Econ 4130
HG Oct. 2019

Lecture note – Introduction to prediction and something about exogeneity in the *iid* model

First about prediction. (See also Rice section 4.4.2. Rice is somewhat thin on prediction - which justifies this note since prediction is important in econometrics.)

The *iid* model will be discussed in example 4 and the appendix 1 below that.

When we want to estimate the future outcome (observation) of a random variable, *X*, we use the term *prediction* instead of "estimation". In contrast, the term *estimation* is reserved for the task of estimating a fixed unknown quantity (parameter) in a population. In the special case when *X* represents a future value in a time series, we usually instead use the term, *forecasting* (Norwegian: *prognose*), for prediction.

1 Case 1 – Prediction of *X* when the distribution of *X* is known

Suppose that $X \sim f(x)$ (*pmf* or *pdf*) where *f* is known. We want to predict (guess) the outcome, X_{obs}, of a future observation of *X* – an outcome that is not yet known. Let *c* denote our prediction (guess).

What is the best way to choose *c*?

The answer to this question does not have a universally valid solution, but will depend on the criterion of "best" that we use. Such a criterion is usually formulated in terms of the average value of a suitable *loss function*. The following three criteria (especially *crit.*1) are among the most common:

Criterion 1 Loss function = $L_1(c) = (X-c)^2 = (\text{prediction error})^2$.
Choose *c* to minimize expected loss = $E\left[(X-c)^2\right]$ (also called "the mean squared error" or MSE):
Answer: Best prediction is the expectation, $c = \mu = E(X)$.

Criterion 2 Loss function = $L_2(c) = |X-c| = |\text{prediction error}|$.
Choose *c* to minimize expected loss = $E|X-c|$.
Answer: Best prediction is $c = m = \text{median}(X)$.

Criterion 3 (Mostly used for discrete distributions).
Loss function, $L_3(c) = \begin{cases} 0 & \text{if } c = X_{obs} \\ 1 & \text{if } c \neq X_{obs} \end{cases}$
Choose *c* to minimize expected loss =
$E(L_3(c)) = 0 \cdot P(X=c) + 1 \cdot P(X \neq c) = 1 - P(X=c)$

Answer: Best prediction is c = the most likely observation = the c that makes $P(X=c)$ largest
(also called *the mode* in the distribution of X).

Example of criterion 3: Suppose the pmf of X is given by

x	1	2	3
$f(x)$	0.5	0.3	0.2

The best guess (prediction) of X, according to *crit.*3, is $c=1$. This way to predict is probably the common way applied intuitively by most people (being risk averters…).

Proof of criterion 1. We have (using that $\mu-c$ is a constant):

$$MSE = E\left(L_1(c)\right) = E\left[(X-c)^2\right] = E\left[(X-\mu+\mu-c)^2\right] =$$

(1) $$= E\left[(X-\mu)^2\right] + 2(\mu-c)E(X-\mu) + (\mu-c)^2 = E\left[(X-\mu)^2\right] + (\mu-c)^2 =$$

$$= \operatorname{var}(X) + (\mu-c)^2$$

which shows that minimum MSE is obtained by choosing $c=\mu$. (End of proof).

Proof of criterion 2 (optional reading): The proof in the discrete case is a little tricky - so we skip that and assume that X is a continuous rv with pdf, $f(x)$ and cdf $F(x)$. Then

$$E\left[L_2(c)\right] = E\left|X-c\right| = \int_{-\infty}^{\infty} |x-c|\, f(x)dx = \int_{-\infty}^{c} |x-c|\, f(x)dx + \int_{c}^{\infty} |x-c|\, f(x)dx =$$

$$= \int_{-\infty}^{c} (c-x) f(x)dx + \int_{c}^{\infty} (x-c) f(x)dx = c\int_{-\infty}^{c} f(x)dx - c\int_{c}^{\infty} f(x)dx - \int_{-\infty}^{c} xf(x)dx + \int_{c}^{\infty} xf(x)dx =$$

$$= cP(X \le c) - cP(X > c) - \int_{-\infty}^{c} xf(x)dx + \int_{c}^{\infty} xf(x)dx =$$

$$= c\left[2F(c)-1\right] - \int_{-\infty}^{c} xf(x)dx + \int_{c}^{\infty} xf(x)dx$$

Using that $\frac{d}{dc}\int_{-\infty}^{c} xf(x)dx = cf(c)$ and $\frac{d}{dc}\int_{c}^{\infty} xf(x)dx = -cf(c)$, we get

$$\frac{d}{dc}E\left[L_2(c)\right] = \cdots \text{fill in} \cdots = 2F(c)-1$$

showing that minimum is obtained for $2F(c)-1=0$, or $c = \text{median}(X)$

(End of proof)

Example 2 Consider the simple case of *X* being the result of throwing a fair die, with pmf, $f(x) = P(X = x) = 1/6$ for $x = 1,2,\ldots,6$. An easy calculation gives us $\mu = E(X) = 3.5$ which, according to *crit.*1 should be the best prediction of X_{obs}. This is, of course, nonsense since 3.5 is not possible as an outcome. This, however, does not mean that *criterion* 1 itself is nonsense. Note that *crit.* 1 just says to minimize $E\left[(X-c)^2\right] = \sum_{x=1}^{6}(x-c)^2\cdot\frac{1}{6}$ when *c* is one of the permitted values, $1,2,\ldots,6$.

