Statistics refresher: interactions

Brice Ozenne (brice.ozenne@nru.dk)

[1] Neurobiology Research Unit, University Hospital of Copenhagen, Rigshospitalet.

[2] Section of Biostatistics, Department of Public Health, University of Copenhagen.

January 29th, 2024 - IPSY meeting

Why interactions?

Studying the effect of more than one exposure on an outcome

- e.g. how X_1 and X_2 affect the mean outcome Y
  $\mathbb{E}[Y] = f(X_1, X_2)$

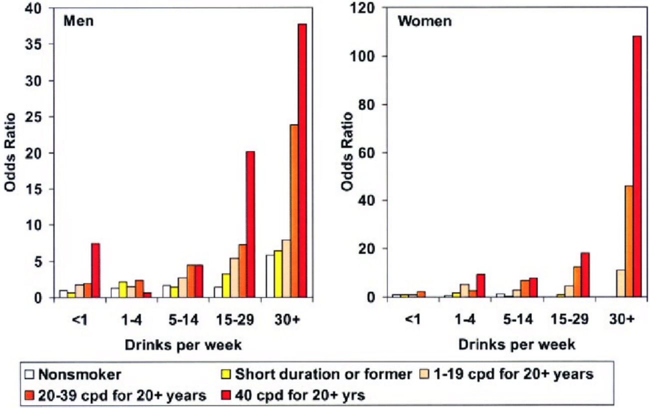
Exposures may not act independently!

- necessity: oxygen **and** fuel **and** heat are necessary to ignite fire
- reinforcement or inhibition: combining exposures lead to additional mechanisms enhancing or diminishing the effect
  - smoking and drinking on the risk of cancer (next slide)
  - chocolate [0.5mm] chips [0.5mm] but chips with chocolate [0.5mm]

Real life example (Blot WJ, 1988)

Smoking and drinking lead to high risk of pharyngeal cancer

- more than the addition of the (separate) effect of smoking and drinking.



Example of no interaction model

Without interaction we fit (linear) models such as:

$$\mathbb{E}[Y] = \alpha + \beta_1 X_1 + \beta_2 X_2$$

Example dataset[1]:

```
df <- subset(vitaminD,country %in% c("Denmark","Finland"))
e.lm <- lm(log10(vitd) ~ country + bmi, data = df)
summary(e.lm)
```

```
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.282737   0.100673  12.742   <2e-16 ***
countryFinland 0.032126   0.037434   0.858   0.3920
bmi            0.009672   0.003860   2.506   0.0132 *
```

[1] http://staff.pubhealth.ku.dk/~linearpredictors/datafiles/VitaminD.csv

What is easy without interactions? (1/3)

```
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.282737   0.100673  12.742   <2e-16 ***
countryFinland 0.032126   0.037434   0.858   0.3920
bmi            0.009672   0.003860   2.506   0.0132 *
```

1. `bmi`: is the bmi effect,
   i.e. the typical difference in log10 of the vitamin D level between two person from the same country but with a bmi differing by 1.

$$\beta_2 = \mathbb{E}[Y|X_1 = x_1 + 1, X_2 = x_2] - \mathbb{E}[Y|X_1 = x_1, X_2 = x_2]$$
$$\widehat{\beta}_2 \approx 0.001352$$

- similarly for the country effect

What is easy without interactions? (2/3)

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.282737   0.100673  12.742   <2e-16 ***
countryFinland 0.032126   0.037434   0.858   0.3920
bmi            0.009672   0.003860   2.506   0.0132 *
```

1. `bmi`: is the bmi effect
2. `bmi`: p-value is (very) similar to a likelihood ratio test

```
e.lm0 <- lm(log10(vitd) ~ country, data = df)
anova(e.lm, e.lm0)
```

```
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    162 9.3237
2    163 9.6852 -1  -0.36148 6.2807 0.01319 *
```

What is easy without interactions? (3/3)

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.282737   0.100673  12.742   <2e-16 ***
countryFinland 0.032126   0.037434   0.858   0.3920
bmi            0.009672   0.003860   2.506   0.0132 *
```

