Logistic regression.
Some more on confounders & colliders.

Manuela Zucknick
Oslo Center for Biostatistics and Epidemiology, UiO
manuela.zucknick@medisin.uio.no

MF9130 – Introductory Statistics
May 11, 2023

Recap

	Disease	No disease
Exposed	a	c
Not exposed	b	d

- Odds for disease among the exposed

$$\frac{\hat{p}_1}{1-\hat{p}_1} = \frac{a/(a+c)}{c/(a+c)} = \frac{a}{c}$$

- Odds for disease among the non-exposed

$$\frac{\hat{p}_0}{1-\hat{p}_0} = \frac{b/(b+d)}{d/(b+d)} = \frac{b}{d}$$

- Estimated odds ratio

$$OR = \frac{a/c}{b/d} = \frac{a \times d}{b \times c}$$

Example: Smoking and low birth weight (birth.csv)

	LOW $\leq$ 2500	LOW $>$ 2500
SMK $= 1$	30	44
SMK $= 0$	29	86

- $OR = \frac{a \times d}{b \times c} = \frac{30 \times 86}{2944} = 2.02$
- 95% confidence interval:

$$\left(e^{\ln(\text{OR}) - 1.96 \text{SE}(\ln(\text{OR}))}, e^{\ln(\text{OR}) + 1.96 \text{SE}(\ln(\text{OR}))}\right),$$

where $\text{SE}(\ln(\text{OR})) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$

- With numbers from table: $(1.08, 3.78)$

Regression analysis

- Response variable (dependent variable) Y,
- Predictor variables (independent variables) $X_1, \ldots, X_n$,

- Want to establish a simple formula that provides good predictions of the outcomes of Y based on the outcomes of $X_1, \ldots, X_n$,

Example: multiple linear regression

$$Y = \beta_0 + \beta_1 X_1 + \ldots \beta_n X_n$$

- Y continuous variable, and $X_1, \ldots, X_n$ continuous or categorical,
- Example (birth.csv):
    - Y birth weight,
    - X_1 Weight of mother,
    - X_2 Smoking,
    - Hypertension,
    - Age.

Logistic regression

- Response variable is dichotomous, a variable that typically is 1 if a person has a given disease, and 0 if it does not,

- $p = P(Y = 1|x_1, \ldots, x_n)$ is the (conditional) probability that the person has the disease,
- $1 - p = P(Y = 0|x_1, \ldots, x_n)$ is the (conditional) probability that the person does not have the disease,
- $0 \leq p \leq 1$.

Logistic regression

- Assume that p depends on the outcomes $x_1, \ldots, x_n$,
- We want to describe the function

$$p = p(x_1, \ldots, x_n),$$

- Works better to go through odds:

$$\text{Odds} = \frac{p}{1 - p}$$

Logistic regression

- Model for odds:

$$\frac{p}{1-p} = \exp(\beta_0 + \beta_1 x_1 + \ldots \beta_n x_n)$$

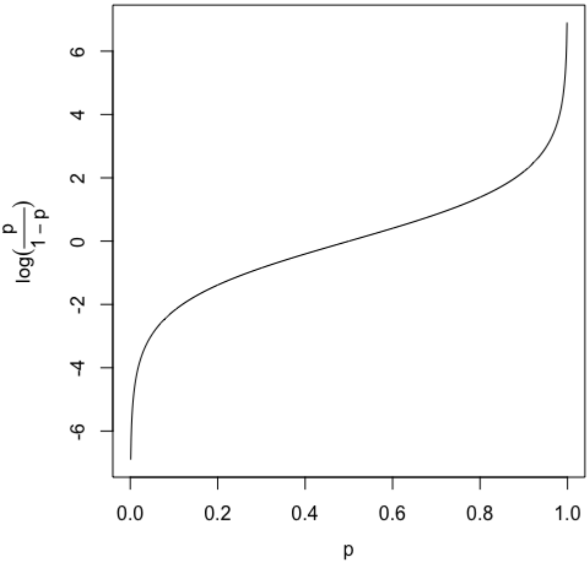
- Apply logarithm on both sides:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \ldots \beta_n x_n,$$