Similarly, *crit.*2 says to choose *c* among $1,2,\ldots,6$, so that $E\left|X-c\right| = \sum_{x=1}^{6}|x-c|\cdot\frac{1}{6}$ becomes as small as possible. Using, e.g., Excel we get the table

c	1	2	3	4	5	6
$E\left[(X-c)^2\right]$	9.2	5.2	3.2	3.2	5.2	9.2
$E\left\|X-c\right\|$	2.5	1.8	1.5	1.5	1.8	2.5
$E\left[L_3(c)\right]$	5/6	5/6	5/6	5/6	5/6	5/6

Hence, both *crit.*1 and *crit.*2 lead to two equally good solutions for the best prediction, i.e., $c = 3$ or $c = 4$. *Crit.*3 says that any of $1,2,\ldots,6$ will do as best prediction.

In the continuous case, if, e.g., $X \sim N(169, 36)$, then both *crit.*1 and 2 would lead to 169 as the best prediction of a future X_{obs}. Remember that the expectation and the median are equal in symmetric distributions.

If $X \sim \exp(\lambda = 0.5)$, we have that $E(X) = \frac{1}{\lambda} = 2$ and $\text{median}(X) = \frac{\ln 2}{\lambda} = 1.39$, so that *crit.*1 would lead to 2 and *crit.*2 to 1.39 as the best prediction respectively of a future X_{obs}.

Example 3

Suppose (X,Y) is a pair of rv's with joint pdf, $f(x,y)$, which is known. We want to predict *Y* for a given chosen value of *X*, i.e., for $X = x$. The relevant distribution for *Y* to use for the prediction is then clearly the conditional distribution of *Y* given $X = x$, with pdf, $f(y\,|\,x) = f(x,y)/\,f_1(x)$. The best prediction is then, according to criterion 1, given by the expectation in this distribution, i.e., $E(Y\,|\,x)$. For example, using the distribution in Rice defined on page 75,

$$(X,Y) \sim f(x,y) = \begin{cases} \frac{12}{7}(x^2+xy) & \text{for } 0 \le x \le 1 \quad 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Suppose we want to predict *Y* when *X* has the value, $X = 0.75$. In the lecture we derived the conditional pdf (valid for $0 \le x \le 1$):

$$f(y \mid x) = \frac{x+y}{x+0.5} \quad \text{for } 0 \le y \le 1$$

and the expectation, $\mu(x) = E(Y \mid x) = \frac{x+2/3}{2x+1}$. Hence, the best (following *crit.*1) prediction of *Y* is $\mu(0.75) = E(Y \mid 0.75) = \frac{0.75+2/3}{1.5+1} = 0.567$.

2 Case 2 – Prediction of *X* when the distribution of *X* is unknown

We want to predict *X* when the distribution with pdf $f(x)$ and $\mu = E(X)$ are unknown. We will use the most common criterion 1 for choosing our predictor. The task is, thus, to find the best prediction, $\hat{X}$, of *X* based on,

Criterion 1 Choose $\hat{X}$ to minimize $MSE = E\left[(X-\hat{X})^2\right]$.

If μ were known, we would use $\hat{X} = \mu$, of course. μ being unknown, the natural thing to do is to estimate μ as good as we can from available data and use the estimate, $\hat{\mu}_{obs}$, as our prediction. It turns out that this intuitive procedure is the correct one based on *crit.*1, as the following theoretical argument shows: Suppose that the available data are observations of $X_1, X_2, \ldots, X_n$, which we denote as the vector $D = (X_1, X_2, \ldots, X_n)$.

> **Elaboration (optional reading until the relation (*) next page).** For simplicity we assume that the future *X* to be predicted is independent of *D*, implying that the conditional pdf, $f(x \mid x_1, x_2, \ldots, x_n) = f(x \mid d) = f(x)$ does not depend on *d*.[1] This implies further that $E(X \mid d) = E(X) = \mu$.
>
> Now, our task becomes to construct our predictor $\hat{X} = h(D)$ as a function of *D* in the best way, i.e., such that $MSE = E\left[(X-\hat{X})^2\right] = E\left[(X-h(D))^2\right]$ is minimized.
>
> Applying the theorem of total expectation on the *MSE*, we get

[1] This assumption is typically reasonable for cross section data where *D* often consists of iid variables. For time series data, however, where X_i represents the value of the series at time point *i*, and *X* the value of the series at a future time point, there will often be dependence between *X* and *D*. The conclusion of the argument, however, will still be the same, i.e., the best prediction of *X* will simply be the best estimate of $E(X \mid d)$.

