H. Goldstein
07.02.2013

Supplementary lecture note

1. Modeling relationships between the rv's *X* and *Y*

We saw in the lecture that when modeling a relationship between two random variables (rv's), *X* and *Y* (trying to explain *Y* by *X*), it is sometimes a good idea to start with the right side of

(1) $f(x, y) = f_c(y \mid x) f_X(x)$

(where $f(x, y)$ is the joint pdf of (X, Y), $f_c(y \mid x)$ is the conditional pdf of $(Y \mid X = x)$, X being fixed to a value *x*, and $f_X(x)$ the marginal pdf of *X*.)

There are, however, also cases where the modelling is done directly from the left side of the factorization. Starting from the left side of (1), a linear relationship between *X* and *Y* is sometimes proposed directly as

(2) $Y = \alpha + \beta X + e$

where the random error term, *e*, is uncorrelated with *X*, has expectation, $E(e) = 0$, and constant variance, $\mathrm{var}(e) = \sigma_e^2$.

As clarified by Harald Cramér in his famous textbook from 1945, **Mathematical Methods of Statistics**, the relation (2) *should not be taken as an (additional) assumption* since it actually follows (is true) from any model for $f(x, y)$, (as long as moments up to second order exist) – which will be shown as follows:

(3) **Model:** Let (X, Y) have a joint distribution given by $f(x, y)$, such that expectations, variances and covariance exist. Write them as the following five population quantities

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

I will show that the statement (2) is automatically valid in model (3) without extra assumptions:

(4) Define[1] α and β by: $\beta \overset{Def}{=} \dfrac{\sigma_{XY}}{\sigma_X^2}$ and $\alpha \overset{Def}{=} \mu_Y - \beta \mu_X$

[1] Note that the definition of α and β, which are population quantities, corresponds exactly to the OLS estimates based on a sample of observations of (X, Y).

(5) After this, define the rv, *e*, by: $e \overset{Def}{=} Y - \alpha - \beta X$

From this we have

(6) $Y = \alpha + \beta X + e$

Now the rv, *e*, can be interpreted as an error term in the sense of (2), since we shall prove that $E(e) = 0$ and $\text{cov}(e, X) = 0$:

Proof of $E(e) = 0$ **and** $\text{cov}(e, X) = 0$**:**

We find from STAT1 –rules

$$E(e) = E(Y - \alpha - \beta X) = \mu_Y - \alpha - \beta\mu_X \overset{(4)}{=} \mu_Y - (\mu_Y - \beta\mu_X) - \beta\mu_X = 0$$

From this we get

$$\begin{aligned} \text{cov}(e, X) &= E\left(eX\right) - E(e) \cdot E(X) = E\left(eX\right) = E\left((Y - \alpha - \beta X) \cdot X\right) = \\ &= E\left(XY - \alpha X - \beta X^2\right) = E\left(XY\right) - \alpha\mu_X - \beta E\left(X^2\right) \end{aligned}$$

Substituting for α and β and using that $E\left(XY\right) = \text{cov}(X,Y) + E(X) \cdot E(Y) = \sigma_{XY} + \mu_X \mu_Y$, and $E\left(X^2\right) = \sigma_X^2 + \mu_X^2$, we get

$$\text{cov}(e, X) = \sigma_{XY} + \mu_X \mu_Y - \left(\mu_Y - \frac{\sigma_{XY}}{\sigma_X^2}\mu_X\right) \cdot \mu_X - \frac{\sigma_{XY}}{\sigma_X^2}\left(\sigma_X^2 + \mu_X^2\right) = \cdots = 0$$ **End of proof.**

So the conditions in (2) are automatically fulfilled.