1. bmi: is the bmi effect
2. bmi: p-value is (very) similar to a likelihood ratio test
3. Estimate/p-value are unaffected by centering continuous covariates
   (except the intercept)

```
summary(lm(log10(vitd) ~ country + I(bmi-25), data = df))
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.524548   0.027168  56.117   <2e-16 ***
countryFinland 0.032126   0.037434   0.858   0.3920
I(bmi - 25)    0.009672   0.003860   2.506   0.0132 *
```

What is difficult with interactions? (1/4)

```
e.lmI <- lm(log10(vitd) ~ country * bmi, data = df)
summary(e.lmI)
```

```
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.313965   0.148551   8.845 1.55e-15 ***
countryFinland     -0.024308   0.200474  -0.121    0.904
bmi                 0.008429   0.005814   1.450    0.149
countryFinland:bmi  0.002233   0.007792   0.287    0.775
```

1. The significance levels **look** different between considering or not interactions
   - different statistical hypotheses are considered
   - more complex model can have lower (more parameters to estimate) or higher power (less residual noise), depending on the data.

What is difficult with interactions? (2/4)

```
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.313965   0.148551   8.845 1.55e-15 ***
countryFinland     -0.024308   0.200474  -0.121    0.904
bmi                 0.008429   0.005814   1.450    0.149
countryFinland:bmi  0.002233   0.007792   0.287    0.775
```

1. The significance levels **look** different between considering or not interactions
2. The p-value of `bmi` does **not** assess the evidence for a bmi effect on the outcome
   - only the effect for a specific country (here Denmark)
   - a LRT vs. a model without `bmi` does

```
anova(e.lmI, e.lm0)
```

```
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    161 9.3190
2    163 9.6852 -2  -0.36623 3.1636 0.04491 *
```

What is difficult with interactions? (3/4)

```
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.313965   0.148551   8.845 1.55e-15 ***
countryFinland     -0.024308   0.200474  -0.121    0.904
bmi                 0.008429   0.005814   1.450    0.149
countryFinland:bmi  0.002233   0.007792   0.287    0.775
```

1. The significance levels **look** different between considering or not interactions
2. The p-value of `bmi` does **not** assess the evidence for a bmi effect on the outcome
3. Estimate/p-value are affected by centering continuous covariates (except the interaction)

```
summary(lm(log10(vitd) ~ country * I(bmi-25), data = df))
```

```
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.524693   0.027250  55.953   <2e-16
countryFinland           0.031513   0.037601   0.838    0.403
I(bmi - 25)              0.008429   0.005814   1.450    0.149
countryFinland:I(bmi - 25) 0.002233   0.007792   0.287    0.775
```

What is difficult with interactions? (4/4)

```
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        1.313965   0.148551   8.845 1.55e-15 ***
countryFinland    -0.024308   0.200474  -0.121    0.904
bmi                0.008429   0.005814   1.450    0.149
countryFinland:bmi 0.002233   0.007792   0.287    0.775
```

1. The significance levels **look** different between considering or not interactions
2. The p-value of `bmi` does **not** assess the evidence for a bmi effect on the outcome
3. Estimate/p-value are affected by centering covariates (except the interaction)
4. One has to report several estimates for a given exposure:
   - BMI effect in Denmark: 0.008429
   - BMI effect in Finland: 0.008429+0.002233=0.010662

