TA session 1

The identification problem

Identification example: your own movements and your movements in a mirror – which drives which or do both move due to an external stimulus? This reflection problem (Manski, 1993) arises if you try to interpret the common observation that individuals belonging to the same group tend to behave similarly. Three hypotheses have been proposed to explain this phenomenon:

1. *endogenous effects*: 'propensity of an individual to behave in some way varies with the prevalence of the behaviour in the group'.
2. *contextual effects*: 'propensity of an individual to behave in some way varies with the distribution of background characteristics in the group'.
3. *correlated effects*: 'individuals in the same group tend to behave similarly because they face similar environments or have similar individual characteristics'.

Why would we care about what generates observed patterns of group behaviour? One reason is that different processes have different ramifications for *public policy*. Data alone cannot reveal which hypothesis might be correct, so to draw conclusions we need to combine empirical evidence (data) with assumptions. This is an *identification problem.*

Manski (2007) distinguishes identification and statistical inference as follows:

> 'Studies of identification seek to characterize the conclusions that could be drawn if one could use the sampling process to obtain an unlimited number of observations. Studies of statistical inference seek to characterize the generally weaker conclusions that can be drawn from a finite number of observations.' (Manski, 2007: 3)

Logically, identification precedes inference much like the study of probability precedes that of statistics. Koopmans (1949: 132) introduced the term 'identification' into econometric literature as follows.

> 'In our discussion we have used the phrase "a parameter that can be determined from a sufficient number of observations." We shall now define this concept more sharply, and give it the name *identifiability* of a parameter. Instead of reasoning, as before, from "a sufficiently large number of observations" we shall base our discussion on a hypothetical knowledge of the probability distribution of the observations, as defined more fully below. It is clear that exact knowledge of this probability distribution cannot be derived from any finite number of observations. Such knowledge is the limit approachable but not attainable by extended observation. By hypothesizing nevertheless the full availability of such knowledge, we obtain a clear separation between problems of statistical inference arising from the variability of finite samples, and problems of identification in which we explore the limits to which inference even from an infinite number of observations is suspect.'

Data and assumptions lead to conclusions. To paraphrase Manski, we can only overcome identification problems by making stronger assumptions or by initiating new sampling processes that yield different kinds of data rather than gathering more of the same kind of data.

Law of diminishing credibility

Definition. *The Law of Diminishing Credibility*: the credibility of inference decreases with the strength of the assumptions made. (Manski, 2007: 3)

Extrapolation

The problem of *extrapolation* is that of predicting off the *support* where we say that a covariate value x_0 is *on the support of the distribution* $P(x)$ if there is positive probability of observing x arbitrarily close to x_0; a covariate value x_0 is *off the support* of $P(x)$ if there is zero probability of observing x within some neighbourhood of x_0. Mathematically, t is *in the support of* P if:

$$P[t-\delta \leq y \leq t+\delta] > 0 \, \forall \delta > 0$$

Consider the problem of extrapolation – i.e. prediction off the support, in particular the problem of predicting a random variable y conditional on x where x only takes values at $\{0, 1, 2, 3, 5, 6, 7\}$. We can compute a tighter confidence interval for the mean of $y|x$ with 1000 observations of (y, x) than we can with 100 observations. Figures 1 and 2 represent each case. The width of the confidence intervals relates to a statistical problem, since we can estimate $E(y|x)$ more precisely with more data. However, at $x = 5$, the confidence interval is infinite irrespective of sample size, which means we are dealing with an identification problem.

External validity

A related concept to extrapolation is generalisability or *external validity*.

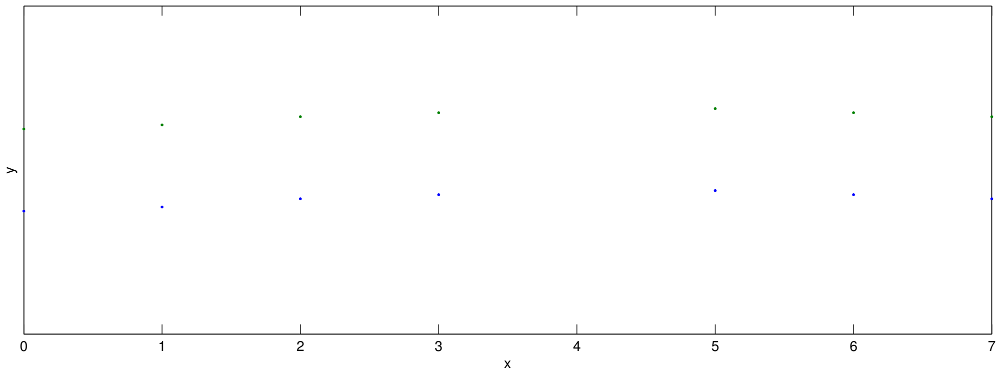

Figure 1: Confidence interval for $E(y|x)$ when $n = 100$.

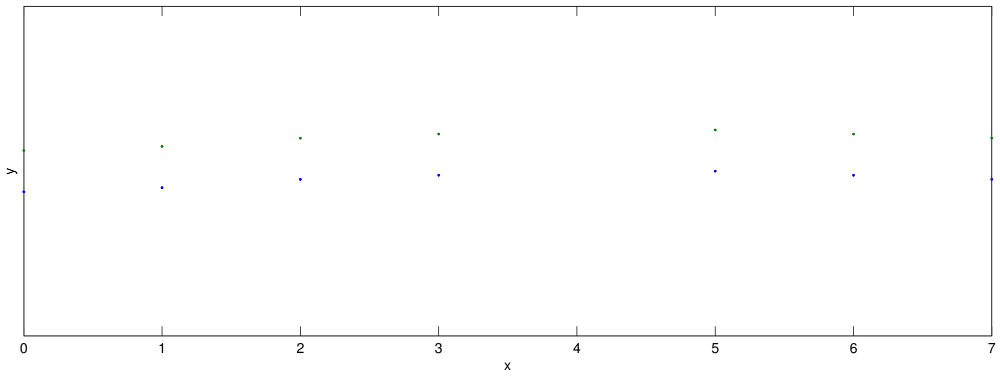

Figure 2: Confidence interval for $E(y|x)$ when $n = 1000$.

Definition. An experiment is said to have *external validity* if the distribution of outcomes realized by a treatment group is the same as the distribution of outcomes that would be realized in an actual program. (Manski, 2007:27)

Definitions of identification

Consider the model

$$y = x'\beta + \epsilon$$
$$E(\epsilon|x) = 0$$

where x is $k \times 1$, alternatively defined by

$$E[y|x] = x'\beta$$

Suppose we have data $(y_i, x_i)_{i=1}^{N}$ where N is extremely large and an independently and identically distributed (iid) sample. Here with identification, we always work under the assumption that you can observe the population (i.e. $N = \infty$) so $E(y|x) = x'\beta$ is 'known'. Our question relates to the identification of β in the model above with given data; what can we learn about β?

$$\left\{\begin{matrix}\text{Model}\\ +\\ \text{Data}\end{matrix}\right\} \overset{\substack{\text{Identification}\\ \text{Analysis}}}{\Longrightarrow} \text{ Information about } \beta$$

Definition. A parameter $b \in \mathbf{R}^k$ is *identified relative to* β if $P_X\{x : x'b \neq x'\beta\} > 0$.

Definition. In the model above, β is *point identified* if $\forall b \neq \beta$, b is identified relative to β.

The following is an alternative definition of a parameter begin identified.

Definition. The parameter vector $\boldsymbol{\theta}_0$ is *identified* if for any other parameter vector $\boldsymbol{\theta} \in \Theta$, the set

$$\{y | f(y|\boldsymbol{\theta}) \neq f(y|\boldsymbol{\theta}_0)\}$$

has positive probability.

Example (How to verify or check for identifiability)**.** Suppose that we observe (x, y) where

$$y = \begin{cases} 0 & \alpha x + \epsilon < \gamma \\ 1 & \alpha x + \epsilon \geq \gamma \end{cases}$$

where $x \perp\!\!\!\perp \epsilon$, $\epsilon \sim N(\mu, \sigma^2)$ and x has some known distribution, which doesn't depend on any of $\alpha, \gamma, \mu, \sigma^2$.

(a) We claim first that $\theta = (\alpha, \gamma, \mu, \sigma^2)$ is not identified:

Proof. We have to check whether, for any θ, there exists $\theta' \neq \theta$ s.t. $\forall x$, $P(y = 0|x; \theta) = P(y = 0|x; \theta')$.

Note that we can restrict ourselves only to conditional distributions instead of the multivariate ones, since we know the distribution of x.

Then $P(y = 0|x; \theta) = P(\alpha x + \epsilon < \gamma|x) = P(\frac{\epsilon-\mu}{\sigma} < \frac{\gamma-\alpha x-\mu}{\sigma}|x) = \Phi(\frac{\gamma-\alpha x-\mu}{\sigma})$ where Φ is the cumulative normal distribution function.