We also need an expression for $\sigma_e^2 = \text{Var}(e)$. Having established that *e* and *X* are uncorrelated, we have

$$\sigma_Y^2 = \text{var}(Y) = \text{var}\left(\alpha + \beta X + e\right) = \beta^2 \text{var}(X) + \text{var}(e) = \beta^2 \sigma_X^2 + \sigma_e^2$$

Hence, introducing the correlation between *X* and *Y*, $\rho = \dfrac{\sigma_{XY}}{\sigma_X \sigma_Y}$, we find

$$\sigma_e^2 = \sigma_Y^2 - \beta^2 \sigma_X^2 = \sigma_Y^2 - \frac{\sigma_{XY}^2}{\sigma_X^4}\sigma_X^2 = \sigma_Y^2 - \frac{\sigma_{XY}^2}{\sigma_X^2} = \sigma_Y^2 - \frac{\sigma_{XY}^2}{\sigma_X^2 \sigma_Y^2} \cdot \sigma_Y^2 = \sigma_Y^2 - \rho^2 \sigma_Y^2$$

or

(7) $\sigma_e^2 = \sigma_Y^2(1-\rho^2)$

(7) is a famous relationship that, among other things, offers an interpretion of ρ^2 (in the population) as a measure of the fraction of the total variance, σ_Y^2, of *Y* that is explained by the linear relationship (2).

Note that another bonus that we get from (7) is a proof that $-1 \le \rho \le 1$ (assuming $\mathrm{var}(Y) > 0$). This follows trivially since $\sigma_e^2 = \mathrm{var}(e) \ge 0$ implies $\rho^2 \le 1$.

2. Interpretation of (2)

Cramér derived (2) as the solution of a minimization problem. He showed that the values given of α and β, minimize the function, $Q(\alpha,\beta) = E\left[(Y-\alpha-\beta X)^2\right]$ (assuming (3) only), and that the minimizing values are unique[2]. This represents one interpretation. I will give you another supplementary interpretation.

Write the linear part of (2) as $g(X)$, where *g* is the linear function, $g(x) = \alpha + \beta x$. As shown, this function is always well defined under model (3). A common misunderstanding appears to be a tendency to interprete $g(x)$ as the conditional expectation, $\mu(x) = E(Y \mid x)$, which is wrong in general. To claim that $g(x) = \mu(x)$, requires additional assumptions in model (3). In general $\mu(x)$ does not have to be linear at all, as examples 1 and 2 below, show. However, there is a certain relationship between the two functions $\mu(x)$ and $g(x)$. To see this, make first $\mu(x)$ random by replacing *x* by the rv *X*, leading to the rv, $\mu(X)$. Then, add and subtract this in *Q*, giving

$$
\begin{aligned}
Q(\alpha,\beta) &= E\left[(Y-g(X))^2\right] = E\left[(Y-\mu(X)+\mu(X)-g(X))^2\right] = \\
&= E\left[(Y-\mu(X))^2\right] + 2E\left[(Y-\mu(X))(\mu(X)-g(X))\right] + E\left[(\mu(X)-g(X))^2\right]
\end{aligned}
$$

Using the theorem of iterated expectations, we can show that the first term is equal to, $E\left[\sigma^2(X)\right]$, where $\sigma^2(x) = \mathrm{var}(Y \mid x)$ is the conditional variance function.

> [You may try to show this as an exercise. Write first $E\left[(Y-\mu(X))^2\right] = E\left[E\left((Y-\mu(X))^2 \mid X\right)\right]$ and use the two-step approach described in the lecture. On step one, find the function of *x*,
> $E\left((Y-\mu(X))^2 \mid X = x\right) = E\left((Y-\mu(x))^2 \mid x\right) \overset{def}{=} \mathrm{var}(Y \mid x) = \sigma^2(x)$, noting that fixing *X* to the value *x*, turns, the rv $\mu(X)$ into the constant value, $\mu(x) = E(Y \mid x)$. On step two, replace *x* by the rv *X* and take expectations.]