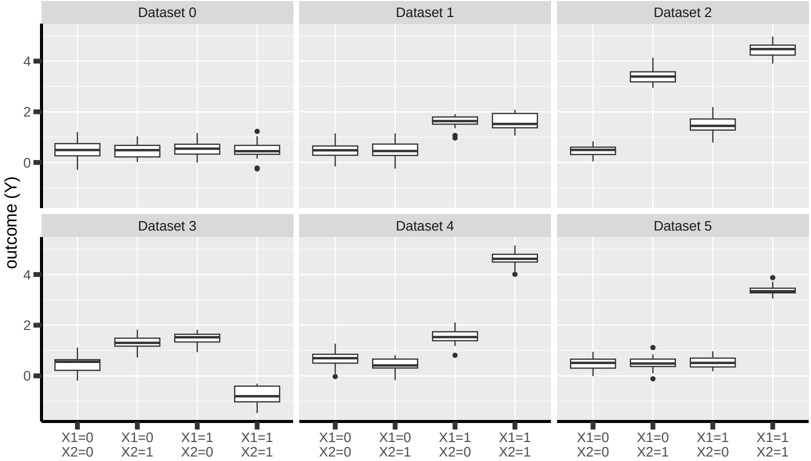
General advise

Make a graphical representation of the data and the model fit

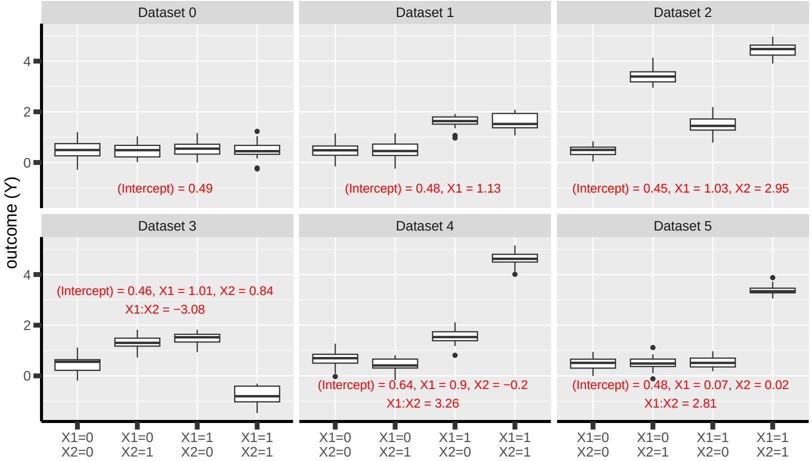
- to retrieve what each regression coefficient means
- possibly in a simplified model, i.e., without covariates

Interaction between binary variables

Illustrative datasets



Estimated regression coefficients



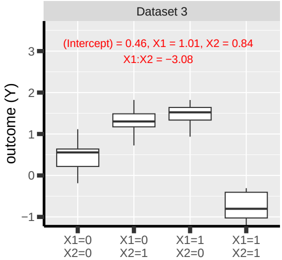
Expliciting the statistical model

$$
\begin{aligned}
\mathbb{E}\left[Y\right] &= \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma X_1 : X_2 \\
&= \begin{cases}
\alpha & \text{when } X_1 = X_2 = 0 \\
\alpha + \beta_2 & \text{when } X_1 = 0, X_2 = 1 \\
\alpha + \beta_1 & \text{when } X_1 = 1, X_2 = 0 \\
\alpha + \beta_1 + \beta_2 + \gamma & \text{when } X_1 = X_2 = 1
\end{cases}
\end{aligned}
$$

Expliciting the statistical model

$$\mathbb{E}[Y] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma X_1 : X_2$$
$$= \begin{cases} \alpha & \text{when } X_1 = X_2 = 0 \\ \alpha + \beta_2 & \text{when } X_1 = 0, X_2 = 1 \\ \alpha + \beta_1 & \text{when } X_1 = 1, X_2 = 0 \\ \alpha + \beta_1 + \beta_2 + \gamma & \text{when } X_1 = X_2 = 1 \end{cases}$$