- Or equivalently:

$$p(x_1, \ldots, x_n) = \frac{\exp(\beta_0 + \beta_1 x_1 + \ldots \beta_n x_n)}{1 + \exp(\beta_0 + \beta_1 x_1 + \ldots \beta_n x_n)}.$$

The logit function



Example

- Want to identify risk factors for low birth weight,
- “birth.csv” contains data on 189 women,
- Response variable [LOW]: 1 means $\leq 2500g$ and 0 means ≥ 2500,
- Some explanatory variables:

AGE	Mother’s age,
LWT	Weight before pregnancy,
ETH	Ethnicity,
SMK	Smoking during pregnancy.

Example (cont.): logistic regression

- χ^2-test gives a significant association ($p = 0.026$),
- We can use logistic regression to estimate the odds ratio,
- p is the risk of low birth weight,
- x is the smoking status of the mother,

- The model:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x$$

Logistic regression and odds ratio

- Odds for smokers

$$\text{Odds}_{X=1} = e^{\beta_0+\beta_1\cdot 1}$$

- Odds for non-smokers

$$\text{Odds}_{X=0} = e^{\beta_0+\beta_1\cdot 0}$$

- Odds ratio:

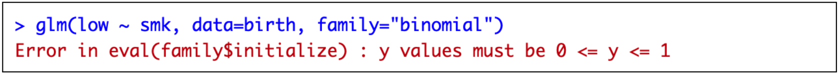
$$\text{OR} = \frac{\text{Odds}_{X=1}}{\text{Odds}_{X=0}} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}.$$

- Logistic regression gives estimated odds ratio.

Logistic regression in R

- Dependent variable: LOW. Independent variable: SMK.
- We use the command `glm(..., family="binomial")` (glm for **generalized linear model**)

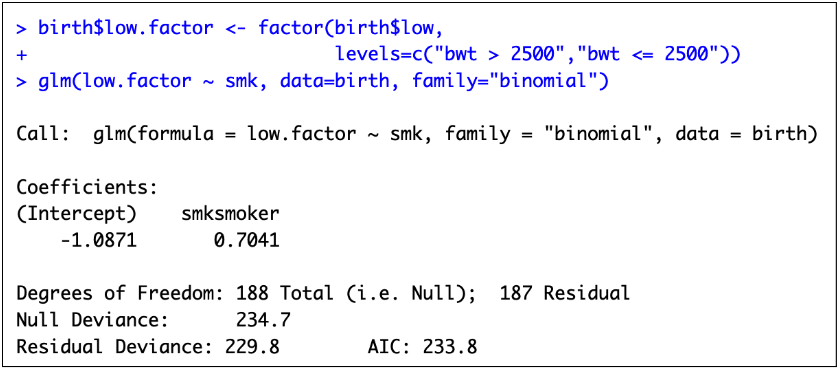
- Note that the dependent variable needs to be coded as 0/1 or be a factor variable.
- Here, LOW is a character variable, which results in an error message. LOW needs to be transformed.



```
> glm(low ~ smk, data=birth, family="binomial")
Error in eval(family$initialize) : y values must be 0 <= y <= 1
```