$$MSE = E\left[(X-h(D))^2\right] = E\left\{E\left[(X-h(D))^2 \mid D\right]\right\} = Eg(D)$$

where $g(d)$ is the function,

$$g(d) = E\left[(X-h(D))^2 \mid D=d\right] = E\left[(X-h(d))^2 \mid D=d\right]$$

noting that $h(d)$ is just a constant in the conditional distribution given $D=d$. Now we can do the same thing as in the proof of *crit.*1 on page 2, adding and subtracting the expected value, $E(X \mid d)$, inside $(X-h(d))^2$, and we get

$$g(d) = E\left[(X-E(X \mid d))^2 \mid D=d\right] + \left(E(X \mid d) - h(d)\right)^2$$

As we saw above, $E(X \mid d) = E(X) = \mu$, giving

$$g(d) = E\left[(X-\mu)^2 \mid D=d\right] + \left(\mu - h(d)\right)^2$$

But, also $(X-\mu)^2$ is independent of *D*, so we can drop the conditioning in the first term in $g(d)$, giving

$$g(d) = E\left[(X-\mu)^2\right] + (\mu - h(d))^2 = \sigma^2 + (\mu - h(d))^2 \text{ , where } \sigma^2 = \mathrm{var}(X) .$$

Now it only remains to replace *d* by the rv *D* in $g(d)$ and take expectation (*the optional reading ends here*).

We then get

$$(*) \qquad MSE = Eg(D) = E\left[\sigma^2 + (\mu - h(D))^2\right] = \sigma^2 + E\left[(\mu - h(D))^2\right].$$

where $D = (X_1, X_2, \ldots, X_n)$ and $\hat{X} = h(D) = h(X_1, X_2, \ldots, X_n)$ an arbitrary predictor of *X*. From this result we can conclude

Conclusion

> The best predictor, $\hat{X} = h(D)$, according to *crit.*1 is a function of the data (*D*) that minimizes $E\left[(\mu - h(D))^2\right]$, or, in other words, $h(D)$ is the same as *the best estimator*, $\hat{\mu} = h(D)$, of μ, minimizing $E\left[(\mu - \hat{\mu})^2\right]$.

This ends the theoretical elaboration confirming the intuition we started with, namely that the problem of predicting *X* reduces to the problem of estimating $\mu = E(X)$.

This also tells us how to predict (using *crit.*1) a response variable *Y* based on a given value, $X = x$, of an explanatory variable, X:

- If $E(Y \mid x)$ is known, the best prediction of *Y* is given by $E(Y \mid x)$.

- If $E(Y \mid x)$ is unknown, the best prediction of *Y* is given by the best possible estimate, $\hat{E}(Y \mid x)$, of $E(Y \mid x)$.

For example, if $E(Y \mid x) = \alpha + \beta x$, the best prediction of *Y*, given $X = x$, is $\hat{Y} = \hat{\alpha} + \hat{\beta} x$, where $\hat{\alpha}, \hat{\beta}$ are good estimators, for example OLS in the case of homoscedasticity. This is the reason that Stata (and other packages) usually call the expression $\hat{\alpha} + \hat{\beta} x$ for "predicted *Y*".

Note that, even if the prediction problem gives the same answer (estimate) as an estimation problem, it is wrong to look at the two problems as the same. The difference between prediction and estimation does not appear in the estimates, but rather in the uncertainty. The uncertainty of a prediction (measured by a so called "prediction interval") is larger than the uncertainty of the corresponding estimation (as measured by a confidence interval). This will be illustrated in the following example 4 with data.

Example 4

We will use the mother/daughter data collected from Stat 1 lectures (Econ2130) in the period 2010-2012. (The data, consisting of $n = 125$ observations, can be downloaded as an Excel-file from http://folk.uio.no/haraldg/ under the heading Econ2130.) In addition to prediction we will also use the example as an opportunity to discuss the important *iid* model often used for cross section data.

Let *X* denote the height of the mother and *Y* the height of the daughter for a randomly chosen pair from the population. This we express by saying, $(X,Y) \sim f(x,y)$ where the pdf $f(x,y)$ represents the *population distribution*.

The problem is to predict the height of the daughter (*Y*) for a particular pair where the mother has height, $X = x_0 = 160$. (It could, for example, be a case where the mother has only one child, a daughter who is just a small child, and we want to predict how tall the daughter will be when she grows up.)

We will use criterion 1 as our guide line for "best prediction".

The first model assumption is that we have a *representative* sample from the population. More precisely:

(1) The data are observations of an *iid* sample of random pairs, $(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)$, where $(X_i,Y_i) \sim f(x_i,y_i)$ for $i = 1,2,\ldots,n$.

Note that this implicitely implies *representativity* in the sense that the sample would have been non-representative if, e.g., the common distribution of (X_i,Y_i) $i = 1,2,\ldots,n$, were different from the population distribution *f*.

There are strong reasons to postulate that the regression function for *Y* with respect to *X* (i.e., $\mu(x) = E(Y \mid x)$) is a linear function of x in this situation. (For example, assuming that (X,Y) is bivariate normally distributed, which is an empirically well founded assumption for homogeneous height data, the linearity of the regression function is automatically fulfilled – even the assumption of homoscedasticity.) So we assume

(2) Regression function: $\mu(x) = E(Y \mid x) = \alpha + \beta x$, where α, β are unknown constants.
(3) Homoscedasticity: $\sigma^2(x) = \text{var}(Y \mid x) = \sigma^2$ - constant.