Similarly, the second term in *Q*, can be shown to be 0. Hence, we can write

[2] Actually he showed (2) in a more general setting with an arbitrary number of explanatory variables, and not only one as here.

$$Q(\alpha,\beta)=E\left[\sigma^2(X)\right]+E\left[(\mu(X)-\alpha-\beta X)^2\right]$$

Since the minimization Q with respect to α and β does not affect the first term, we can now interprete the function $g(x)$ as the best linear approximation to $\mu(x)$ in an expected squared error sense.

3 Estimation of $g(x)=\alpha+\beta x$ in (2)

The joint distribution of (X,Y), determined by the joint pdf, $f(x,y)$, may be called *the population distribution* (as suggested by Ragnar in the lecture), and we want information on this distribution based on a representative sample (data) from the population expressed as " n observations[3], $(x_1^o,y_1^o),(x_2^o,y_2^o),\ldots,(x_n^o,y_n^o)$ of (X,Y), sampled independently". A model for this is the *iid* model where we consider the observations as observations of n random pairs, $(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)$ assumed to be *iid* pairs. This means that we assume the n pairs to be independent random pairs, identically distributed as (X,Y), and with the common pdf, $f(x,y)$.

Then, using the construction of (2) for each pair[4], we get the (simple regression) model for data (known from several books),

$$Y_i=\alpha+\beta X_i+e_i,\quad i=1,2,\ldots,n$$

where the error terms, $e_1,e_2,\ldots,e_n$, are independent and identically distributed with expectation 0 and constant variance, $\operatorname{var}(e_i)=\sigma_e^2$. This specification can be well treated (estimated) by the OLS method.

4. Example 1

We can construct the joint pdf, $f(x,y)$, for the random pair, (X,Y) using the factorization, $f(x,y)=f_c(y\,|\,x)f_X(x)$, (see (1)), where the two factors on the right can be modeled as we wish – independently of each other. For example, suppose that X is uniformly distributed over the interval [0,1] with pdf

$$f_X(x)=\begin{cases}1 & \text{for } 0\le x\le 1\\ 0 & \text{for } x \text{ outside } [0,1]\end{cases}$$

This implies that, $E(X)=\mu_X=1/2$, and $\operatorname{var}(X)=\sigma_X^2=1/12$.

[3] The upper index ^{o} signifying that the observations are concrete numbers.
[4] Note that, since all the pairs have the same distribution, α and β will be the same for each pair.

Fixing X to a number x between 0 and 1, we assume that $Y \mid (X = x)$ is normally distributed with expectation, $2x^2$, and variance 1 (in short $(Y \mid X = x) \sim N(2x^2, 1)$). This implies that the (true) regression function is $\mu(x) = E(Y \mid x) = 2x^2$ (well defined for x in [0,1] only since the conditional pdf, $f_c(y \mid x)$, is not defined for x outside [0,1]).

> [**Small technical point:** Note that, even if $f_c(y \mid x)$ is not defined for x outside [0,1], we define (as is usual in practice) $f(x, y) = f_c(y \mid x) f_X(x) = 0$ for x outside [0,1], since $f_X(x) = 0$ there. Hence, the joint pdf becomes
>
> $$f(x, y) = \begin{cases} \dfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y - 2x^2)^2} & \text{for } 0 \le x \le 1 \text{ and } -\infty < y < \infty \\ 0 & \text{for any } (x, y) \text{ where } x \text{ is outside } [0,1] \end{cases}$$
>
> Note that $f(x, y)$ is concentrated (i.e., > 0) in the strip which goes all along the y-axis determined by $0 \le x \le 1$ and 0 outside the strip. Note also that (X, Y) is *not* jointly normally distributed (the joint normal pdf is never 0 and would imply that both marginal distributions are normal).]