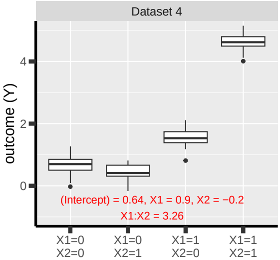
$$\mathbb{E}[Y|X_1 = 0, = X_2 = 0] = 0.46$$
$$\mathbb{E}[Y|X_1 = 0, = X_2 = 1] = 0.46 + 0.84 = 1.3$$
$$\mathbb{E}[Y|X_1 = 1, = X_2 = 0] = 0.46 + 1.01 = 1.47$$
$$\mathbb{E}[Y|X_1 = 1, = X_2 = 1] = 0.46 + 0.84 + 1.01 - 3.08$$
$$= -0.77$$



Expliciting the statistical model

$$\mathbb{E}[Y] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma X_1 : X_2$$
$$= \begin{cases} \alpha & \text{when } X_1 = X_2 = 0 \\ \alpha + \beta_2 & \text{when } X_1 = 0, X_2 = 1 \\ \alpha + \beta_1 & \text{when } X_1 = 1, X_2 = 0 \\ \alpha + \beta_1 + \beta_2 + \gamma & \text{when } X_1 = X_2 = 1 \end{cases}$$

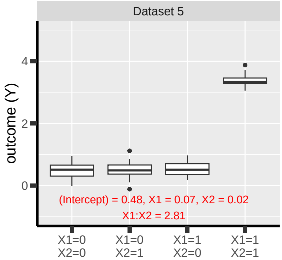
$$\mathbb{E}[Y|X_1 = 0, = X_2 = 0] = 0.64$$
$$\mathbb{E}[Y|X_1 = 0, = X_2 = 1] = 0.64 - 0.2 = 0.44$$
$$\mathbb{E}[Y|X_1 = 1, = X_2 = 0] = 0.64 + 0.9 = 1.54$$
$$\mathbb{E}[Y|X_1 = 1, = X_2 = 1] = 0.64 - 0.2 + 0.9 + 3.26$$
$$= 4.6$$



Expliciting the statistical model

$$\mathbb{E}[Y] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma X_1 : X_2$$
$$= \begin{cases} \alpha & \text{when } X_1 = X_2 = 0 \\ \alpha + \beta_2 & \text{when } X_1 = 0, X_2 = 1 \\ \alpha + \beta_1 & \text{when } X_1 = 1, X_2 = 0 \\ \alpha + \beta_1 + \beta_2 + \gamma & \text{when } X_1 = X_2 = 1 \end{cases}$$

$$\mathbb{E}[Y|X_1 = 0, = X_2 = 0] = 0.48$$
$$\mathbb{E}[Y|X_1 = 0, = X_2 = 1] = 0.48 + 0.07 = 0.55$$
$$\mathbb{E}[Y|X_1 = 1, = X_2 = 0] = 0.48 + 0.02 = 0.50$$
$$\mathbb{E}[Y|X_1 = 1, = X_2 = 1] = 0.48 + 0.07 + 0.02 + 2.81$$
$$= 3.38$$



Expliciting the effect of one exposure

Given the model

$$\mathbb{E}\left[Y|X_1, X_2\right] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma X_1 : X_2$$

The effects of X_1 are:

$$\mathbb{E}\left[Y|X_1 = 1, X_2 = 0\right] - \mathbb{E}\left[Y|X_1 = 0, X_2 = 0\right] \quad \text{(among } X_2 = 0\text{)}$$
$$= (\alpha + \beta_1) - (\alpha) = \beta_1$$
$$\mathbb{E}\left[Y|X_1 = 1, X_2 = 1\right] - \mathbb{E}\left[Y|X_1 = 0, X_2 = 1\right] \quad \text{(among } X_2 = 1\text{)}$$
$$= (\alpha + \beta_1 + \beta_2 + \gamma) - (\alpha + \beta_2) = \beta_1 + \gamma$$

Expliciting the effect of one exposure

Given the model

$$\mathbb{E}\left[Y|X_1,X_2\right] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma X_1 : X_2$$