If α, β were known, we would use $\mu(x_0) = \mu(160) = \alpha + \beta x_0 = \alpha + \beta \cdot 160$ as our best prediction. Since they are unknown, however, we will need to estimate them in order to obtain a prediction.

- For the sample, from assumptions (1), (2), (3), it follows that for $i = 1, 2, \ldots, n$, $E(Y_i \mid x_i) = \mu(x_i) = \alpha + \beta x_i$ and $\text{var}(Y_i \mid x_i) = \sigma^2$.
- Notice here that x_i represents the observed value of the rv X_i and, as such, a fixed number.
- Introducing error terms, $e_i = Y_i - \alpha - \beta x_i$, we can write this equivalently as $Y_i = \alpha + \beta x_i + e_i$ for $i = 1, 2, \ldots, n$, where $E(e_i \mid x_i) = 0$ and $\text{var}(e_i \mid x_i) = \text{var}(Y_i - \alpha - \beta x_i \mid x_i) = \text{var}(Y_i \mid x_i) = \sigma^2$ (using that $\alpha + \beta x_i$ is a constant in the conditional distribution given $X_i = x_i$).
- Hence it appears that we can use the simple regression model presented in Stat 1 (Econ2130) where the values, $x_1, x_2, \ldots, x_n$, of the explanatory variable are assumed to be fixed numbers.

That this is allowed will be justified theoretically in the appendix 1 below. The fixation of the explanatory variable to the observed values is a highly convenient measure, simplifying the estimation problem considerably, and is used a lot in econometrics when dealing with exogenous explanatory variables. *But, of course, it needs a theoretical justification…*

Thus our model for the data is the same as in the Stat 1 course:

(4) $Y_i = \alpha + \beta x_i + e_i$ for $i = 1, 2, \ldots, n$

where the error terms $e_1, e_2, \ldots, e_n$ are *iid* rv's with $e_i \sim N(0, \sigma^2)$, and where $x_1, x_2, \ldots, x_n$ are fixed numbers.

Note that the normality of e_i follows automatically if (X_i, Y_i) is bivariate normally distributed (a reasonable assumption here) which implies that the conditional distributions are normal as well. So $(e_i \mid x_i) \sim N\left[E(e_i \mid x_i),\ \text{var}(e_i \mid x_i)\right] = N(0, \sigma^2)$.

Note also that e_i is a *non observable* variable (also called a *latent* variable) since it depends on the unknown α, β. However, it can be estimated (i.e., predicted) by the predictor (called *residual*), $\hat{e}_i = Y_i - \hat{\alpha} - \hat{\beta} x_i$ for $i = 1, 2, \ldots, n$. In general, residuals contain important information when judging the realism of some model and is therefore often used as a diagnostic tool for this purpose (i.e., a tool for studying potential weaknesses of the model).

Introducing the population parameters,
$\mu_X = E(X), \quad \mu_Y = E(Y), \quad \sigma_X^2 = \text{var}(X), \quad \sigma_Y^2 = \text{var}(Y), \quad \sigma_{XY} = \text{cov}(X,Y)$,
these can be estimated by the sample analogues
$\bar{X}, \ \bar{Y}, \ S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \ S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2, \ S_{xy}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$
respectively. (It can be proven that if (X,Y) is bivariate normal these will be equal to the maximum likelihood estimators except that $1/(n-1)$ in the variance/covariance formulae's is replaced by $1/n$.)

Introducing the correlation between *X* and *Y*, $\rho = \sigma_{XY}/(\sigma_X \sigma_Y)$, this can be estimated by the sample analogue, $r = S_{xy}/(S_x S_y)$ (which will, actually, be equal to the mle estimator in bivariate normal case).
Using model (4) we can replace all X_i by the observed values, x_i, and consider $\bar{x}$ and $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ as constants and $S_{xy}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})$. The OLS estimators are then given by $\hat{\beta} = \frac{S_{xy}}{s_x^2}, \quad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}, \quad \hat{\mu}(x) = \hat{\alpha} + \hat{\beta}x$, and the regression variance,

$\sigma^2 = \text{var}(Y \mid x) = \sigma_Y^2(1-\rho^2)$, is estimated by $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{e}_i^2 = \frac{n-1}{n-2}\left(S_y^2 - \hat{\beta}^2 s_x^2\right)$
(see, e.g., the Stat 1 course for formulae's). Estimates are given in the table

Table of estimates. (Sample size, $n = 125$)

Parameter	Estimate	Value
μ_X	$\bar{x}$	166.9
μ_Y	$\bar{y}$	167.6
σ_X	$s_x = \sqrt{s_x^2}$	5.8232
σ_Y	$s_y = \sqrt{s_y^2}$	5.5938
σ_{XY}	s_{xy}	11.7315
β	$\hat{\beta} = s_{xy}/s_x^2$	0.346
α	$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$	109.96
$\mu(160)$	$\hat{\alpha} + \hat{\beta} \cdot 160$	165.3

According to this, our prediction of the daughter's height when she grows up is 165.3.