To see that θ is not identified, take $\theta' = (k\alpha, k\gamma, k\mu, k^2\gamma^2)$, for some $k > 0$. Then we have $P(y = 0|x; \theta) = P(y = 0|x; \theta')$. □

(b) Thus, first normalize $\sigma^2 = 1$. Then the parameters (α, γ, μ) are still not identified (although α alone is identified):

Proof. Take $\theta' = (\alpha, \gamma + A, \mu + A)$ and we see that $\Phi(\gamma + A - \alpha x - \mu - A) = \Phi(\gamma - \alpha x - \mu)$. □

(c) If we normalise again $\gamma = 0$, then (α, μ) are finally identified:

Proof. Suppose that $\forall x, \Phi(-(\alpha x + \mu)) = \Phi(-(\alpha' x + \mu'))$. Since Φ is 1-1, it follows that, $\forall x, \alpha x + \mu = \alpha' x + \mu' \implies (\alpha - \alpha')x = \mu' - \mu$. This implies that $\alpha = \alpha'$ and $\mu' = \mu$. □

Conditional prediction

The joint probability (frequency) distribution of $(y, x \in Y \times X)$ across population is $P(y, x)$. A person is drawn at random from the subpopulation of people with a specified value of x. The problem is to predict his value of y. $P(y|x)$ can be interpreted as:

1. the distribution of y conditional on x, viewed as a function of x;
2. the distribution of y conditional on x, evaluated at a specified value of x;
3. the probability that y takes a given value conditional on x viewed as a function of x;
4. the probability that y takes a given value conditional on x evaluated at a specified value of x.

The particular interpretation will be clear from the context and sometimes I may write $P(y)$.

Estimation of best predictors from random samples

Whenever we can observe (y, x) at random from the population of interest, we might ask how we can learn about the conditional distribution $P(y|x)$ or at least the value of a best predictor of y given x. Even if we assume nothing about the form of the distribution $P(y, x)$, random sampling will reveal $P(y, x)$. For studying conditional prediction, we will now look at empirical distributions and illustrate use of some of the above concepts, in addition to the analogy principle, which loosely implies using sample statistics for population counterparts and then calling on results from asymptotic theory to justify these sample statistics. The empirical distribution $P_N(y, x)$ is the sample analog and natural estimate of $P(y, x)$. It is a multinomial distribution placing equal mass $\frac{1}{N}$ on each of N observations $[(y_i, x_i) : i = 1, \ldots, N]$; if a particular value of (y, x) recurs in the data, it receives multiple $\frac{1}{N}$ weights. It is natural to use $P_N(y, x)$ to estimate $P(y, x)$ since the empirical distribution estimates the probability $P[(y, x) \in A]$ that $E(y, x)$ falls in some set A by estimating the fraction of observations of (y, x) that fall in the set A:

$$P_N[(y, x) \in A] = \frac{1}{N} \sum_{i=1}^{N} 1[(y_i, x_i) \in A] \xrightarrow{\text{as}} P[(y, x) \in A]$$

where the convergence follows by the SLLN; more specifically, it can be shown that (ICBST) the empirical distribution for this probability converges almost surely to $E[1[(y, x) \in A]]$, which turns out to be $P[(y, x) \in A]$; the proof of the second part of this statement is as follows:

Proof.

$$\begin{aligned} E[1[(y, x) \in A]] &= 1.P[(y, x) \in A] + 0.P[(y, x) \notin A] \\ &= P[(y, x) \in A] \end{aligned} \qquad \square$$

This essentially means that we can interpret probability as the expectation of an indicator function. Since with random sampling, we can learn $P[(y, x) \in A]$ even if we knew nothing before about it's value and this holds for every set A, we can learn the distribution $P(y, x)$. Let us now look at the three cases. The lesson from this section will be that we can only do non-parametric estimation on the support.

1. x_0 is on the support of $P(x)$ and $P(x = x_0) > 0$.
2. $P(x = x_0) = 0$ but x_0 is on the support.
3. x_0 is off the support of $P(x)$.

In the first case where covariates have positive probability, i.e. when $P(x = x_0) > 0$, we have that the conditional empirical probability is given by:

$$
\begin{aligned}
P_N(y \in B | x = x_0) &= \frac{\sum_{i=1}^N 1[y_i \in B, x_i = x_0]}{\sum_{i=1}^N 1[x_i = x_0]} \\
&= \frac{\frac{1}{N}\sum_{i=1}^N 1[y_i \in B, x_i = x_0]}{\frac{1}{N}\sum_{i=1}^N 1[x_i = x_0]} \qquad (1)
\end{aligned}
$$

Observe that the numerator converges almost surely to $P(y \in B, x = x_0)$ by SLLN and the denominator converges almost surely to $P(x = x_0)$, which is positive in this case. So, by the Contraction Mapping Theorem and Bayes Theorem, the RHS of (1) converges almost surely to $P(y \in B|x = x_0)$:

$$
\frac{P(y \in B, x = x_0)}{P(x = x_0)} \overset{\text{Bayes}}{=} P(y \in B|x = x_0)
$$

which holds for every set B so we can learn about the conditional distribution $P(y|x = x_0)$. So, (sample) empirical quantiles (e.g. mean and median) converge to population quantiles. Also remember that for the SLLN and CMT, we need functions to be continuous.

Example.

$$
E_N(y|x = x_0) = \frac{\sum_{i=1}^N y_i \cdot 1[x_i = x_0]}{\sum_{i=1}^N 1[x_i = x_0]} = \frac{\frac{1}{N}\sum_{i=1}^N y_i \cdot 1[x_i = x_0]}{\frac{1}{N}\sum_{i=1}^N 1[x_i = x_0]} \qquad (2)
$$

The numerator of the RHS in (2) converges almost surely to $E(y \cdot 1[x = x_0])$ as N increases by SLLN, which equals $E(y|x = x_0)P(x = x_0)$ and the denominator converges to $P(x = x_0)$. Given that $P(x = x_0) > 0$, by the CMT:

$$
E_N(y|x = x_0) \xrightarrow{\text{a.s.}} E(y|x = x_0)
$$

Similarly, $M_N \xrightarrow{\text{a.s.}} M$.

In the second case where covariates have zero probability, i.e. when $P(x = x_0) = 0$ but x_0 on the support of $P(x)$ (e.g. continuous distributions). Let $\rho(x_i, x_0)$ measure distance between covariate of interest x_0 and an observed value x_i. When x is scalar, $\rho(x_i, x_0) = |x_i - x_0|$. When x vector, it could be any reasonable measure of distance between x_i and x_0, e.g. Euclidean distance between these vectors. Let d_N denote *bandwidth*. The subscript N on d_N indicates that bandwidth is a function of sample size. Estimate $E(y|x = x_0)$ by the sample mean of y among the observations for which $\rho(x_i, x_0) < d_N$.

Definition. The *local average* or *uniform kernel estimate* is:

$$
\theta_N(x_0, d_N) \equiv E_N(y|x = x_0) = \frac{\sum_{i=1}^N y_i \cdot 1[\rho(x_i, x_0) < d_N]}{\sum_{i=1}^N 1[\rho(x_i, x_0) < d_N]}
$$

where d_N is a sample-size dependent *bandwidth* selected by the researcher conveying the idea of restricting attention to observations where x_i is near x_0; sometimes d_N is written as $d_N(x_0)$ and called the *local bandwidth selection* to emphasise the case where we do not use the same bandwidth everywhere.

A basic finding of modern nonparametric regression analysis is that the uniform kernel estimate $E_N(y|x = x_0) \xrightarrow{\text{a.s.}} E[y|x = x_0]$ provided the following four conditions hold:

1. $E(y|x)$ varies continuously with x near x_0.

2. $V(y|x)$ bounded for x near x_0.

3. Tighten bandwidth d_N as sample size N increases.

4. Do not tighten bandwidth d_N too rapidly as sample size N increases.

Conditions (i) and (ii) are minimum regularity conditions; we can always choose bandwidths so (iii) and (iv) hold.

