Sheet 2

~~1. On the basis of this research, we can ~~conclude~~ know that there is notable an association in the sample~~

1. This research tells us that there's an association in the sample between frequent chocolate consumption and lower weight. If the sample was representative, we can also infer with some degree of confidence that such an association exists in the UK population.

However, we cannot interpret this result causally without some more work. There is no particular reason to think we have zero/minimal selection bias here, because very plausibly there's some confounder (e.g. wealth) which is causally linked to higher chocolate consumption and lower weight (i.e. $E[uX] \neq 0$). So, it's not ~~very~~ justified to say that ~~any~~ any particular person's weight (or the average person's) would decrease by eating more chocolate. (Indeed, it seems highly improbable that this is the causal relationship.)

3. $bw = 3395 - 15\,cig + \hat{u}$

a) $\hat{bw}(0) = 3395$ (what notation would you normally use here?) what you wrote is great. i.e. $\hat{bw}(\cdot)$

$\hat{bw}(20) = 3095$

So, on average, an increase in cigarettes smoked of 1 pack/day is associated with a ~~reducing~~ reduction of birthweight by 300g. ✓ (is this suitably non-causal-sounding?) Yes

b) No, because OR is unlikely to hold. ✓

In particular, I would expect that ~~negative assoc~~ $Cov(u, cig) < 0$, ✓ because mothers who smoke more are likely to be more deprived, which lowers birthweight.

c) ✓ Solving for cig, we get -7. But clearly this is impossible as cig is non-negative.

- Note that, in general, if you reverse the direction of the regression, the coefficients won't merely be inverses. So if you wanted to predict cig from bw, you should reverse and re-run the regression.

- According to this model, for the expected birthweight to be 3500g, the mother + situation would need to differ from our sample's average in more ways than just cig, i.e. via a change which means $\mathbb{E}[u_i] \neq 0$, like higher income, etc.

OK. But note that the above doesn't imply 3500 can't happen. In fact it could be in your sample.

4a) ~~We can take FOCs, letting $C :=$~~

Let $C := \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$.

Note that because n is not a choice ~~Note that as n is not~~ variable,

Solving ~~via~~ $\arg\min_{b_0, b_1} \frac{1}{n} C$ also yields the solutions to $\arg\min_{b_0, b_1} C$.

Conveniently, $\frac{1}{n} C$ is simply the mean squared ~~prediction~~ error in the sample.

Taking FOCs, $\frac{\partial C}{\partial b_0} = \frac{2}{n} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i) \stackrel{!}{=} 0$ (i)

$\frac{\partial C}{\partial b_1} = -\frac{2}{n} \sum_{i=1}^{n} X_i (Y_i - b_0 - b_1 X_i) \stackrel{!}{=} 0$ (ii), solved by $(\hat{\beta}_0, \hat{\beta}_1)$

From (i), ~~one~~ $\hat{\beta}_0$ ~~which solves this~~ $= \frac{1}{n} \sum_{i=1}^{n} Y_i - \frac{1}{n} \sum_{i=1}^{n} \hat{\beta}_1 X_i$.

~~by rearranging of (i).~~

But this, by definition, means $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$, with $\bar{Z}$ the sample mean.

From (ii), $\frac{1}{n} \sum_{i=1}^{n} [X_i (Y_i - (\bar{Y} - \hat{\beta}_1 \bar{X}) - \hat{\beta}_1 X_i)] = 0$

so $\frac{1}{n} \sum (X_i (Y_i - \bar{Y})) + \hat{\beta}_1 \frac{1}{n} \sum ((\bar{X} - X_i)(X_i)) = 0$

$\Rightarrow \hat{\beta}_1 = \frac{1}{n} \cdot \sum [X_i (Y_i - \bar{Y})] \div (\frac{1}{n} \sum [X_i (X_i - \bar{X})])$

~~But $\hat{Cov}(X, Y) = \frac{1}{n}$~~

And by definition, $\hat{Cov}(X, Y) = \frac{1}{n} \sum [(X_i - \bar{X})(Y_i - \bar{Y})]$

don't totally get this (*) step though it'd probably make sense if I wrote out the summation expansion

as in slides $= \frac{1}{n} \sum [X_i (Y_i - \bar{Y})]$

So as required, $\hat{\beta}_1 = \frac{\hat{Cov}(X, Y)}{\hat{Var}(X)}$

It follows from the fact that for any R.V. Z, $\sum (Z_i - \bar{Z}) = 0$

Let $\bar{\hat{u}} := \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i = 0$ also.

b) From (i) and (ii) we have, by construction,
$\sum_{i=1}^{n} \hat{u}_i = \sum_{i=1}^{n} x_i \hat{u}_i = 0$, evaluated at $(\hat{\beta}_0, \hat{\beta}_1)$.

$$\hat{\text{cov}}(x, \hat{u}) := \sum_{i=1}^{n}\left[(x_i - \bar{x})(\hat{u}_i - \bar{\hat{u}})\right]$$

if we scaled by $\frac{1}{n}$, this part would also be 0, right?

$$= \underbrace{\sum x_i \hat{u}_i}_{0} - \bar{x}\underbrace{\sum \hat{u}_i}_{0} - \underbrace{\bar{\hat{u}}}_{0}\left(\left(\sum^{n} x_i\right) - \bar{x}\right)$$

$$= 0$$

realised when to write x vs x_i? random

c) $$\hat{\text{corr}}(x, Y) := \frac{\hat{\text{cov}}(x, Y)}{\sqrt{\hat{\text{var}}(x) \cdot \hat{\text{var}}(Y)}}$$

(*) cf d).

In the case where $\hat{\text{Var}}(Y) = \hat{\text{var}}(x)$, then $\hat{\beta}_1$ is exactly the sample correlation between x and y. Otherwise, it's the sample correlation scaled by the ratio of standard deviations $\frac{s_x}{s_Y}$.