Note that the estimate, $\hat{\mu}(160)_{obs} = 165.3$ now has two interpretations:

- It is the (OLS) prediction of a single case of daughter's height when the mother is 160 cm.
- It is the (OLS) estimate of the mean in the population of daughter's heights with mothers all being 160 cm.

Evaluation of uncertainty (in terms confidence/prediction intervals).

95% confidence interval (CI) for $\mu(x_0)=\mu(160)$ (review from the basic course, Econ2130):

According to results given in the Stat 1 course, $T_1=\dfrac{\hat{\mu}(x_0)-\mu(x_0)}{\mathrm{SE}(\hat{\mu}(\mathrm{x}_0))}\overset{\text{exactly}}{\sim} t_{n-2}$ distributed (i.e., *t*-distribution with *n*-2 degrees of freedom). When *n* is so large as 125, the t_{123}-distribution is practically almost identical with the *N*(0,1) distribution, so we may as well state $T_1\sim N(0,1)$.

$\hat{\mu}(x_0)=\hat{\alpha}+\hat{\beta}x_0$ is unbiased and with variance

(5) $$\operatorname{var}(\hat{\mu}(x_0))=\sigma^2\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{(n-1)s_x^2}\right)$$

The formulae for the (estimated) standard error is

(6) $$\mathrm{SE}(\hat{\mu}(x_0))=\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{(n-1)s_x^2}}$$

where, as derived in the basic course, $\hat{\sigma}^2=\dfrac{1}{n-2}\sum_{i=1}^{n}\hat{e}_i^2=\dfrac{n-1}{n-2}\left(S_y^2-\hat{\beta}^2 s_x^2\right)$. Using $T_1\sim N(0,1)$, we get

$$0.95=P(-1.96\le T_1\le 1.96)=P\left(-1.96\le\frac{\hat{\mu}(x_0)-\mu(x_0)}{\mathrm{SE}(\hat{\mu}(\mathrm{x}_0))}\le 1.96\right)=$$
$$=P\left[\hat{\mu}(x_0)-1.96\cdot\mathrm{SE}(\hat{\mu}(\mathrm{x}_0))\le\mu(x_0)\le\hat{\mu}(x_0)+1.96\cdot\mathrm{SE}(\hat{\mu}(\mathrm{x}_0))\right]$$

Using $x_0=160$, the estimates become $\hat{\sigma}^2_{obs}=(5.2396)^2$ and

$\mathrm{SE}(\hat{\mu}(160))|_{obs}=\hat{\sigma}_{obs}\sqrt{\dfrac{1}{n}+\dfrac{(160-\bar{x})^2}{(n-1)s_x^2}}=\hat{\sigma}_{obs}\cdot(0.139)=0.728$, giving

A 95% CI for $\mu(160)$**:**

(7) $$\left[\hat{\mu}(160)\pm 1.96\cdot\mathrm{SE}(\hat{\mu}(160))\right]_{obs}=165.3\pm 1.96\cdot 0.728=\left[163.9,\ 166.7\right]$$

A 95% prediction interval (PI) for $Y\,|\,X=160$**:**

Instead of the standard error of $\hat{\mu}(x_0)$, we now use the square root of mean squared error (MSE) of the predictor $\hat{Y}=\hat{\mu}(x_0)$:

$$MSE = E\left[(Y-\hat{\mu}(x_0))^2 \mid X = x_0\right] = E\left[(Y-\mu(x_0)+\mu(x_0)-\hat{\mu}(x_0))^2 \mid X = x_0\right] = \cdots =$$
$$= E\left[(Y-\mu(x_0))^2 \mid x_0\right] + E\left[(\mu(x_0)-\hat{\mu}(x_0))^2\right] = \mathrm{var}(Y \mid x_0) + \mathrm{var}(\hat{\mu}(x_0)) =$$
$$= \sigma^2 + \sigma^2\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{(n-1)s_x^2}\right) = \sigma^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{(n-1)s_x^2}\right)$$

Notice that an addition to the variance of $\hat{\mu}(x_0)$ appears that does not get smaller when n increases.