Hence the true regression function, $\mu(x)$, is not linear (but part of a parabola, see fig. 1 below). To determine the best linear approximation described in (2), we need as in (3), $E(X) = \mu_X = 1/2$, $E(Y) = \mu_Y$, $\operatorname{var}(X) = \sigma_X^2 = 1/12$, $\operatorname{var}(Y) = \sigma_Y^2$, and $\operatorname{cov}(X, Y) = \sigma_{XY}$. The obvious tool here is the iterated expectation theorem:

$$\mu_Y = E(Y) = E\left[E(Y \mid X)\right] = E\left[2X^2\right] = 2\left[\operatorname{var}(X) + \left(E(X)\right)^2\right] = 2\left(\frac{1}{12} + \frac{1}{4}\right) = \frac{2}{3}$$

$$\sigma_{XY} = \operatorname{cov}(X, Y) = E(XY) - \mu_X \mu_Y = E(XY) - \frac{1}{3}$$

$$E(XY) = E\left[E(XY \mid X)\right] = E\left[X \cdot E(Y \mid X)\right] = E\left[X \cdot 2X^2\right] = 2E\left[X^3\right]$$

> [**Note.** To understand this manipulation, it is best to use the two-stage approach described in the lecture. **Step 1:** Find the conditional function behind the inner expectation first. $c(x) = E\left(XY \mid X = x\right) = E\left(xY \mid X = x\right) = x \cdot E(Y \mid x)$. This works since the value x is just a constant in the distribution of $Y \mid (X = x)$. Hence, $c(x) = xE(Y \mid x) = 2x^3$. **Step 2:** Replace x by the rv X in $c(x)$ and take expectation: $c(X) = X \cdot E(Y \mid X) = 2X^3$. The theorem of iterated expectations tells us that taking the expectation of $c(X)$ gives us $E\left(XY\right)$.]

$$E(XY) = 2E\left(X^3\right) = 2\int_0^1 x^3 f_X(x)dx = 2\int_0^1 x^3 dx = 2 \cdot \frac{1}{4} = \frac{1}{2}$$

Hence

$$\sigma_{XY} = E(XY) - \frac{1}{3} = \frac{1}{2} - \frac{1}{3} = \frac{1}{6}$$

$$\sigma_Y^2 = \text{var}(Y) = E\left[\text{var}(Y\,|\,X)\right] + \text{var}\left[E(Y\,|\,X)\right] = E\left[1\right] + \text{var}\left[2X^2\right] = 1 + 4\cdot\text{var}\left(X^2\right)$$

Hence

$$\sigma_Y^2 = 1 + 4\left(E\left[X^4\right] - \left(E\left[X^2\right]\right)\right) = 1 + 4\left(\int_0^1 x^4 \cdot 1dx - \left(\int_0^1 x^2 \cdot 1dx\right)^2\right) = 1 + 4\cdot\frac{4}{45} = 1 + \frac{16}{45}$$

We can now find the best linear approximation, $g(x) = \alpha + \beta x$, to $\mu(x) = E(Y\,|\,x)$, where