The effects of X_1 are:

$$\mathbb{E}\left[Y|X_1=1,X_2=0\right] - \mathbb{E}\left[Y|X_1=0,X_2=0\right] \quad (\text{among } X_2=0)$$

$$= (\alpha+\beta_1) - (\alpha) = \beta_1$$

$$\mathbb{E}\left[Y|X_1=1,X_2=1\right] - \mathbb{E}\left[Y|X_1=0,X_2=1\right] \quad (\text{among } X_2=1)$$

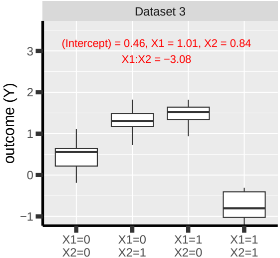
$$= (\alpha+\beta_1+\beta_2+\gamma) - (\alpha+\beta_2) = \beta_1+\gamma$$

$$\mathbb{E}\left[Y|X_1=1,=X_2=0\right] - \mathbb{E}\left[Y|X_1=0,=X_2=0\right]$$

$$= 1.62 - 0.61 = 1.01$$

$$\mathbb{E}\left[Y|X_1=1,=X_2=1\right] - \mathbb{E}\left[Y|X_1=0,=X_2=1\right]$$

$$= -0.62 - 1.45 = -2.97$$



Expliciting the effect of one exposure

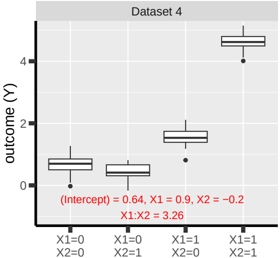
Given the model

$$\mathbb{E}\left[Y|X_1, X_2\right] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma X_1 : X_2$$

The effects of X_1 are:

$$\mathbb{E}\left[Y|X_1 = 1, X_2 = 0\right] - \mathbb{E}\left[Y|X_1 = 0, X_2 = 0\right] \quad \text{(among } X_2 = 0\text{)}$$
$$= (\alpha + \beta_1) - (\alpha) = \beta_1$$
$$\mathbb{E}\left[Y|X_1 = 1, X_2 = 1\right] - \mathbb{E}\left[Y|X_1 = 0, X_2 = 1\right] \quad \text{(among } X_2 = 1\text{)}$$
$$= (\alpha + \beta_1 + \beta_2 + \gamma) - (\alpha + \beta_2) = \beta_1 + \gamma$$

$$\mathbb{E}\left[Y|X_1 = 1, = X_2 = 0\right] - \mathbb{E}\left[Y|X_1 = 0, = X_2 = 0\right]$$
$$= 1.54 - 0.64 = 0.9$$
$$\mathbb{E}\left[Y|X_1 = 1, = X_2 = 1\right] - \mathbb{E}\left[Y|X_1 = 0, = X_2 = 1\right]$$
$$= 4.6 - 0.44 = 4.16$$



Expliciting the effect of one exposure

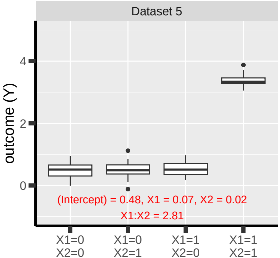
Given the model

$$\mathbb{E}\left[Y|X_1, X_2\right] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma X_1 : X_2$$