Remark (Curse of Dimensionality)**.** We should choose the bandwidth d_N so MSE tends to zero, i.e. variance and bias tend to zero. Large d_N is good for variance in that we get a big sample and so standard deviation reduces at rate $\sqrt{n}$, but then the bias could be large, i.e. $E(y|\rho(x, x_0) < d_N) - E_N(y|x = x_0)$ could be large when d_N is large. When d_N is small, however, variance increases because there are fewer observations in each cell. The *curse of dimensionality* rears its head in that a larger dimension for x does not affect bias but it does affect variance because there tends to be fewer observations lying inside bandwidths of radius d_N, so variance is still high. With non-parametric estimation, the price to pay is generally that the rate of convergence will be slower than $\sqrt{n}$. Stone (1981) showed that the best rate of convergence you can achieve from non-parametric estimation gets tougher as the dimension of x increases. So the curse of dimensionality takes the following form: the best achievable rate of convergence diminishes as the dimension of covariates increases. Of course, one solution would be to look at semi-parametric models such as linear index models, e.g. the regression of $E(y|x) = g(x)$ where we only know that g is continuous; say x is in a three dimension space, then $g(x) = g(x_1, x_2, x_3) = g'(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)$ for a linear index model where $g \neq g'$, i.e. linear index models wipe out the curse of dimensionality; note that semi-parametric models make assumptions about the regression functions: g is parametric while g' is non-parametric.

In the third case, x_0 is off the support of $P(x)$, i.e.

$$\exists d_0 > 0 : P[\rho(x_i, x_0) < d_0] = 0$$

e.g. when $x_0 = 5$ in figure 1. One can only do non-parametric estimation on the support. Now, data alone reveal nothing about $P(y|x = x_0)$. So, we are in the case of *extrapolation*: predicting y when x_0 is off the support, i.e. making predictions away from data. We need *global* assumptions for identifying power in the case of extrapolation, i.e. we need to restrict $E(\cdot|\cdot)$ globally. We can invoke *invariance* assumptions, such that y behaves the same at x_0 as at x_1 on the support of $P(x)$, i.e.

$$P(y|x = x_0) = P(y|x = x_1)$$

Definition. Using a *counterfactual* entails expressing what has not happened but what might or would happen if circumstances, i.e. data were different.

Example. What would the consequences be for the US had the Paulson plan not been enacted? What would the consequences have been for Europe had we not decided to bail out the banks? Surely we would be better off? What about if Senator John McCain had been president instead of Barack Obama? These are hypothetical situations. This has much to do with the inherent problem in economics that experiments are not as readily available as in the natural sciences.

Example (Predicting Criminality)**.** Selective incapacitation implies that sentencing of convicts should be linked with predictions regarding their future criminality. The RAND study by Greenwood & Abrahamse (1982) found that using a sample of 2200 prison and jail inmates in 1978 across California, Michigan and Texas, those with backgrounds such as previous convictions, drug use and unemployment predict high rates of future offenses and part of their research team then suggested that those with such backgrounds should receive longer prison terms. This was very controversial, especially when this prediction approach became part of a legislative proposal for selective incapacitation. The part of the controversy that relates to econometrics concerns the external validity from the RAND results to other groups, places and sentencing policies. The findings hold for this particular cohort of prisoners in these three states for the given sentencing policies, but that does not imply that they would still apply to other cohorts of prisoners in other states or to criminals who would be sentenced under alternative policies such as selective incapacitation.

Remark. Sometimes, hopefully rarely, researchers misinterpret correlation with causation. However, if treatments are randomly assigned, then causation becomes more justifiable. For instance, if there were no selection issues and people were truly randomly assigned to different treatments, then if data showed that outcomes under one treatment were 'better' than outcomes under another treatment, we could conclude that there is a causal link: the first treatment improves the outcome relative to the second.

Failures off the support are inherently not detectable. Theory becomes important. The first function of theory is to allow extrapolation and the second function function of theory is to improve the sampling precision for estimation on the support. Causal interpretation can be 'dodgy', while prediction is usually better. Unfortunately, cases where theory is most testable are generally least needed – learning $P(y|x)$ on the support of $P(x)$, while cases where theory is least testable are generally most needed – learning $P(y|x)$ off the support of $P(x)$.

Returning to the local average estimate, a more general kernel estimator is the following.

Definition. The *local weighted average* or *kernel estimate* is

$$E_N(y|x = x_0) = \frac{\frac{1}{N}\sum_{i=1}^{N} y_i K\left[\frac{\rho(x_i,x_0)}{d_N}\right]}{\frac{1}{N}\sum_{i=1}^{N} K\left[\frac{\rho(x_i,x_0)}{d_N}\right]}$$

where $0 \leq K(\cdot)$ is inversely related to $\rho(x_i, x_0)/d_N$. The uniform kernel can be seen to be a special case where

$$1\left[\frac{\rho(x_i, x_0)}{d_N} < 1\right]$$

and here all observations get used and are given the same weight, hence the name 'uniform' kernel.

Remark. Choosing bandwidth d_N can be extremely subjective. Best practice is to report multiple estimates or use data dependent automated rules to choose the bandwidth, e.g. cross-validation. Cross-validation involves fixing the bandwidth, estimating the regression on each of the N possible subsamples (each of size $N - 1$ and then in each case we use the estimate to predict $y|x$ for the observation that was left out. The resulting bandwidth, the cross-validated bandwidth yields the best predictions of the left-out values of y. Increasing the bandwidth typically reduces variance but increases bias, which is not good for the MSE – this is another manifestation of the curse of dimensionality as N increases.

Best predictors under given loss functions

Recall from lectures that the *best predictor* p of the random variable Y given other random variables X minimises a *loss function* $\mathcal{L}(\cdot)$. We usually want to minimise the expected loss conditional on the random variates X:

$$\min E[\mathcal{L}(y-p)|x]$$

The *best predictor* solves this minimisation problem, so choosing a best predictor is a decision problem whose solution depends on the objective, i.e. the best predictor is determined by $\mathcal{L}(.)$ and $P(y|x)$. It can be shown that when the loss function is a square loss function, the best predictor is the mean. In this case, the solution may not exist, however. It can also be shown that with absolute loss functions, the best predictor is the median. In what follows, let $u \equiv y - p$.

Lemma. *Under square loss, the best predictor is the mean.*

Proof. Let $L(\cdot)$ be the square loss function, i.e. $\mathcal{L}(u) = u^2$, $\mu = E(y)$ and $\mu \neq \theta \in \mathbb{R}$. Then

$$\begin{aligned}
E(y-\theta)^2 &= E[(y-\mu)+(\mu-\theta)]^2 \\
&= E(y-\mu)^2 + (\mu-\theta)^2 + 2(\mu-\theta)E(y-\mu) \\
&= E(y-\mu)^2 + (\mu-\theta)^2 > E(y-\mu)^2
\end{aligned}$$

Therefore, μ uniquely minimises the expected loss. □

Lemma. *Under absolute loss, the best predictor is the median.*

Proof. Let $\mathcal{L}(\cdot)$ be the absolute loss function, i.e. $\mathcal{L}(u) = |u|$ and $m \equiv \min\{\theta : P(y \leq \theta) \geq \frac{1}{2}\}$ be the median of y where $m \in \mathbb{R}$. Let us first compare the expected loss at m with that at any $\theta < m$:

$$\begin{aligned}
E[|y-\theta|] - E[|y-m|] &= E[|y-\theta| - |y-m|] \\
&= (\theta-m)P(y \leq \theta) \\
&\quad + E[2y - (\theta+m)|\theta < y < m]P(\theta < y < m) \\
&\quad + (m-\theta)P(y \geq m) \\
&\geq (\theta-m)P(y \leq \theta) + (\theta-m)P(\theta < y < m) \\
&\quad + (m-\theta)P(y \geq m) \\
&= -(m-\theta)P(y < m) + (m-\theta)P(y \geq m) \\
&= (m-\theta)[P(y \geq m) - P(y < m)]
\end{aligned}$$

Since by definition $P(y < m) \leq \frac{1}{2}$, the final expression is nonnegative. Finally, let us compare the expected loss at m with that at any $\theta > m$:

$$\begin{aligned}
E[|y-\theta|] - E[|y-m|] &= E[|y-\theta| - |y-m|] \\
&= (\theta-m)P(y \leq m) \\
&\quad + E[(\theta+m) - 2y|m < y < \theta]P(m < y < \theta) \\
&\quad + (m-\theta)P(y \geq \theta) \\
&\geq (\theta-m)P(y \leq m) + (m-\theta)P(m < y < \theta) \\
&\quad + (m-\theta)P(y \geq \theta) \\
&= (\theta-m)P(y \leq m) - (\theta-m)P(m < y) \\
&= (\theta-m)[P(y \leq m) - P(m < y)]
\end{aligned}$$

Since by definition $P(y \leq m) \geq \frac{1}{2}$, the final expression is nonnegative. □

Stata

`kernreg` is a nonparametric kernel estimator contained within Arie Beresteanu & Charles Manski's 'bounds' package for Stata. In 1, we estimate $E(y|m = 1)$ using a uniform kernel with a bandwidth of 0.5. In 2, we estimate $E(y|m = 1)$ using a Gaussian kernel with the same bandwidth.

```
kernreg y m, g(yubar) at(1) w1(0.5) rec
```

Listing 1: Stata code for kernreg with a uniform kernel

```
kernreg y m, g(yubar) at(1) w1(0.5) gau
```

Listing 2: Stata code for kernreg with a Gaussian kernel

Bounding quantities

The following two theorems, the law of total probability and the law of iterated expectations are highly useful for the study of identification. We can use them to express bounds on quantities such as probabilities and expectations.