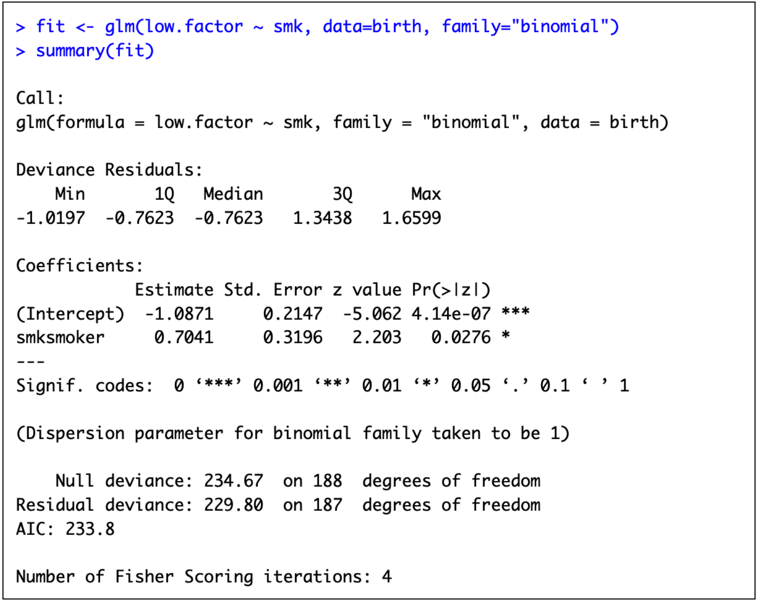
Logistic regression in R

- We decide to make a new factor variable out of LOW. Be careful to make sure that normal birthweight `bwt > 2500` is used as the reference category!



```
> birth$low.factor <- factor(birth$low,
+                             levels=c("bwt > 2500","bwt <= 2500"))
> glm(low.factor ~ smk, data=birth, family="binomial")

Call:  glm(formula = low.factor ~ smk, family = "binomial", data = birth)

Coefficients:
(Intercept)    smksmoker
    -1.0871       0.7041

Degrees of Freedom: 188 Total (i.e. Null);  187 Residual
Null Deviance:      234.7
Residual Deviance: 229.8        AIC: 233.8
```

Use the summary() function for more output



```
> fit <- glm(low.factor ~ smk, data=birth, family="binomial")
> summary(fit)

Call:
glm(formula = low.factor ~ smk, family = "binomial", data = birth)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0197  -0.7623  -0.7623   1.3438   1.6599

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.0871     0.2147  -5.062 4.14e-07 ***
smksmoker     0.7041     0.3196   2.203   0.0276 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 229.80  on 187  degrees of freedom
AIC: 233.8

Number of Fisher Scoring iterations: 4
```

- The model: $\log(\text{Odds}) = \beta_0 + \beta_1 \cdot \text{SMK}$,
- The *first column* gives the estimates of the regression coefficients, $\hat{\beta}_0 = -1.087$ and $\hat{\beta}_1 = 0.704$,
- The *second column* gives their standard errors, $\widehat{\text{SE}}(\hat{\beta}_0) = 0.215$ and $\widehat{\text{SE}}(\hat{\beta}_1) = 0.320$,

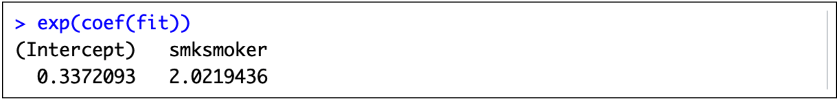
- The odds ratio can also be computed from $\hat{\beta}_1$ (and the CIs):

$$\widehat{\text{OR}} = e^{\hat{\beta}_1} = e^{0.704} = 2.02,$$

(and the same for the lower and upper bound of the 95% CI).

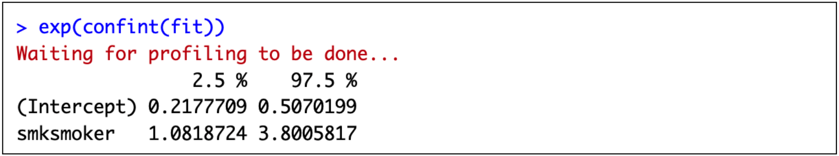
For the odds ratio and its confidence interval, we exponentiate the output

- Odds ratios:



```
> exp(coef(fit))
(Intercept)   smksmoker
  0.3372093   2.0219436
```

- 95% confidence intervals of the odds ratios:



```
> exp(confint(fit))
Waiting for profiling to be done...
                2.5 %    97.5 %
(Intercept) 0.2177709 0.5070199
smksmoker   1.0818724 3.8005817
```