The effects of X_1 are:

$$\mathbb{E}\left[Y|X_1 = 1, X_2 = 0\right] - \mathbb{E}\left[Y|X_1 = 0, X_2 = 0\right] \quad \text{(among } X_2 = 0\text{)}$$
$$= (\alpha + \beta_1) - (\alpha) = \beta_1$$
$$\mathbb{E}\left[Y|X_1 = 1, X_2 = 1\right] - \mathbb{E}\left[Y|X_1 = 0, X_2 = 1\right] \quad \text{(among } X_2 = 1\text{)}$$
$$= (\alpha + \beta_1 + \beta_2 + \gamma) - (\alpha + \beta_2) = \beta_1 + \gamma$$

$$\mathbb{E}\left[Y|X_1 = 1, = X_2 = 0\right] - \mathbb{E}\left[Y|X_1 = 0, = X_2 = 0\right]$$
$$= 0.55 - 0.48 = 0.07$$
$$\mathbb{E}\left[Y|X_1 = 1, = X_2 = 1\right] - \mathbb{E}\left[Y|X_1 = 0, = X_2 = 1\right]$$
$$= 3.38 - 0.50 = 2.88$$



Different parametrisation in R

```
df$bmi25 <- factor(df$bmi>25)
summary(lm(log10(vitd) ~ country * bmi25, data = df))
```

```
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.51886    0.03839  39.566   <2e-16 **
countryFinland            -0.03368    0.05395  -0.624   0.5333
bmi25TRUE                  0.01363    0.05429   0.251   0.8021
countryFinland:bmi25TRUE   0.12716    0.07488   1.698   0.0914 .
```

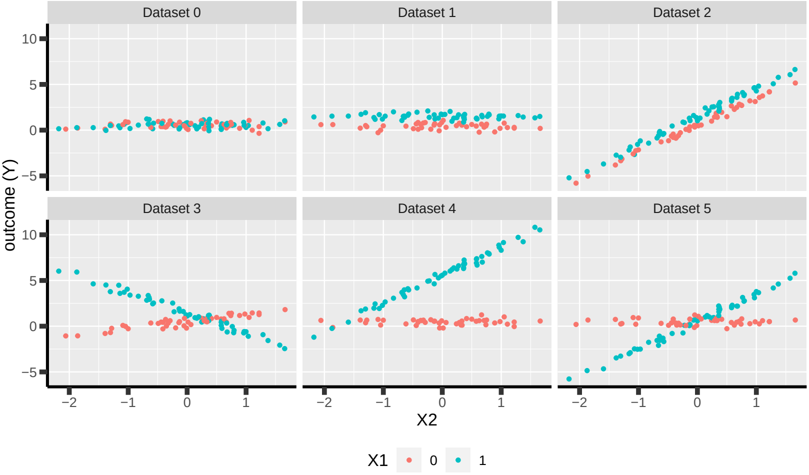
```
summary(lm(log10(vitd) ~ 0+country:bmi25, data = df))
```

```
                           Estimate Std. Error t value Pr(>|t|)
countryDenmark:bmi25FALSE   1.51886    0.03839   39.57   <2e-16 *
countryFinland:bmi25FALSE   1.48519    0.03791   39.18   <2e-16 *
countryDenmark:bmi25TRUE    1.53249    0.03839   39.92   <2e-16 *
countryFinland:bmi25TRUE    1.62598    0.03497   46.50   <2e-16 *
```

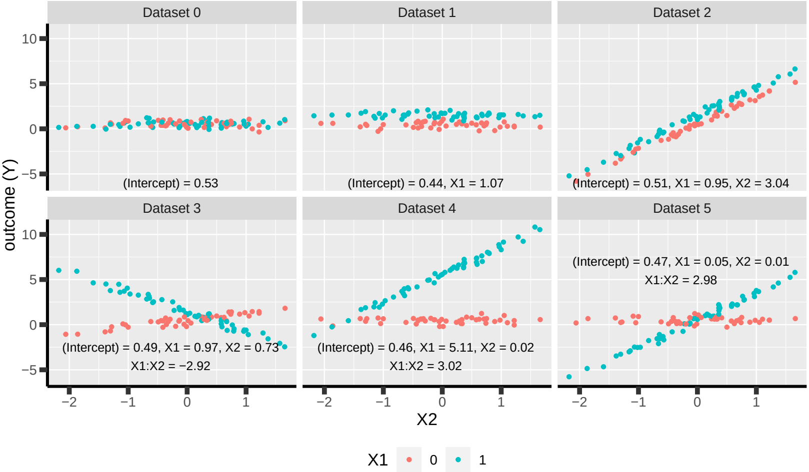
Same models expressed differently!