Law of total probability

Definition (LTP)**.** Recall from lectures that the for a *sample space* Ω, which is the set of all possible outcomes from an experiment, elements of the sample space called *events* $B_1, \ldots, B_N$ where $B_i \in \Omega$ for all i form a *partition* of Ω if the following two conditions are met:

1. $B_i \cap B_j = \emptyset$ for all $i \neq j$, i.e. events do not overlap – this property is called *pairwise disjointness.*
2. $\cup_i B_i = \omega$, i.e. the collection of all the events *covers* the sample space exactly.

Theorem 1. *Let $B_1, B_2, \ldots, B_n$ be a partition of the sample space Ω such that $P(B_i) > 0$ for all events in the partition. Then for any event A, we have the law of total probability:*

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

The *law of total probability* known by other names such as the *partition law* or the *law of the extension of conversation* states that for any events A and B

$$P(A) = P(A \cap B) + P(A \cap B^c)$$

For discrete random variables

$$P(Y = y) = \sum_{x \in Supp(X)} P(Y = y|X = x)P(X = x)$$

Law of iterated expectations

Theorem 2 (LIE)**.** *The law of iterated expectations (LIE) can be stated as*

$$E(Y) = E_X(E(Y|X)) = \sum_{x \in Supp(X)} E(Y|X = x)P(X = x)$$

Incomplete (missing) data

Let (y, z, x) be such that y is an outcome to be predicted, x are covariates and define

$$z = \begin{cases} 1 & \text{if } y \text{ is observed} \\ 0 & \text{else} \end{cases}$$

Draw N people at random from population. For each $i = 1, \dots, N$, the outcome y_i is observable if $z_i = 1$ and missing if $z_i = 0$. The objective is to use available data to learn about $P(y|x)$ at a specified value of x on $Supp(P(x))$. We can use the LTP to express the missing data problem more clearly:

$$P(y|x) = P(y|x, z = 1)P(z = 1|x) + \underbrace{P(y|x, z = 0)}_{\text{missing}} P(z = 0|x)$$

The missing $P(y|x, z = 0)$, which is unknown implies that we are dealing with an identification problem. Denote $P(y|x, z = 0) = \gamma \in \Gamma_Y$, the identification region is:

$$H[P(y|x)] = [P(y|x, z = 1)P(z = 1|x) + \gamma P(z = 0|x), \gamma \in \Gamma_Y]$$

and note that

$$P(z = 0|x) < 1 \Longrightarrow H[\cdot] \subsetneq \Gamma_Y$$

where Γ_Y denotes the set of all probability distributions on the set Y. H is a proper subset of Γ_Y, i.e. $H \subsetneq \Gamma_Y$ when $P(z = 0|x) < 1$ and is the single distribution $P(y|x, z = 1)$ when $P(z = 0|x) = 0$. Therefore, $P(y|x)$ is *partially identified* when $0 < P(z = 0|x) < 1$ and is *point identified* when $P(z = 0|x) = 0$.

Empirical research often has an objective of inferring a parameter of the outcome distribution, e.g. $E(y|x)$. Let $\theta(\cdot)$ be a function mapping probability distributions on Y into $\mathbb{R}$ and consider the parameter $\theta[P(y|x)]$. The identification region for this parameter is the set of all values it may take when $P(y|x)$ varies over all of its feasible values, so $H\{\theta[P(y|x)]\} = \{\theta(\eta), \eta \in H[P(y|x)]\}$.

Now looking at the identification of event probabilities, we can once again use the LTP to express the missing data problem more clearly:

$$P(y \in B|x) \overset{\text{LTP}}{=} P(y \in B|x, z = 1)P(z = 1|x) + \underbrace{P(y \in B|x, z = 0)}_{\in[0,1]} P(z = 0|x)$$

The worst case bound on $P(y \in B|x)$:

$$\begin{aligned} P(y \in B|x, z = 1)P(z = 1|x) &\leq P(y \in B|x) \\ &\leq P(y \in B|x, z = 1)P(z = 1|x) + P(z = 0|x) \end{aligned} \tag{3}$$

Equivalently, letting L_B and U_B denote lower and upper bounds, respectively, we have that

$$\begin{aligned} L_B :& \quad P(y \in B|x, z = 1)P(z = 1|x) \\ U_B :& \quad P(y \in B|x, z = 1)P(z = 1|x) + P(z = 0|x) \end{aligned}$$

U_B and L_B are the largest and smallest feasible values of $P(y \in B|x)$ and hence are called *sharp* bounds:

$$H[P(y \in B|x)] = [P(y \in B|x, z = 1)P(z = 1|x), P(y \in B|x, z = 1)P(z = 1|x) + P(z = 0|x)]$$

Observe that the width of the identification interval is given by $U_B - L_B = P(z = 0|x)$; hence, data is informative unless y is always missing. Note that the width of interval may vary with x but it does not vary with the set B.

Example (Bounding Probability of Exiting Homelessness)**.** Piliavin & Sosin (1988) wanted to know the probability of homelessness six months after already being homeless. Let $y = 1$ if the individual has a home six months later and $y = 0$ denote the opposite; x are background attributes. The goal is learn $P(y = 1|x)$ and the missing data problem arises due to not being able to locate part of the original sample six months later. Let $x = sex$. 106 men were sampled originally and 64 of these men were found six months later, 21 of which were no longer homeless. Therefore, the empirical probability estimate of $P(y = 1|male, z = 1) = \frac{21}{64}$ and that of $P(z = 1|male) = \frac{64}{106}$. So, the estimate of the bound on $P(y = 1|male)$ is $[\frac{21}{106}, \frac{63}{106}] \approx [0.20, 0.59]$. 31 women were sampled originally and 14 of these women were found six months later, 3 of which were no longer homeless. Therefore, the empirical probability estimate of $P(y = 1|female, z = 1) = \frac{3}{14}$ and that of $P(z = 1|female) = \frac{14}{31}$. So, the estimate of the bound on $P(y = 1|female)$ is $[\frac{3}{31}, \frac{20}{31}] \approx [0.10, 0.65]$.

The following is a corollary of LIE, which states that $E_X(E(Y|X)) = E(Y)$. The corollary holds since $E(E(h(Y, X)|X)) = E(h(Y, X))$.

Corollary.

$$E[g(y)|x] \overset{LIE}{=} E[g(y)|x, z = 1]P(z = 1|x) + E[g(y)|x, z = 0]P(z = 0|x)$$

Assume g is bounded and let

$$\begin{aligned} g_0 &= \inf_{y \in Supp(Y)} g(y) \\ g_1 &= \sup_{y \in Supp(Y)} g(y) \end{aligned}$$

$g(y)$ will be bounded as long as $P(z = 0|x) > 0$. If $g_1 = \infty$ or $g_0 = -\infty$, i.e. the case of unboundedness, then we need more assumptions for inference.

The width of the interval will be $(g_1 - g_0)P(z = 0|x)$ and

$$\begin{aligned} H[E[g(y)|x]] = [&E(g(y)|x, z = 1)P(z = 1|x) + g_0P(z = 0|Zx), \\ &E(g(y)|x, z = 1)P(z = 1|x) + g_1P(z = 0|x)] \end{aligned} \tag{4}$$

which is a proper subset of $[g_0, g_1]$ when $P(z = 0|x) < 1$, so it is informative. The severity of missing data is directly proportional to $P(z = 0|x)$.

Example. As an application of (4), let $B \subset Y$ and $g(y) = 1[y \in B]$. Then $g_0 = 0, g_1 = 1, E[g(y)|x] = P(y \in B|x)$ and $E[g(y)|x, z = 1] = P(y \in B|x, z = 1)$. So, (4) is an alternative version of the bound given by (3) on$P(y \in B|x)$.

Remark. When $g(\cdot)$ is unbounded from above or below, (4) still holds but has different implications when $P(z = 0|x) > 0$. The lower bound on $E[g(y)|x]$ is $-\infty$ if $g_0 = -\infty$ and ∞ if $g_1 = \infty$. The identification region has infinite width but is still informative if $g(\cdot)$ is bounded from at least one side. As Manski put it, ‘the presence of missing data makes credible assumptions a prerequisite for inference on the mean of an unbounded random variable.’ (2007: 44)

Missing at random

To tighten our bounds, we may invoke different distributional assumptions about the nature of our uncertainty regarding missing data. The most common though generally implausible assumption is the *missing at random* or *conditional statistical independence* assumption.