Let us call the estimated version of the square root of MSE for $\mathrm{SE}(\hat{Y})$. Then

$$\mathrm{SE}(\hat{Y}) = \hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{(n-1)s_x^2}} \text{ with observed value, } \mathrm{SE}(\hat{Y})_{obs} = 5.287$$

Now, in the normal case (Y being normally distributed), it can be proven[2] that

$T_2 = \dfrac{Y-\hat{\mu}(x_0)}{\mathrm{SE}(\hat{Y})} \overset{\text{exactly}}{\sim} t_{n-2}$ - which we here may identify with the $N(0,1)$ distribution. Hence, the same calculation as above gives us

$$0.95 = P(-1.96 \le T_2 \le 1.96) = P\left(-1.96 \le \frac{Y-\hat{\mu}(x_0)}{\mathrm{SE}(\hat{Y})} \le 1.96\right) =$$
$$= P\left[\hat{\mu}(x_0) - 1.96\cdot\mathrm{SE}(\hat{Y}) \le Y \le \hat{\mu}(x_0) + 1.96\cdot\mathrm{SE}(\hat{Y})\right]$$

So the 95% “prediction interval” (PI) for Y becomes, $\hat{\mu}(x_0) \pm 1.96\cdot\mathrm{SE}(\hat{Y})$. The observed value of the PI for $x_0 = 160$ is

$$\left[\hat{\mu}(160) \pm 1.96\cdot\mathrm{SE}(\hat{Y})\right]_{obs} = \left[165.3 \pm 1.96\cdot 5.287\right] = \left[154.9,\ 175.7\right]$$

which is larger than the corresponding CI for $\mu(160)$.

Appendix 1 Theoretical justification that the observed values of the explanatory variable can be considered as fixed numbers in the regression model

The starting point is the iid model in (1) repeated here

(8) The data are observations of an *iid* sample of random pairs, $(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)$, where $(X_i,Y_i) \sim f(x_i,y_i)$ for $i = 1,2,\ldots,n$.

[2] Technical details (optional reading) about the construction of *t*-distributed rv’s (as well as *F*-distributed rv’s) can be found in Rice chapter 6.

where *f* represents the population distribution for (X,Y). This implies that the joint pdf of all the variables (denoted by $\bar{f}$) can be written[3]

(9) $\bar{f}(x_1,y_1,x_2,y_2,\ldots,x_n,y_n) = f(x_1,y_1)f(x_2,y_2)\cdots f(x_n,y_n)$

We can also factorize $\bar{f}$ using the conditional distribution of the Y_i's given all the $X_i = x_i$

(10) $\bar{f}(x_1,y_1,x_2,y_2,\ldots,x_n,y_n) = f(y_1,y_2,\ldots,y_n \mid x_1,x_2,\ldots,x_n)\cdot f_1(x_1,x_2,\ldots,x_n)$

where f_1 is the marginal joint pdf of all the X_i's obtained by integrating out all the y_i's from $\bar{f}$. If all the arguments, x_i, y_i's are equal to the observed values of X_i, Y_i, (10) gives the likelihood function of our data with log likelihood

(11) $l(\theta) = \ln \bar{f} = \ln f(y_1,y_2,\ldots,y_n \mid x_1,x_2,\ldots,x_n) + \ln f_1(x_1,x_2,\ldots,x_n)$

where $\theta = (\alpha,\beta,\sigma^2,\theta_4,\ldots,\theta_r)$ is the vector of all the parameters in the model. We notice that, from the model formulation (4), that all the three parameters of interest, α,β,σ^2, occur in the first term only in (11) and not in the second term, so that, when we develop the mle estimators for these parameters by derivation, the derivative of the second term ($\ln f_1$) will always be zero.

In other words, the maximization of the full likelihood function with respect to the three parameters of interest, α,β,σ^2, is equivalent to maximizing the conditional likelihood, $f(y_1,y_2,\ldots,y_n \mid x_1,x_2,\ldots,x_n)$ with respect to α,β,σ^2. **Hence we can consider the observations, $x_1,x_2,\ldots,x_n$, as given fixed numbers when estimating α,β,σ^2.**

This argument, referring to the maximum likelihood principle for estimating the parameters of interest, is a common argument used in the econometric literature for justifying the consideration of the observed values, x_i's, as given fixed numbers in the model.

[3] The factorization actually uses a slightly more general concept of independence than given in the lecture. Namely that one group of rv's is independent of another group of rv's if the joint pdf of all variables is the product of the two marginal joint pdf's for the two groups. For example, (X_1,Y_1) and (X_2,Y_2) are independent pairs if the joint pdf can be factorized, $\bar{f}(x_1,y_1,x_2,y_2) = f_1(x_1,y_1)f_2(x_2,y_2)$, where f_1, f_2 are the marginal joint pdf's of the two pairs respectively obtained by integrating out the other variables from $\bar{f}$. this implies (in the same way as in the lecture) that any event formulated in terms of the first pair (e.g., $X_1 < Y_1$) is independent of any event formulated in terms of the second pair (e.g., $|X_2/Y_2| > 3$). It also implies that X_1, X_2 are independent as well as Y_1, Y_2.

It may also be mentioned that the fact that the parameters of interest do not occur in the marginal pdf f_1 sometimes is used as one (of several possible) statistical definitions of *X* being "exogenous".