Interaction between binary and continuous variables

Illustrative datasets



Estimated regression coefficients



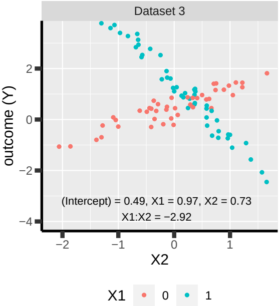
Expliciting the statistical model

$$\mathbb{E}\left[Y|X_1, X_2\right] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma X_1 : X_2$$
$$= \begin{cases} \alpha + \beta_2 X_2 & \text{when } X_1 = 0 \\ (\alpha + \beta_1) + (\beta_2 + \gamma) X_2 & \text{when } X_1 = 1 \end{cases}$$

Expliciting the statistical model

$$\mathbb{E}\left[Y|X_1, X_2\right] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma X_1 : X_2$$
$$= \begin{cases} \alpha + \beta_2 X_2 & \text{when } X_1 = 0 \\ (\alpha + \beta_1) + (\beta_2 + \gamma) X_2 & \text{when } X_1 = 1 \end{cases}$$

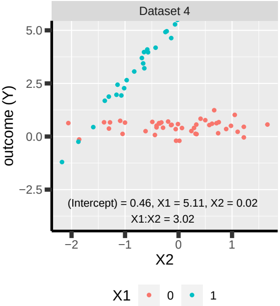
$$\mathbb{E}\left[Y|X_1 = 0, X_2\right] = 0.49 + 0.73 X_2$$
$$\mathbb{E}\left[Y|X_1 = 1, X_2\right] = (0.49 + 0.97) + (0.73 - 2.92) X_2$$
$$= 1.46 - 2.19 X_2$$



Expliciting the statistical model

$$\mathbb{E}[Y|X_1, X_2] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma X_1 : X_2$$
$$= \begin{cases} \alpha + \beta_2 X_2 & \text{when } X_1 = 0 \\ (\alpha + \beta_1) + (\beta_2 + \gamma) X_2 & \text{when } X_1 = 1 \end{cases}$$

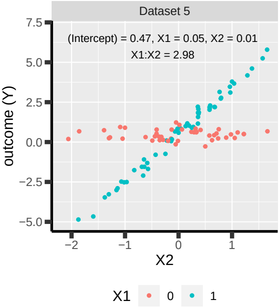
$$\mathbb{E}[Y|X_1 = 0, X_2] = 0.46 + 0.02 X_2$$
$$\mathbb{E}[Y|X_1 = 1, X_2] = (0.46 + 5.11) + (0.02 + 3.02) X_2 = 5.57 + 3.04 X_2$$



Expliciting the statistical model

$$\mathbb{E}\left[Y|X_1, X_2\right] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma X_1 : X_2$$
$$= \begin{cases} \alpha + \beta_2 X_2 & \text{when } X_1 = 0 \\ (\alpha + \beta_1) + (\beta_2 + \gamma) X_2 & \text{when } X_1 = 1 \end{cases}$$

$$\mathbb{E}\left[Y|X_1 = 0, X_2\right] = 0.47 + 0.01 X_2$$
$$\mathbb{E}\left[Y|X_1 = 1, X_2\right] = (0.47 + 0.05) + (0.01 - 2.98) X_2$$
$$= 0.52 + 2.99 X_2$$



Expliciting the effect of one exposure

Given the model

$$\mathbb{E}\left[Y|X_1, X_2\right] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma X_1 : X_2$$

The effects of X_1 is:

$$\mathbb{E}\left[Y|X_1 = 1, X_2 = x\right] - \mathbb{E}\left[Y|X_1 = 0, X_2 = x_2\right] = \beta_1 + \gamma x_2$$

i.e. 0 at $x = -\beta_1/\gamma$ and non-0 otherwise.