Definition. The *missing at random* (MAR) or *conditional statistical independence* assumption is:

$$\begin{aligned} &P(y|x, z = 1) = P(y|x, z = 0) = P(y|x) \\ \Longrightarrow\ &E[y|x, z = 1] = E[y|x, z = 0] = E[y|x] \end{aligned}$$

The missing at random assumption is an example of a *non refutable* assumption – any assumption about $P(y|x, z = 0)$ is not refutable.

Definition. 'Any assumption that directly restricts the distribution $P(y|x, z = 0)$ of missing data is *non-refutable.*' (Manski, 2007: 46; my emphasis)

To see how refutable assumptions may be tested statistically, consider the assumption $E[g(y)|x] \in R_1 \subset \mathbb{R}$. We can reject the this hypothesis if $H_N\{E[g(y)|x)]\}$ is sufficiently far from R_1.

MAR implies that the following identification region contains one element, $P(y|x, z = 1)$:

$$H_0[P(y|x)] \equiv [P(y|x, z = 1)P(z = 1|x) + \gamma P(z = 0|x), \gamma \in \Gamma_{0Y}]$$

Assumptions placed on $P(y|x, z = 0) \in \Gamma_{0Y}$ are not empirically testable (they are nonrefutable though you may be able to argue why it is not plausible for $P(y|x, z = 0)$ to lie in Γ_Y) versus $P(y|x) \in \Gamma_Y$ is better since it can be empirically tested. Interesting assumptions include those positing that $P(y|x)$ lies in a specific set of distributions, Γ_{1Y}. Data alone imply that $P(y|x) \subset H[P(y|x)]$. Combining the data with the assumption $P(y|x) \subset \Gamma_{1Y}$:

$$H_1[P(y|x)] \equiv H[P(y|x)] \cap \Gamma_{1Y}$$

If $\cap\Gamma_{1Y}$ is zero, then $P(y|x)$ cannot lie in Γ_{1Y} and so we have made the wrong assumption, but if $\cap\Gamma_{1Y}$ is nonzero, this does not imply that we accept but rather that we do not reject. Data may be refuted or may 'not be refuted'. When we use the term 'nonrefutable', we ask whether there could be an ex-ante probability that the intersection is the null set – if yes, then our assumption is refutable, else it is nonrefutable. Refutability concerns logic and is a property of assumptions and data whereas and credibility is a property of assumptions and the researcher, so there is an element of subjectivity involved with credibility.

With no assumption on $P(y|x)$, $P(y|x) \in H[P(y|x)]$. Now assume that $P(y|x) \in \Gamma_Y$. After making this assumption, $P(y|x) \in H[P(y|x)] \cap \Gamma_Y = H_1[P(y|x)]$. If $H_1[P(y|x)]$ is a strict subset of $H[P(y|x)]$, then the assumption is said to have *identifying power.* If $H_1[P(y|x)] = 0$, then the assumption is refutable. If $H_1[P(y|x)] \neq 0$, then the assumption is non refutable; however, this does not mean that the assumption is true!

Treatment response

Analysis of treatment response is an interesting problem of prediction with missing outcomes: predict outcomes that would occur if alternative treatment rules were applied to a population. We can at most observe the realised outcomes, i.e. outcomes experienced under received treatments, but not the counterfactual outcomes, i.e. outcomes that would be experienced under other treatments. For example, if a group are ill and there are two treatments, *viz.* drugs or surgery where the outcome of interest is life span, then we may wish to predict the life spans that might occur should all patients of a certain type be treated by drugs. However, the only available data on realised life spans would involve some patients that were treated by drugs and the rest by surgery. Another example relates to economic policy where workers displaced from a plant closure were either retrained or assisted in job search where the outcome of interest might be income. We may wish to learn the incomes that might occur if all workers with particular backgrounds were retrained and then compare these incomes with those that would occur if the same workers were given assistance in job search instead. However, available data on realised incomes will most likely involve a subgroup of workers that were retrained and another group, each of who were given job assistance.

More formally, let T be the set of all feasible treatments and each member of the study population possess covariates $x_j \in X$ and a *response function* $y_j(\cdot) : T \longrightarrow Y$ mapping mutually exclusive and exhaustive treatments $t \in T$ into outcomes $y_j(t) \in Y$. Therefore, $y_j(t)$ is the outcome person j would experience if s/he were to receive treatment t.[1] The subscript j on $y_j(\cdot)$ allows treatment response to be heterogeneous across members of the population, i.e. they need not respond to treatment in the same way. Also, treatment response is individualistic, i.e. the outcome that person j experiences is independent of treatments other people receive, hence the notation $y_j(\cdot)$. Let $z_j \in T$ be person j's received treatment, so $y_{\equiv} y_j(z_j)$ is the realised outcome, while counterfactual outcomes are denoted by $[y_j(t), t \neq z_j]$. Observation may reveal $P(y, z|x)$ of realised outcomes and treatments for people with covariates x, while the distribution of outcomes that would occur if all people with covariate x received treatment t is denoted by $P[y(t)|x]$; so, to predict outcomes under a policy of treatment t for people with covariates x, we must infer $P[y(t)|x]$.

Definition. The *selection problem* refers to the problem of identification of outcome $P[y(t)|x]$ given a knowledge of $P(y, z|x)$.

With treatment response, where $z = t$ is treatment t, from data alone, the problem can be written as:

$$\begin{aligned} P[y(t)|x] &\overset{\text{LTP}}{=} P[y(t)|x, z = t]P(z = t|x) \\ &\quad + \underbrace{P[y(t)|x, z \neq t]}_{\text{unobserved}} P[z \neq t|x] \\ &= P(y|x, z = t)P(z = t|x) \\ &\quad + P[y(t)|x, z \neq t]P(z \neq t|x) \end{aligned}$$

where $P[y(t)|x, z \neq t]$ corresponds to missing outcomes and the remaining quantities after the second equality are known. The identification region using empirical evidence alone is given by

$$H[P(y(t)|x)] = [P(y|x, z = t)P(z = t|x) + \gamma P(z \neq t|x); \gamma \in \Gamma_Y]$$

Note that $P[y(t)|x] = P(y|x, z = t)$ if treatment selection is random, but generally differ otherwise. The primer is the distribution of outcomes that would occur if everyone with covariates x received treatment t, while the latter is the distribution of outcomes that occur for people who have covariates x and actually receive treatment t.

To learn about policies mandating different treatments for people with covariates x, we would like to know about $\{P[y(t)|x], t \in T\}$. The identification region using only data is

$$H\{P[y(t)|x], t \in T\} = \times_{t \in T} H\{P[y(t)|x]\}$$

[1] $y_j(t)$ is also called a *potential, latent* or *conjectural* outcome.

We learn more about $P[y(t)|x]$ but less about $P[y(t')|x], t' \neq t$, the more often that treatment t is selected in the study population. Furthermore, data alone cannot answer the question as to whether outcomes vary with treatment since counterfactuals are unobservable and observation of realised treatments and outcomes is uninformative regarding all the counterfactual outcome distributions $\{P[y(t)|x, z \neq t], t \in T\}$. It may be possible that in fact for each person j, y_j, the person's realised outcome is the same as the potential outcome under any treatment $y_j(t), t \in T$. Therefore, data alone cannot refute the hypothesis that $P[y(t)|x], t \in T$ are all the same, i.e. that hypothesis is nonrefutable.

Definition. Focus on two treatments t and t'. The *average treatment effect* (ATE) is

$$E[y(t)|x] - E[y(t')|x]$$

Remark. Note that the hypothesis $ATE = 0$ is non refutable with data alone since counterfactual observations are missing it is possible that $y_j(t) = y_j(t')\ \forall j$. This hypothesis only becomes refutable if we combine the data with sufficiently strong distributional assumptions, e.g. *randomisation of treatment*:

$$P(y(t)|x, z = t) = P(y(t)|x, z \neq t)$$

which gives point identification. The distributional assumption of statistical independence is the selection problem analog of MAR, i.e.

$$P(y(t)|x) = P(y|x, z = t) = P(y(t)|x, z \neq t)$$

is almost only credible in classical randomized experiments. Equivalently

$$z = t \perp\!\!\!\perp y|x$$

When treatment selection is random, we have point identification of $P(y(t)|x)$ as mentioned before. From LIE and the observability of realised outcomes:

$$\begin{aligned} E[y(t)|x] - E[y(t')|x]& \\ &= E(y|x, z = t)P(z = t|x) \\ &\quad + \underbrace{E[y(t)|x, z \neq t]}_{\in[y_0,y_1])} P(z \neq t|x) \\ &\quad - E(y|x, z = t')P(z = t'|x) \\ &\quad - \underbrace{E[y(t')|x, z = t']}_{\in[y_0,y_1]} P(z \neq t'|x) \end{aligned}$$

So the identification region for ATE is

$$\begin{aligned} H\{E[y(t)|x] - E[y(t')|x]\}& \\ &= \big[E(y|x, z = t)P(z = t|x) + y_0 P(z \neq t|x) \\ &\quad - E(y|x, z = t')P(z = t'|x) - y_1 P(z \neq t'|x), \\ &\quad E(y|x, z = t)P(z = t|x) + y_1 P(z \neq t|x) \\ &\quad - E(y|x, z = t')P(z = t'|x) - y_0 P(z \neq t'|x)\big] \end{aligned}$$

which necessarily contains zero and its width is given by

$$(y_1 - y_0)[P(z \neq t|x) + P(z \neq t'|x)] = (y_1 - y_0)[2 - P(z = t|x) - P(z = t'|x)]$$

i.e. the width of the interval depends on the fraction of the population under study that receive treatments t and t'. Since the sum of the fraction is one, the width of the interval is between $(y_1 - y_0)$ and $2(y_1 - y_0)$. When t and t' are the only feasible treatments, the width is $(y_1 - y_0)$. When there is no data, ATE lies in $[y_0 - y_1, y_1 - y_0]$ and the width is $2(y_1 - y_0)$. Therefore, data alone restricts the ATE to half its logically possible range.