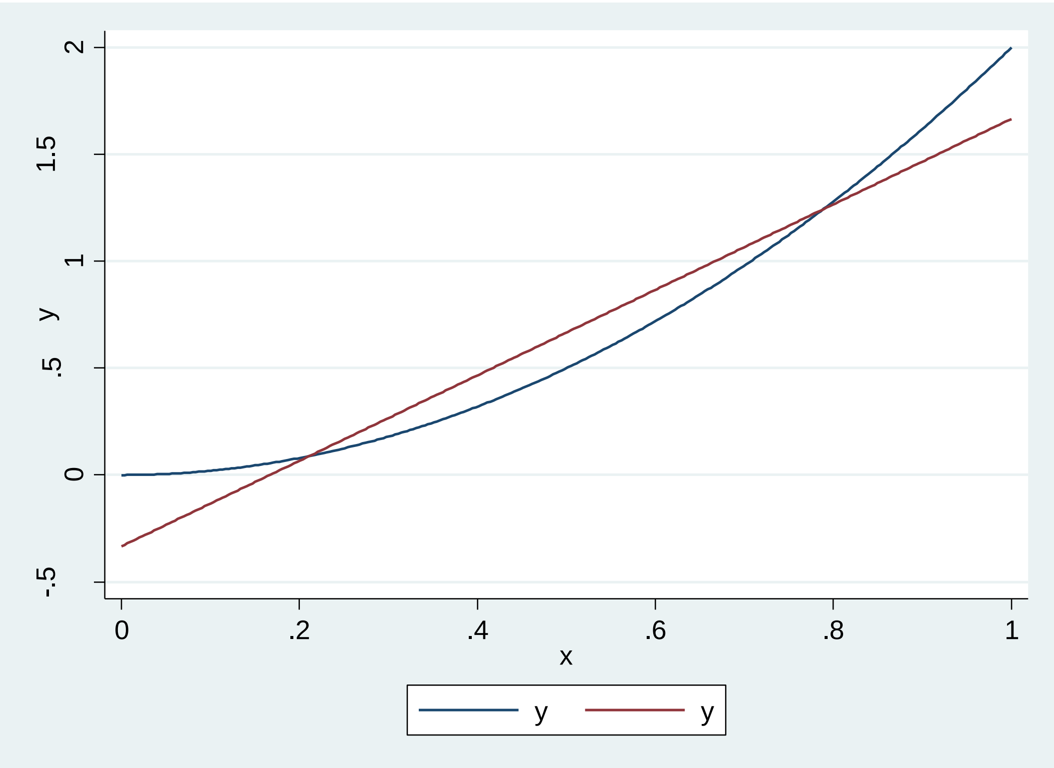
$$\beta = \frac{\sigma_{XY}}{\sigma_X^2} = \frac{1/6}{1/12} = 2$$

$$\alpha = \mu_Y - \beta\mu_X = \frac{2}{3} - 2\cdot\frac{1}{2} = -\frac{1}{3}$$

or

$$g(x) = -\frac{1}{3} + 2x$$

Figure 1 $g(x) = \alpha + \beta x$ and regression function $\mu(x) = E(Y\,|\,x)$



Stata command: twoway (function y=2*x^2, range(0 1)) (function y=-1/3+2*x, range(0 1))

Example 2 Exercise

A. Repeat the calculations as in example 1, now assuming

(i) $X \sim$ uniformly distributed over [0, 2]

[implying, $f_X(x) = 1/2$ for $0 \le x \le 2$, $\mu_X = E(X) = 1$, and $\sigma_X^2 = \text{var}(X) = 1/3$]

(ii) $Y \mid x \sim N\left(2(x-1)^2, x^2\right)$

[implying, $\mu(x) = E(Y \mid x) = 2(x-1)^2$ and $\sigma^2(x) = \text{var}(Y \mid x) = x^2$]

Hint. Verify that $\mu_Y = \dfrac{2}{3}, \quad \sigma_Y^2 = \dfrac{76}{45}, \quad \sigma_{XY} = -\dfrac{1}{3}$.

Developing σ_Y^2, you may need to find $\text{var}\left[(X-1)^2\right] = E\left[(X-1)^4\right] - \left(E\left[(X-1)^2\right]\right)^2$.

Note that, for example, $E\left[(X-1)^4\right] = \int_0^2 (x-1)^4 f_X(x)dx = \int_0^2 (x-1)^4 \cdot \frac{1}{2} dx = \dfrac{1}{5}$

B. Find the best linear approximation, $g(x) = \alpha + \beta x$, to the true regression, $\mu(x) = E(Y \mid x)$, and plot both functions in the same graph.

C. How would you go about to simulate (make the computer draw) *n* independent observations of the random pair, (X, Y)? (**Hint:** Utilize the right side of (1).)

(Try it (!), using (e.g.) Excel or Stata for *n* = 50, and make a scatter plot.)