(*) am a bit unsure about the connection to when $\mathbb{E}[Y|X]$ is nonlinear and $\hat{\beta}_1$ only being an approximation to the average change. As in, it's locally an approximation to $\frac{\partial Y}{\partial X}$ in a Taylor sense?

Good question → will discuss this in class.

This fits with its interpretation as the average number-of-units increase in Y associated with a unit increase in x, in the sample.

d) Substituting at $x = \bar{x}$ we have

$$y = \underbrace{(\bar{Y} - \hat{\beta}_1 \bar{x})}_{\hat{\beta}_0} + (\hat{\beta}_1 \cdot \bar{x})$$

$$= \bar{Y} \quad \text{as required}$$

(again, how to notate this?)

5. $Y_i = \beta_0 + \beta_1 x_i + u_i$ $\qquad \text{Cov}(x_i, u_i) = 0; \quad \mathbb{E}[u_i] \neq 0$

how come the subscript i's show up here?

If your sample is iid, it doesn't matter... i.e. you can drop the subscript

If we regress Y_i on X_i we'll end up with population parameters $\beta_0 = \mathbb{E}[Y] - \beta_1 \mathbb{E}[X]$; $\beta_1 = \frac{\text{Cov}(X_i, Y_i)}{\text{Var}(X_i)}$

$$\text{Cov}(X_i, Y_i) := \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}[Y])] \equiv \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

$$\text{Var}(X_i) \equiv \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

simpler here to use Cov rules directly.

Given that $Y_i = \beta_0 + \beta_1 X_i + u_i$,

$$Cov(X,Y) = \mathbb{E}[X_i(\beta_0 + \beta_1 X_i + u_i)] - \mathbb{E}[X_i]\mathbb{E}[\beta_0 + \beta_1 X_i + u_i]$$

by linearity of $\mathbb{E}$ $= \underbrace{\left(\mathbb{E}[X_i u_i] - \mathbb{E}[X_i]\mathbb{E}[u_i]\right)}_{Cov(X_i, u_i)} + \left(\mathbb{E}[X_i\beta_0 + \beta_1 X_i^2] - \mathbb{E}[X_i]\mathbb{E}[\beta_0 + \beta_1 X_i]\right)$

as β_0, β_1 are const. $= 0 + \beta_1\left(\mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2\right) + \beta_0\left[\underbrace{\mathbb{E}[X_i] - \mathbb{E}[X_i]}_{0}\right]$

$= \beta_1 Var(X_i)$

But then $\rho_1 = \dfrac{Cov(X,Y)}{Var(X)} = \beta_1$, exactly as required.

Q6. (If this is the regression model, why are the coefficients written as β, not ρ?)

When we regress Y on X, we obtain the ~~correlation~~ coefficient $\rho_1 = \dfrac{Cov(X,Y)}{Var(X)}$.

Given OR, i.e. $\mathbb{E}(u_i) = 0$ and $\mathbb{E}(X_i u_i) = 0$ in the causal model, β_1 solves the normal equations which pin down ρ_1 from FOCs, and so $\rho_1 = \beta_1$.

If we regress X on Y, we instead have $\tilde{\rho}_1 = \dfrac{Cov(X,Y)}{Var(Y)}$.

From 5., $Cov(X,Y) = \beta_1 Var(X_i)$ when OR is met.

And $Var(Y) = \beta_1^2 Var(X_i) + Var(u_i) + 2\beta_1 \underbrace{Cov(X_i, u_i)}_{0}$ by Cov properties

So $\tilde{\rho}_1 = \dfrac{\beta_1 Var(X_i)}{\beta_1^2 Var(X_i) + Var(u_i)}$ which $= \frac{1}{\beta_1} \iff Var(u_i) = 0$

Thus they do not coincide unless there are no other determinants (why is $\mathbb{E}[\mathbb{E}[Y|X]|X] = \mathbb{E}[Y|X]$ true? From Q6, and James Duffy appendix). Conditioning property about $g(X)$?

of Y besides X, i.e. there ~~are~~ is never any residual captured by u.

pls bring this up in class

8. $\mathbb{E}[(\mathbb{E}[Y|X] - (b_0 + b_1 X))^2] =: C(b_0, b_1)$

RTP: (β_0, β_1) solves $\underset{b_0, b_1}{\text{argmin}}\ C(b_0, b_1)$ (I)

By construction, (β_0, β_1) solves $\underset{b_0, b_1}{\text{argmin}}\ \underbrace{\mathbb{E}[(Y-(b_0+b_1X))^2]}_{Q}$

From the law of iterated expectations, we know that $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]]$. Let $m(X) := \mathbb{E}[Y|X]$.

Taking FOCs for (I), $\frac{\partial C}{\partial b_0} = -2\mathbb{E}[m(X) - (b_0 + b_1 X)] \overset{!}{=} 0$ (i)

$\frac{\partial C}{\partial b_1} = -2\mathbb{E}[X(m(X) - (b_0 + b_1 X))] \overset{!}{=} 0$ (ii)

From (i), $\mathbb{E}[m(X)] = \mathbb{E}[b_0 + b_1 X]$

(ii), $\mathbb{E}[m(X)X] = \mathbb{E}[(b_0 + b_1 X)X] = b_1\mathbb{E}[X^2] + b_0\mathbb{E}[X]$ } by linearity

But by LIE, these are simply the normal equations which we obtain to solve Q. So as the solution is unique, the same parameters will minimise both cost functions.

(i.e., the normal equations
$\mathbb{E}[Y - \beta_0 - \beta_1 X] = 0$
$\mathbb{E}[X(Y - \beta_0 - \beta_1 X)] = 0$)