Mock exam questions

Exercise. The following exercise has been taken from Manski (2011). Suppose a researcher observes the scores of the population of foreign PhD students who take and the composition part of the TOEFL examination. For each student, the researcher observes the following:

$$y = \text{the score (passing scores are 4,5,6 say)}$$

$$x = \begin{cases} 1 & \text{if the student has a mathematics or science bachelor's degree} \\ 0 & \text{otherwise} \end{cases}$$

The population distribution $P(y, x)$ is shown in table 1:

	Test score			
Degree	$y = 4$	$y = 5$	$y = 6$	Totals
$x = 0$	0.20	0.40	0.15	0.75
$x = 1$	0.05	0.10	0.10	0.25
Total	0.25	0.50	0.25	1.00

Table 1: Distribution of foreign PhD students' marks for composition part of TOEFL

1. Find a best predictor of y given $x = 0$ under absolute loss.
2. Find a best predictor of y given $x = 1$ under square loss.
3. Find a best predictor of x given $y = (4 \text{ or } 5)$, under square loss.
4. Find a best predictor of x given $y = 6$, under absolute loss.
5. Suppose that the researcher observes $P(y = 6|x = 0) = 0.2$ and $P(y = 6|x = 1) = 0.4$ and then states that

   > The data appears to suggest that obtaining a mathematics or science degree substantially increases the chance that a student obtains the highest examination score. The estimated effect of a mathematics or science degree is to increase the probability of scoring 6 from 0.2 to 0.4.

   Is this statement accurate in describing the empirical results? Explain.

Solution (Conditional Prediction)**.**

1. The best predictor of Y conditional on X under *absolute loss* is the *median* of Y conditional on X, where the median is defined to be

   $$M(Y|X = x) = \min_t : P(Y \leq t|X = x) \geq \frac{1}{2} \tag{5}$$

   We need to find the probability density function (pdf) of $(Y|X = 0)$ to calculate the cumulative density function (cdf) of $(Y|X = 0)$ so that we can find $M(Y|X = 0)$ as defined by (5). The pdf of $(Y|X = 0)$ is given by

   $$P(Y = 4|X = 0) = \frac{P(Y = 4, X = 0)}{P(X = 0)} = \frac{0.20}{0.75} = 0.27$$
   $$P(Y = 5|X = 0) = \frac{P(Y = 5, X = 0)}{P(X = 0)} = \frac{0.40}{0.75} = 0.53$$
   $$P(Y = 6|X = 0) = \frac{P(Y = 6, X = 0)}{P(X = 0)} = \frac{0.15}{0.75} = 0.20$$

So, the cdf of $(Y|X=0)$ is given by

$$\begin{aligned}P(Y\leq 4|X=0)&=0.27\\P(Y\leq 5|X=0)&=0.80\\P(Y\leq 6|X=0)&=1.00\end{aligned}\tag{6}$$

From the cdf, we can see that $\underline{M(Y|X=0)=5}$, which is the best predictor under absolute loss.

2. The best predictor of $(Y|X)$ under *square loss* is the *mean*, $E(Y|X)$. To calculate the best predictor of $(Y|X=1)$ under square loss, i.e. to calculate $E(Y|X=1)=\sum_y y\cdot P(Y|X=1)$, we need to find the pdf of $(Y|X=1)$. The pdf of $(Y|X=1)$ is given by

$$\begin{aligned}P(Y=4|X=1)&=\frac{P(Y=4|X=1)}{P(X=1)}=\frac{0.05}{0.25}=0.2\\P(Y=5|X=1)&=\frac{P(Y=5|X=1)}{P(X=1)}=\frac{0.10}{0.25}=0.4\\P(Y=6|X=1)&=\frac{P(Y=6|X=1)}{P(X=1)}=\frac{0.10}{0.25}=0.4\end{aligned}$$

Now to calculate the mean

$$\begin{aligned}E(Y|X=1)&=\sum_y y\cdot P(Y|X=1)\\&=4\times 0.2+5\times 0.4+6\times 0.4\\&=5.2\end{aligned}\tag{7}$$

Therefore, the best predictor of $(Y|X=1)$ under square loss is $\underline{E(Y|X=1)=5.2}$.

3. The pdf of $(X|Y=4\vee Y=5)$ is given by

$$\begin{aligned}P(X=0|Y=4\vee Y=5)&=\frac{P(X=0,Y=4\vee Y=5)}{P(Y=4\vee Y=5)}\\&=\frac{0.60}{0.75}=0.8\\P(X=1|Y=4\vee Y=5)&=\frac{P(X=1,Y=4\vee Y=5)}{P(Y=4\vee Y=5)}\\&=\frac{0.15}{0.75}=0.2\end{aligned}$$

The best predictor of $(X|Y=4\vee Y=5)$ under square loss is the mean, which is given by

$$\begin{aligned}E[X|Y=4\vee Y=5]&=\sum_x x\cdot P(X|Y=4\vee Y=5)\\&=0\times 0.8+1\times 0.2\\&=0.2\end{aligned}$$

4. The pdf of $(X|Y=6)$ is given by

$$\begin{aligned}P(X=0|Y=6)&=\frac{P(X=0,Y=6)}{P(Y=6)}=\frac{0.15}{0.25}=0.6\\P(X=1|Y=6)&=\frac{P(X=1,Y=6)}{P(Y=6)}=\frac{0.10}{0.25}=0.4\end{aligned}$$

So, the cdf of $(X|Y = 6)$ is given by

$$\begin{aligned} P(X \leq 0|y = 6) = 0.6 \\ P(X \leq 1|Y = 6) = 1 \end{aligned} \tag{8}$$

The best predictor of $(X|Y = 6)$, under absolute loss is the median, which is given by

$$M(X|Y = 6) = 0$$

5. No, this statement does not accurately describe the empirical finding.
   We can only say that the mathematics and science students on average scored higher than those who did not have such a background. We cannot say that the very fact that they studying mathematics or science increased the probability of a student scoring a 6.
   Asking what would happen to this $E(Y|X)$ when we vary X is akin to a hypothetical change in X, where we have no data and so the researcher has confused correlation with causation and has used a counterfactual (expressing what has not happened but what might or would happen if circumstances, i.e. data, were different). The researcher is in effect extrapolating using the assumption of external validity, which is undermined by the fact that we are only looking at *foreign PhD students* and have no data on the rest of the population of students at large.
   However, if the students were randomly assigned bachelor's degrees in maths / science (an impossibility in this case, but possible in more general cases where X could include randomly distributing answers to the test to some of the students), then the researcher would be correct in saying that an increase in that covariate (e.g. having the answers to the test prior to the test) increases the probability that a student will do better on average than a student who does not have answers to the test. But since the distribution of maths and science degrees is non-random and we are dealing with what actually happened (descriptive) we cannot say that having a maths or science degree increases the probability that a student scores a 6.

Exercise. Let us imagine that for some reason, a fraction of the class' results disappear including all traces of assessment throughout the year and that – again hypothetically – college authorities dictate that we must *impute* a value for these students. Suppose from a class of 120, 6 students have missing results. Let y denote result in percentage terms where students need at least 40 to pass and z denote whether we observe the result or not. The distribution of marks for students we have data on is given in table 2. Take scores in fives for each interval students score in (so $5, 15, 25, \ldots$).

Score	Frequency
0-10	3
11-20	17
21-30	2
31-40	10
41-50	18
51-60	10
61-70	15
71-80	21
81-90	15
91-100	3

Table 2: Distribution of marks.