Results for SMK:
$\widehat{\text{OR}} = 2.02$, 95% CI $= (1.08, 3.80)$, p-value$=0.028$

Additional explanatory variables

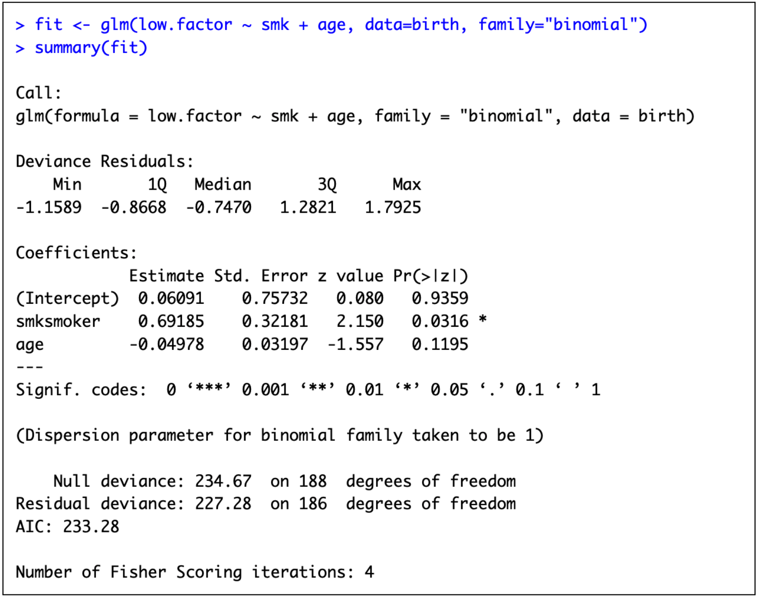
- Want to incorporate age into the regression model,
- The new model:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 \cdot \mathsf{SMK} + \beta_2 \cdot \mathsf{AGE}$$

- Now OR $= e^{\beta_1}$ describes the effect of smoking on the risk of low birth weight, *when adjusted for age*

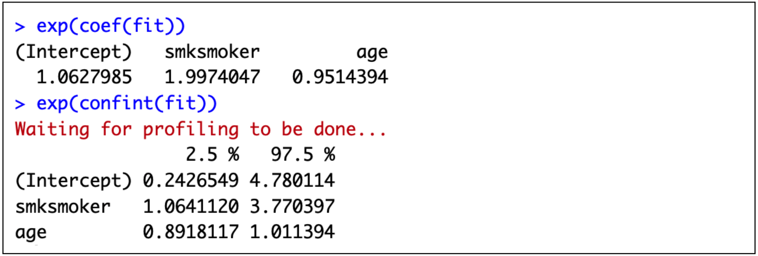
- Comparing two women with the same age, one is smoking and the other is not. The odds for the smoker is e^{β_1} times the odds for the non-smoker.

R output



```
> fit <- glm(low.factor ~ smk + age, data=birth, family="binomial")
> summary(fit)

Call:
glm(formula = low.factor ~ smk + age, family = "binomial", data = birth)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1589  -0.8668  -0.7470   1.2821   1.7925

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.06091    0.75732   0.080   0.9359
smksmoker    0.69185    0.32181   2.150   0.0316 *
age         -0.04978    0.03197  -1.557   0.1195
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 227.28  on 186  degrees of freedom
AIC: 233.28

Number of Fisher Scoring iterations: 4
```

R output



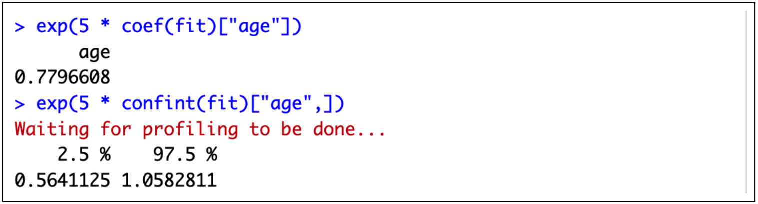
```
> exp(coef(fit))
(Intercept)   smksmoker         age
  1.0627985   1.9974047   0.9514394
> exp(confint(fit))
Waiting for profiling to be done...
                2.5 %   97.5 %
(Intercept) 0.2426549 4.780114
smksmoker   1.0641120 3.770397
age         0.8918117 1.011394
```