In this situation we may simplify the conditional joint f further. First notice that the model (8) implies that $X_1, X_2, \ldots, X_n$ are iid rv's with pdf

$$(12) \qquad f_1(x_1, x_2, \ldots, x_n) = f_2(x_1) f_2(x_2) \cdots f_2(x_n)$$

where $f_2(x_i)$ is the (common) marginal pdf of X_i for $i = 1, 2, \ldots, n$. Using (9)-(11) we get

$$(13) \qquad \begin{aligned} &f(y_1, y_2, \ldots, y_n \mid x_1, x_2, \ldots, x_n) = \frac{\overline{f}(x_1, y_1, x_2, y_2, \ldots, x_n, y_n)}{f_1(x_1, x_2, \ldots, x_n)} = \\ &= \frac{f(x_1, y_1) f(x_2, y_2) \cdots f(x_n, y_n)}{f_2(x_1) f_2(x_2) \cdots f_2(x_n)} = f(y_1 \mid x_1) f(y_2 \mid x_2) \cdots f(y_n \mid x_n) \end{aligned}$$

showing that, given all $X_i = x_i$ for $i = 1, 2, \ldots, n$, then $Y_1, Y_2, \ldots, Y_n$ are independent but with different pdf's

$$Y_i \mid x_1, x_2, \ldots, x_n \sim f(y_i \mid x_i)$$

or

$$f(y_i \mid x_1, x_2, \ldots, x_n) = f(y_i \mid x_i)$$

which exactly represents the model given in (4) and shows that we can drop all x_j's except x_i from the distribution of Y_i - which is actually due to the independence of $X_1, X_2, \ldots, X_n$.

From this we may draw even further important conclusions. If η is one of the three parameters, $\alpha, \beta,$ and $\mu(x) = \alpha + \beta x$, the theory for the model gives that the OLS estimators satisfy

$$T = \left. \frac{\hat{\eta} - \eta}{\mathrm{SE}(\hat{\eta})} \right| x_1, x_2, \ldots, x_n \quad \sim t_{n-2}$$

which we use for inference about η. Since the t_{n-2} distribution does not depend on $x_1, x_2, \ldots, x_n$, we can conclude that *T* is even marginally (without the conditioning) t_{n-2} distributed[4] ! *So confidence intervals and tests about η, developed under the fixed values assumption for $X_1, X_2, \ldots, X_n$, are still valid without the conditioning.*

[4] Remember from the lectures that if $(U, V) \sim f(u, v)$, and the conditional pdf, $f(v \mid u)$ does not depend on *u*, then *U* and *V* must be independent and $f(v \mid u) = f_2(v)$, i.e., the marginal pdf of *V*. This is true (why?) even if *U* consists of several variables.

However, the distribution of $\hat{\eta}$ itself may depend on the conditioning. For example, from the theory of model (4), the OLS estimator $\hat{\beta}$ satisfies (writing $\underline{x}$ for $x_1, x_2, \ldots, x_n$),

$$\text{(14)} \quad \hat{\beta} \mid \underline{x} \quad \sim N\left(\beta, \text{var}(\hat{\beta} \mid \underline{x})\right) = N\left(\beta, \ \frac{\sigma^2}{(n-1)s_x^2}\right)$$

implying that the marginal distribution of $\hat{\beta}$ no longer is normal. On the other hand, (14) implies that

$$E(\hat{\beta} \mid \underline{x}) = \beta \ \text{ (i.e., a constant function of } \underline{x}\text{)}$$

so the theorem of total expectation gives

$$E(\hat{\beta}) = E\left[E(\hat{\beta} \mid \underline{X})\right] = E(\beta) = \beta$$

$\hat{\beta}$ is therefore still unbiased! However, the variance becomes complicated

$$\text{var}(\hat{\beta}) = E\left[\text{var}(\hat{\beta} \mid \underline{X})\right] + \text{var}\left[E(\hat{\beta} \mid \underline{X})\right] = E\left[\frac{\sigma^2}{(n-1)S_x^2}\right] + \text{var}\left[\beta\right] = \frac{\sigma^2}{(n-1)} E\left[\frac{1}{S_x^2}\right]$$

which is complicated. In practice, however, we may look at the expression, $\frac{\sigma^2}{(n-1)} \cdot \frac{1}{S_x^2}$, as an (unbiased) estimator of $\text{var}(\hat{\beta})$.

Appendix 2 (optional reading). Some comments about exogenous variables in iid models

The econometrical consequence of a variable being exogenous is that we can consider this variable as having fixed values (e.g., equal to the observed values) in an econometric model – which often simplifies the estimation problem considerably.

The common reason presented for this is the factorization (10):

$$\text{(15)} \quad \overline{f}(x_1, y_1, x_2, y_2, \ldots, x_n, y_n) = f(y_1, y_2, \ldots, y_n \mid x_1, x_2, \ldots, x_n) \cdot f_1(x_1, x_2, \ldots, x_n)$$

or, in terms of log likelihood,

$$\text{(16)} \quad l(\theta) = \ln \overline{f} = \ln f(y_1, y_2, \ldots, y_n \mid x_1, x_2, \ldots, x_n) + \ln f_1(x_1, x_2, \ldots, x_n)$$

combined with the maximum likelihood principle that, (in a certain sense and under some general conditions), utilizes all available information in the data for estimation.

