})} = \hat{\tau} = 179.56$

As above, the "R-squared" estimates the ratio, $\text{var}\left[E(Y \mid X)\right]/\text{var}(Y)$. Hence, from the output, 91.2% of the variation of Y is explained by X and X^2 in the data.

[**Note** that we will not go into problems of inference (standard deviations, tests, p-values etc.), apart from the estimation itself, at this stage. On the other hand, knowing the meaning of a p-value from the basic statistics course, you don't need to understand how the test is constructed. You can still interpret the p-values in the output, found under the heading, P>|t|. For example the first value, 0.000, gives the p-value for two-sided testing of the coefficient of the variable XM (i.e., test of $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$). Hence, we can conclude that the two beta-estimates are highly significant (i.e., significantly different from zero) while the constant term ($\hat{\alpha}$) found under "_cons", is not significantly different from zero.]

Using the *graph menu* as in figure 2, specifying "`quadratic prediction`" in the final menu (or simply replacing "lfit" with "qfit" in the command that was produced when you made the linear regression fit above), adds a plot of the regression function to the scatter plot as in figure 4:

Figure 4



(i) *Repeat this analysis for women and commodity group 1.*

[**Note:** Although the parabola gave a nice fit in this case, it does not mean that it is necessarily a good model. The parabola model implies for example that there is value X_0 of total expenditure (somewhere between 8000 and 9000), that corresponds to a maximum average expenditure $E(Y \mid X_0)$. For X larger than this optimal X_0, however, the mean expenditure decreases, contrary to intuition (maybe even to economic theory

(?)). It is, of course, possible to model the regression function in such a way that this unfortunate (?) consequence disappears - but this I leave to you to explore when you have obtained more training in econometrics.]

(ii) *Make linear fit plots for commodity groups 2 and 3 for both men and women. In which cases does the linear regression model seem to give a reasonable good fit , and in which cases does it seem to be unrealistic (no exact answers are asked for here - use your common sense)?*

(iii) *Any striking differences between women and men from looking at the plots (no exact answers are asked for here - use your common sense)?*

(iv) *Returning to commodity group 4, estimate the difference in estimated mean expenditure on services between women and men when (a) both have total expenditure HKD 1000, and (b) when both have total expenditure HKD 3000 .*