- Note that OR for smoker vs non-smokers does not change much when we take age into account (from 2.022 to 1.997),
- Interpretation of β_2: Increasing age by 1 year corresponds to multiplying the odds with the factor $e^{\hat{\beta}_2} = 0.951$,
- Age does not seem to have a significant effect, $p = 0.119$.

OR for an increase in AGE by 5 years

- Often we are interested in estimating the change in the outcome for more than 1 year, so for example for $c = 5$ years.
- Then we have: $\widehat{\text{OR}} = e^{c \cdot \hat{\beta}_i}$, and the 95% CI is estimated as:

$$\left(\exp(c \cdot \hat{\beta}_i - 1.96 \cdot c \cdot \widehat{\text{SE}}(\hat{\beta}_i)), \exp(c \cdot \hat{\beta}_i + 1.96 \cdot c \cdot \widehat{\text{SE}}(\hat{\beta}_i))\right)$$



```
> exp(5 * coef(fit)["age"])
      age
0.7796608
> exp(5 * confint(fit)["age",])
Waiting for profiling to be done...
    2.5 %    97.5 %
0.5641125 1.0582811
```

Results for increase in AGE by 5 years:
$\widehat{\text{OR}} = 0.78$, 95% CI $= (0.56,1.06)$, p-value$=0.119$

Note: The p-value is the same as for increase by 1 year. The 95% CI of the OR includes 1, confirming no significance at the 5% level.

Categorical variables with more than two levels

- Are included in the analysis with dummy variables
- Construct two dummy-variables to include ethnicity

ETH	Eth(1)	Eth(2)
White	0	0
Black	1	0
Other	0	1

- A simple univariable model including only ethnicity is then:

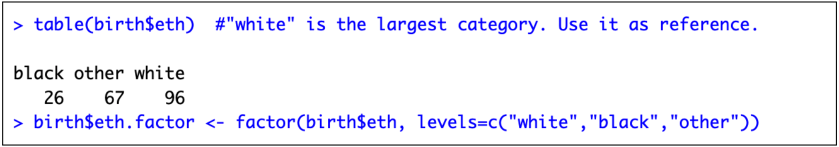
$$\log \frac{p}{1-p} = \beta_0 + \beta_1 \cdot \text{Eth}(1) + \beta_2 \cdot \text{Eth}(2)$$

- A more complicated multivariable model:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 \cdot \text{SMK} + \beta_2 \cdot \text{AGE} + \beta_3 \cdot \text{Eth}(1) + \beta_4 \cdot \text{Eth}(2)$$

Dummy variables in R

- When using a variable with more than 2 categories, we need to decide which category should be the reference.
- Here, we use ”white”, because it is the largest.

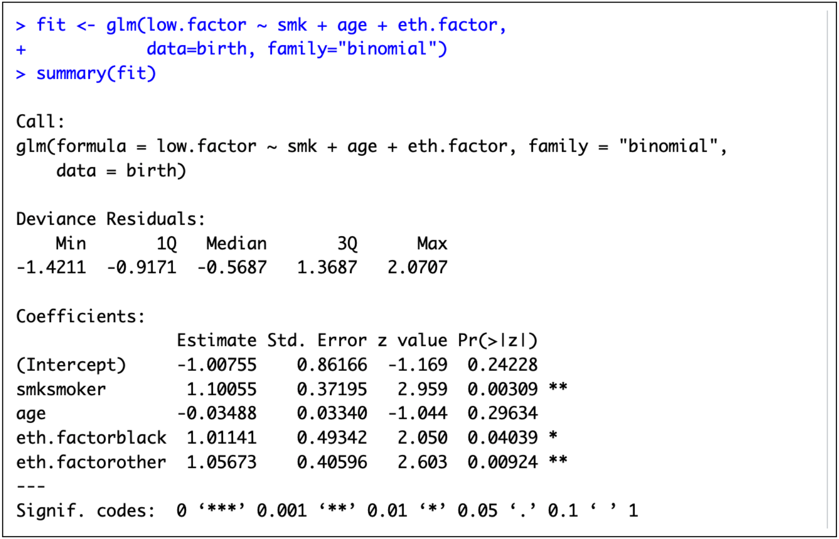


```
> table(birth$eth)  #"white" is the largest category. Use it as reference.

black other white 
   26    67    96 
> birth$eth.factor <- factor(birth$eth, levels=c("white","black","other"))
```