1. What bounds does the probability that a student taken at random from the class passes lie between?

2. What if we assume that the data was missing at random? Is the bound tighter? Why or why not? Is it refutable?
3. What bounds can we place on the average mark in the entire population?
4. Compare your answer from part 3 with an imputation rule that students whose results are missing are given the average mark of the observed results.

Solution (Incomplete Data)**.**
Given that we are taking marks in 'fives', let us write the revised distribution in table 3. There are $N = 120$ students in the class, of whom we have no data on six, so $P(z = 0) = \frac{6}{120} = 0.05$ is the fraction of missing data; remember $P_N(z = 1) = \frac{1}{N}\sum_{i=1}^{N} 1[z_i = 1]$.

Score	Frequency
5	3
15	17
25	2
35	10
45	18
55	10
65	15
75	21
85	15
95	3

Table 3: Revised distribution of marks.

1. To pass, students must get at least 40. Letting B denote the set of all such marks from the revised distribution that corresponding to the students passing, to pass y must be in B:

$$y \in \{45, 55, 65, 75, 85, 95\} \equiv B$$

We want $P(y \in B)$ and can express this using the law of total probability (LTP) as

$$P(y \in B) \overset{\text{LTP}}{=} P(y \in B|z = 1)P(z = 1) + P(y \in B|z = 0)P(z = 0) \tag{9}$$

We know

$$P(z = 0) = 0.05 \implies P(z = 1) = 1 - P(z = 0) = 1 - 0.05 = 0.95$$

and while $P(y \in B|z = 0)$ is the only unknown quantity in (9), because it is a probability, $P(y \in B|z = 0) \in [0, 1]$. We need to calculate $P(y \in B|z = 1)$.

$$\begin{aligned} P_N(y \in B|z = 1) &= \frac{\sum_{i=1}^{N} 1[y_i \in B, z_i = 1]}{\sum_{i=1}^{N} 1[z_i = 1]} \\ &= \frac{18 + 10 + 15 + 21 + 15 + 3}{120 - 6} \\ &= \frac{82}{114} \\ \therefore P(y \in B) &= \frac{92}{114} \times 0.95 + [0, 1] \times 0.05 \\ &\in \left[\frac{41}{60}, \frac{11}{15}\right] \\ &\equiv H[P(y \in B)] \end{aligned}$$

which is our identification region for the probability that a student passes.

2. The assumption of missingness at random (MAR) is

$$P(y|z=1) = P(y|z=0)$$

Observation. Observe that under MAR

$$P(y|z=1) = P(y|z=0) = P(y)$$

To see this, use the law of total probability to expand $P(y)$:

$$\begin{aligned} P(y) &\overset{\text{LTP}}{=} P(y|z=1)P(z=1) + P(y|z=0)P(z=0) \\ &\overset{\text{MAR}}{=} P(y|z=1)[\underbrace{P(z=1)+P(z=0)}_{1}] \\ &= P(y|z=1) \\ &\overset{\text{MAR}}{=} P(y|z=0) \end{aligned}$$

Note that here MAR implies that

$$P(y \in B|z=1) = P(y \in B|z=0)$$

$$\begin{aligned} \therefore P(y \in B) &\overset{\text{LTP}}{=} P(y \in B|z=1)P(z=1) + P(y \in B|z=0)P(z=0) \\ &\overset{\text{MAR}}{=} P(y \in B|z=1) \\ &\overset{1}{=} \frac{82}{114} \end{aligned}$$

Therefore, MAR *point identifies* $P(y \in B)$.
Certainly, the bound is tighter:

$$H_1[P(y \in B)] = \left[\frac{41}{60}, \frac{11}{15}\right] \ni \frac{82}{114} = H_1[P(y \in b)]$$

It is tighter because we impose a strong, *nonrefutable* assumption on the distribution of unknown, missing data; and it is *nonrefutable* because the assumption directly restricts the distribution $P(y \in B|z=0)$ of missing data.

3. We want $E(y)$, so using the law of iterated expectations to expand $E(y)$, we get that

$$E(y) \overset{\text{LIE}}{=} E(y|z=1)P(z=1) + E(y|z=0)P(z=0)$$

We know

$$P(z=1) = 0.95 \quad P(z=0) = 0.05$$

and while $E(y|z=0)$ is unknown, marks must lie within $[0, 100]$. Actually with the assumption of 'fives', we know more:

$$5 \leq E(y|z=0) \leq 95$$

Going even further, we can write this out fully:

$$E(y|z=0) \in \{5, 15, 25, 35, 45, 55, 65, 75, 85, 95\}$$

We need to calculate $E(y|z = 1)$ and can work this out from the revised distribution in table 3. Summing over observed i where I denotes the number of observations:

$$\begin{aligned}
E(y|z=1) &= \frac{1}{I}\sum_i \text{score}_i \times \text{frequency}_i \\
&= \frac{(5)(3)+(15)(17)+(25)(2)+(35)(10)+(45)(18)+(55)(10)+(65)(15)+(75)(21)+(85)(15)+(95)(3}{114} \\
&= \frac{6140}{114} \\
\therefore E(y) &= \frac{6140}{114}(0.95)+[5,95](0.05) \\
&\in \left[\frac{617}{12}, \frac{671}{12}\right] \\
& [51.41\dot{6}, 55.91\dot{6}] \\
&\equiv H[E(y)]
\end{aligned}$$

which is the identification region for the average mark in the class.

4. From part 3

$$\begin{aligned}
E(y|z=1) &\overset{3}{=} \frac{6140}{114} \\
&= E(y|z=0)
\end{aligned}$$

where the second equality follows by the imputation rule in this part of the question does. So now

$$E(y) = \frac{6140}{114}$$

and we get *point* identification rather than *partial* identification. The imputation rule we used utilises a weaker assumption than MAR; it restricts only means, rather than the entire probability distribution of missing data.

Exercise. Modified from Manski (2011). Suppose a researcher wants to use a scale to measure the weight of each student in the population of Trinity undergraduates. Let y denote a student's true weight and let $y*$ be weight as measured on the scale. Let $x = 1$ if a person is vegetarian and $x = 0$ otherwise. Consider the following two assumptions:

A1 The scale is accurate when it is inside the range [100, 200] pounds, but is inaccurate outside this range.

$$\therefore y = \begin{cases} y^* & \text{if } y^* \in [100, 200] \\ y < 100 & \text{if } y^* < 100 \\ y > 200 & \text{if } y^* > 200 \end{cases}$$

A2 The true weight in the population is never below 70 pounds and never exceeds 300 pounds.

Having weighed each member of the population, the researcher reports her findings.

$$\begin{aligned}
P(y^* < 100|x=0)&=0.05 \qquad & E(y^*|x=0)&=180 \\
M(y^*|x=0)&=160 \qquad & P(y^* > 200|x=0)&=0.10 \\
P(y^* < 100|x=1)&=0.10 \qquad & E(y^*|x=1)&=160 \\
M(y^*|x=1)&=150 \qquad & P(y^* > 200|x=1)&=0
\end{aligned}$$

1. What can you conclude about $E(y|x=0)$?
2. What can you conclude about $M(y|x=0)$?
3. Suppose that a nutritionist at Trinity attempt to use these findings to support a proposal that restaurants in Trinity henceforth should serve entirely vegetarian meals. The nutritionist states:

   > The data suggest that being a vegetarian significantly reduces the chance that a student is overweight. The effect of a vegetarian diet is to decrease the probability of weighing more than 200 pounds from 0.10 to 0.

   Is this statement an accurate description of the empirical findings? Explain.