Expliciting the effect of one exposure

Given the model

$$\mathbb{E}\left[Y|X_1, X_2\right] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma X_1 : X_2$$

The effects of X_1 is:

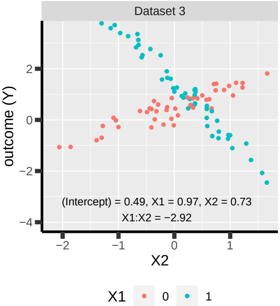
$$\mathbb{E}\left[Y|X_1 = 1, X_2 = x\right] - \mathbb{E}\left[Y|X_1 = 0, X_2 = x_2\right] = \beta_1 + \gamma x_2$$

i.e. 0 at $x = -\beta_1/\gamma$ and non-0 otherwise.

$\mathbb{E}\left[Y|X_1 = 1, = X_2 = x_2\right] - \mathbb{E}\left[Y|X_1 = 0, = X_2 = x_2\right]$
$= 0.97 - 2.92 x_2$

⚠ Non-0 effect of X_1 except at $X_2 \approx 0.33$

Centering X_2 around this value would make $\beta_1 = 0$



Expliciting the effect of one exposure

Given the model

$$\mathbb{E}\left[Y|X_1, X_2\right] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma X_1 : X_2$$

The effects of X_1 is:

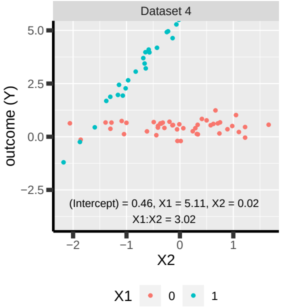
$$\mathbb{E}\left[Y|X_1 = 1, X_2 = x\right] - \mathbb{E}\left[Y|X_1 = 0, X_2 = x_2\right] = \beta_1 + \gamma x_2$$

i.e. 0 at $x = -\beta_1/\gamma$ and non-0 otherwise.

$\mathbb{E}\left[Y|X_1 = 1, = X_2 = x_2\right] - \mathbb{E}\left[Y|X_1 = 0, = X_2 = x_2\right]$
$= 5.11 + 3.02 x_2$

⚠ Non-0 effect of X_1 except at $X_2 \approx -1.5$

Centering X_2 around this value would make $\beta_1 = 0$



Expliciting the effect of one exposure

Given the model

$$\mathbb{E}\left[Y|X_1, X_2\right] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma X_1 : X_2$$

The effects of X_1 is:

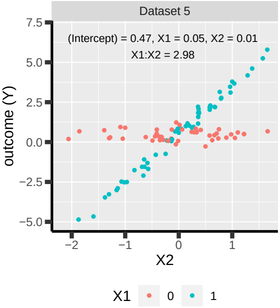
$$\mathbb{E}\left[Y|X_1 = 1, X_2 = x\right] - \mathbb{E}\left[Y|X_1 = 0, X_2 = x_2\right] = \beta_1 + \gamma x_2$$

i.e. 0 at $x = -\beta_1/\gamma$ and non-0 otherwise.

$\mathbb{E}\left[Y|X_1 = 1, = X_2 = x_2\right] - \mathbb{E}\left[Y|X_1 = 0, = X_2 = x_2\right]$
$= 0.05 + 2.98 x_2$

⚠ Non-0 effect of X_1 except at $X_2 \approx 0$

Centering X_2 around this value would make $\beta_1 = 0$



Reference I

Blot WJ, McLaughlin JK, W. D. A. D. G. R. P.-M. S. B. L. S. J. S. A. F. J. J. (1988). Smoking and drinking in relation to oral and pharyngeal cancer. *Cancer Res*, 48(11):3282–3287.