When the parameters of interest (α, β, σ here) *only occur in the first term* ($\ln f(y_1, y_2, \ldots, y_n \mid x_1, x_2, \ldots, x_n)$) on the right hand side of (16), then maximizing the full joint pdf $\overline{f}$ with respect to (w.r.t.) α, β, σ, is equivalent to just maximizing the $\ln f$ on the right side w.r.t. α, β, σ (note that the derivative of the marginal $\ln f_1$ on the right w.r.t. any of the parameters of interest α, β, σ, must then be 0).

Hence, the maximum likelihood principle reduces to maximizing the conditional likelihood, $f(y_1, y_2, \ldots, y_n \mid x_1, x_2, \ldots, x_n)$, where the rv's $X_1, X_2, \ldots, X_n$ are fixed to the observed values, $x_1, x_2, \ldots, x_n$. This is what we mean (statistically) when we say that $X_1, X_2, \ldots, X_n$ (or the corresponding population variable *X*) are (is) exogenous.
However, in order to reach this conclusion, we must be sure that the differentiation of the marginal $\ln f_1(x_1, x_2, \ldots, x_n)$ w.r.t. any of the parameters of interest is 0, i.e., when the other parameters occurring in $\ln f_1(x_1, x_2, \ldots, x_n)$ in no way depend on the parameters of interest – i.e., when the two sets of parameters are, as we call it, *variation independent*, which they are in fact in the present model. The variation independence is shown by the following:

The original parametrization of the bivariate normal population pdf is $\psi = (\mu_x, \sigma_x, \mu_y, \sigma_y, \sigma_{xy})$, where $\sigma_{xy} = \text{cov}(X, Y) = \text{cov}(X_i, Y_i)$. These parameters may take any values as long as the restrictions, $\sigma_x > 0,\ \sigma_y > 0,\ |\rho| \leq 1,$ are fulfilled, (where the correlation is $\rho = \sigma_{xy} \big/ (\sigma_x \sigma_y)$).

The model that we actually use for the mother/daughter data, represents a reparametrization of the distribution with the new parameter vector given by

$$\theta = (\overset{\theta_1}{\alpha, \beta, \sigma}, \overset{\theta_2}{\mu_x, \sigma_x}) = (\theta_1, \theta_2)$$

Here θ_1 represents our parameters of interest and θ_2 the other parameters. That θ_1 and θ_2 are variation independent should be clear considering that the relationship between θ and ψ is one-to-one (written $\theta \leftrightarrow \psi$):

> [**Because:** (i) The relation $\psi \to \theta$: Given any values of $\psi = (\mu_x, \sigma_x, \mu_y, \sigma_y, \sigma_{xy})$, we get unique values of $\theta = (\alpha, \beta, \sigma, \mu_x, \sigma_x)$ by (as seen before),
>
> $\sigma^2 = \text{var}(e_i) = \text{var}(Y_i) = \text{var}(Y) = \sigma_y^2, \qquad \beta = \dfrac{\sigma_{xy}}{\sigma_x^2}$, and $\qquad \alpha = \mu_y - \beta\mu_x$
>
> (ii) The relation $\theta \to \psi$: Given any values of $\theta = (\alpha, \beta, \sigma, \mu_x, \sigma_x)$, we get unique values of $\psi = (\mu_x, \sigma_x, \mu_y, \sigma_y, \sigma_{xy})$ by
>
> $\sigma_y^2 = \sigma^2, \quad \sigma_{xy} = \beta \cdot \sigma_x^2,$ and $\mu_y = \alpha + \beta\mu_x$
>
> From this we can conclude that the parameters characterizing the distribution of mother's height ($\theta_2 = (\mu_x, \sigma_x)$) can be anything without affecting α, β, σ.
>
> (**End of proof**)]

This argumentation enables us to give **a formal definition**

If (i) the parameters of interest (θ_1) occur in the conditional distribution, $f(y_1, y_2, \ldots, y_n \mid x_1, x_2, \ldots, x_n)$ only, and (ii) the parameters of interest, θ_1, are variation independent of the other parameters, θ_2, we say that $X_1, X_2, \ldots, X_n$ (or the corresponding population variable *X*) *are (weakly) exogenous*.

[This implies that $X_1, X_2, \ldots, X_n$ can be considered having fixed values (i.e., the observed values) when making inference about the parameters of interest, and focus our attention on the conditional pdf, $f(y_1, y_2, \ldots, y_n \mid x_1, x_2, \ldots, x_n)$ (which, in our situation (see (13)) reduces to $f(y_1 \mid x_1) f(y_2 \mid x_2) \cdots f(y_n \mid x_n)$].

Further comments.

- In time series data (where the index (*i*) represents time, and where there are dependence between the pairs (X_i, Y_i)), we operate with different concepts (degrees) of exogeneity (e.g.,(a) *weak* exog., (b)*strong* exog., (c)*super* exog.) depending on what the model is used for (e.g., respectively (a) inference of parameters of interest, (b) factual forecasting (based on future predicted X_i's), (c) contra-factual forcasting (based on future X_i-values set arbitrarily).
- In the iid case (with independent pairs as here) the different exogeneity concepts collapse to the same concept as given in the definition above.