Solution (Incomplete Data)**.**

1. We want $E(y|x=0)$ and can use the law of iterated expectations (LIE) to expand it as follows:

$$\begin{aligned} E(y|x=0) &\overset{\text{LIE}}{=} E(y|y^*<100, x=0)P(y^*<100|x=0) \\ &+ E(y|y^*\in[100,200], x=0)P(y^*\in[100,200]|x=0) \\ &+ E(y|y^*>200, x=0)P(y^*>200|x=0) \end{aligned}$$

We know

$$\begin{aligned} &\left.\begin{aligned} P(y^*<100|x=0) &= 0.05 \\ P(y^*>200|x=0) &= 0.1 \end{aligned}\right\} \\ &\implies P(y^*\in[100,200]|x=0) = 1-(0.05+.01) \\ &\qquad\qquad\qquad\qquad\qquad\quad = 0.85 \\ &E(y|y^*<100, x=0) \in [70,100) \\ &E(y|y^*>200, x=0) \in (200,300] \end{aligned}$$

and we need to calculate $E(y|y^*\in[100,200], x=0)$. Observe that

$$E(y|y^*\in[100,200], x=0) = E(y^*|y^*\in[100,200], x=0) \tag{10}$$

By the law of iterated expectations

$$\begin{aligned} \underbrace{E(y^*|x=0)}_{180} &\overset{\text{LIE}}{=} \overbrace{E(y^*|y^*<100, x=0)}^{\in[70,100)}\underbrace{P(y^*<100, x=0)}_{0.05} \\ &+ \underbrace{E(y^*|y^*\in[100,200], x=0)}_{\text{want}}\underbrace{P(y^*\in[100,200], x=0)}_{0.85} \\ &+ \underbrace{E(y^*|y^*>200, x=0)}_{\in(200,300]}\underbrace{P(y^*>200, x=0)}_{0.1} \end{aligned}$$

$$\begin{aligned} &\therefore E(y^*|y^*\in[100,200], x=0) \\ &= \frac{1}{0.85}\{180-[70,100)\times 0.05-(200,300]\times 0.1\} \end{aligned}$$

and combining this with (10)

$$E(y|y^*\in[100,200], x=0) \in \left[\frac{2900}{17}, \frac{3130}{17}\right]$$

$$
\begin{aligned}
\therefore E(y|x=0) &= [70,100) \times 0.05 + \left[\frac{2900}{17}, \frac{3130}{17}\right] \times 0.85 + (200,300] \times 0.1 \\
\therefore H[E(y|x=0)] &= \left(\frac{337}{2}, \frac{383}{2}\right) \\
&= [168.5, 191.5]
\end{aligned}
$$

2. We now want the 'middle value', $M(y|x=0)$.

   Claim 1. $M(y|x=0) = M(y*|x=0) = 160$

   Proof. The definition of the median is

   $$M(y) = \inf\{t : P(y \leq t) \geq \frac{1}{2}\}$$

   First observe that from the definition of y:

   $$
   \begin{aligned}
   P(y^* < 100|x=0) &= P(y < 100|x=0) = 0.05 \\
   P(y^* > 200|x=0) &= P(y > 200|x=0) = 0.1 \\
   P(y^* \in [100,200]|x=0) &= P(y \in [100,200]|x=0) = 0.85
   \end{aligned}
   $$

   $$
   \begin{aligned}
   \therefore P(y < 100|x=0) &= 0.05 \\
   P(y \leq 200|x=0) &= 0.05 + 0.85 = 0.9 \\
   P(y \leq 300|x=0) &= 1 \\
   \implies M(y|x=0) &\in [100,200]
   \end{aligned}
   $$

   Finally note that y^* is accurate for y inside $[100,200]$, i.e. $y = y^*$ for each corresponding y and y^* in $[100,200]$. So there are the same number of elements for $y < 100$ as for $y^* < 100$ and for $y \in [100,200]$ as for $y^* \in [100,200]$ and moreover the elements y are the same as the corresponding y^* inside $[100,200]$. So the 'middle' value for y^* will be the same as the middle value for y and so the medians are the same, i.e.

   $$M(y^*|x=0) = M(y|x=0) = 160$$ □

3. No, this statement does not accurately describe the empirical finding.
   We can only say that vegetarians had on average a lower probability of weighing more than 200 pounds. We cannot say that their diet reduced the probability of weighing more than 200 pounds.
   The nutritionist has confused correlation with causation and has used a counterfactual (expressing what has not happened but what might or would happen if circumstances, i.e. data were different) in suggesting that Trinity should serve entirely vegetarian meals. Perhaps there are other omitted variables, such as exercise, healthier lifestyles, genetic advantages, luck, etc. at play. The nutritionist is in effect extrapolating using the assumption of external validity, which is undermined by the fact that we are only looking at the population of current Trinity undergraduate students. Will these results hold for different cohorts? Will non-vegetarians become healthier by eating vegetarian food while in Trinity? Do students frequent Trinity's restaurants often?
   However, if the Trinity undergraduates were randomly assigned to being vegetarians or non-vegetarians (an impossibility in this case, but possible in cases where we could randomly assign diets to individuals in a cohort and effectively monitor participation), then the nutritionist would be correct in saying that the effect of a vegetarian diet is to lower the probability of weighing over 200 pounds.

Exercise. Suppose that we have data on three mutually exclusive and exhaustive treatments $t = a$, $t = b$ and $t = c$. The outcome of interest is denoted by $y(t)$ where we emphasise the dependence of the response y on the treatment t. Assume $y \in [0, 1]$. Abstract from any issues of finite sample inference. We have the following population distribution:

$$\begin{array}{ccc} P(t=a){=}0.1 & P(t=b){=}0.6 & P(t=c){=}0.3 \\ E(y|t=a){=}0.5 & E(y|t=b){=}0.4 & E(y|t=c){=}0.2 \end{array}$$

1. What is the average treatment effect $E(y(a)) - E(y(b))$?
2. What is the average treatment effect $E(y(b)) - E(y(c))$?
3. What is the average treatment effect $E(y(c)) - E(y(a))$?
4. Calculate the widths of the intervals for each of the previous parts. Comment on the relative magnitudes of these widths.

Solution (Treatment Response)**.**
To answer this question, let us first consider the individual identification regions for $E(y(a))$, $E(y(b))$ and $E(y(c))$. In general by the law of iterated expectations (LIE):

$$\begin{aligned} E(y(t')) &\overset{\text{LIE}}{=} E(y(t')|t = t')P(t = t') + E(y(t')|t \neq t')P(t \neq t') \\ &= E(y|t = t')P(t = t') + E(y(t')|t \neq t')P(t \neq t') \end{aligned}$$

We want $E(y(a))$:

$$E(y(a)) = E(y|t = a)P(t = a) + E(y(a)|t \neq a)P(t \neq a)$$

We know

$$P(t = a) = 0.1 \Longrightarrow P(t \neq a) = 1 - P(t = a) = 1 - 0.1 = 0.9$$

The quantity $E(y(a)|t \neq a)$ is unknown, but since $y \in [0, 1]$, we have that

$$E(y(a)|t \neq a) \in [0, 1]$$

We also know that

$$E(y|t = a) = 0.5$$

$$\begin{aligned} \therefore E(y(a)) &= (0.5)(0.1) + [0, 1](0.9) \\ \therefore H[E(y(a))] &= [0.05, 0.95] \end{aligned}$$

Similarly for $E(y(b))$:

$$E(y(b)) = E(y|t = b)P(t = b) + E(y(b)|t \neq b)P(t \neq b)$$

We know that

$$P(t = b) = 0.6 \Longrightarrow P(t \neq b) = 1 - P(t = b) = 1 - 0.6 = 0.4$$

As before, $E(y(b)|t \neq b) \in [0, 1]$ and we know that

$$E(y|t = b) = 0.4$$

$$\begin{aligned} E(y(b)) &= (0.4)(0.6) + [0, 1](0.4) \\ &\in [0.24, 0.64] \end{aligned}$$

Note that $H[y(b)]$ is tighter than $H[y(a)]$ because more of the population receives treatment b; this is related to the severity of missing data. Once more for $E(y(c))$ where we know

$$\begin{aligned} P(t = c) &= 0.3 \\ P(t \neq c) &= 0.7 \\ E(y|t = c) &= 0.2 \end{aligned}$$

$$\begin{aligned} \therefore E(y(c)) &= (0.2)(0.3) + [0, 1](0.7) \\ &\in [0.06, 0.76] \end{aligned}$$

In what follows, denote LB_t and UB_t as the lower bound and upper bound, respectively of $E(y(t))$, $t \in \{a, b, c\}$. Note that

$$H[E(y(t)) - E(y(t'))] = [LB_t - UB'_t, UB_t - LB'_t]$$

1. $H[E[y(a)] - E[y(b)]] = [-0.59, 0.71]$

2. $H[E[y(b)] - E[y(c)]] = [-0.52, 0.58]$

3. $H[E[y(c)] - E[y(a)]] = [-0.89, 0.71]$

4. The widths of the intervals are 1.3, 1.1 and 1.6, respectively. The width of each interval $H[y(t)]$ depends on the fraction of the study population not receiving treatment t, i.e. unknown relative to t, so it will be wider the lower the fraction of people receiving treatment t is. The width of each interval $H[E[y(t)] - E[y(t')]]$ depends on the fraction of the study population receiving treatments t and t'. Since the largest fraction of the population are covered by treatments b and c than by any other combination of two treatments, their average treatment effect will have the smallest interval – the fraction of the population not receiving either of these two treatments is the smallest $P(t = a) = 0.1$. Treatments a and c together cover only 40% of the population so their average treatment effect will have the largest interval – the fraction of the population not receiving either of these two treatments is the largest $P(t = b) = 0.6$. In between these two extremes is the combination of treatments a and b.

   Remark. Note that the hypothesis of zero average treatment effects is nonrefutable, unless we combine this with strong distributional assumptions. We can see that zero is contained in each of the three intervals for the average treatment effects.