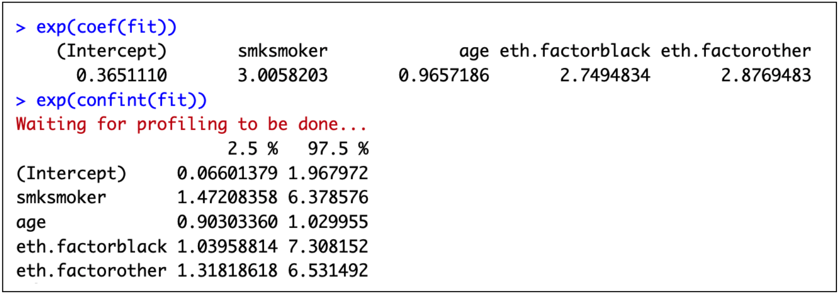
Dummy variables in R

- See R output on the next slides.

- ETH becomes statistically significant in the model with AGE and SMK ($p = 0.0193$)
- The *adjusted odds ratios* are $\widehat{\mathrm{OR}} = 2.75$ for *black vs white* and $\widehat{\mathrm{OR}} = 2.88$ for *other vs white*



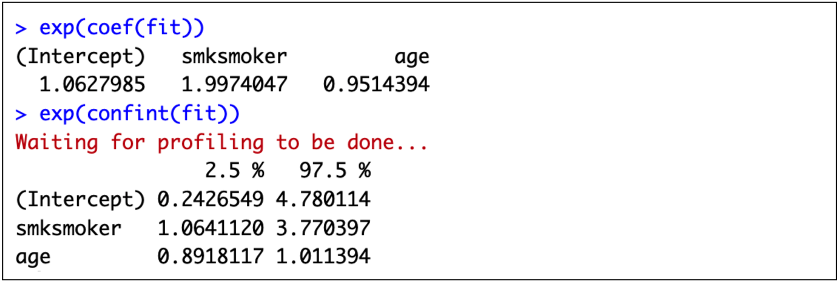
```
> fit <- glm(low.factor ~ smk + age + eth.factor,
+            data=birth, family="binomial")
> summary(fit)

Call:
glm(formula = low.factor ~ smk + age + eth.factor, family = "binomial",
    data = birth)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4211  -0.9171  -0.5687   1.3687   2.0707

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.00755    0.86166  -1.169  0.24228
smksmoker        1.10055    0.37195   2.959  0.00309 **
age             -0.03488    0.03340  -1.044  0.29634
eth.factorblack  1.01141    0.49342   2.050  0.04039 *
eth.factorother  1.05673    0.40596   2.603  0.00924 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



```
> exp(coef(fit))
    (Intercept)       smksmoker             age eth.factorblack eth.factorother 
      0.3651110       3.0058203       0.9657186       2.7494834       2.8769483 
> exp(confint(fit))
Waiting for profiling to be done...
                     2.5 %   97.5 %
(Intercept)     0.06601379 1.967972
smksmoker       1.47208358 6.378576
age             0.90303360 1.029955
eth.factorblack 1.03958814 7.308152
eth.factorother 1.31818618 6.531492
```

ETH is a confounding variable

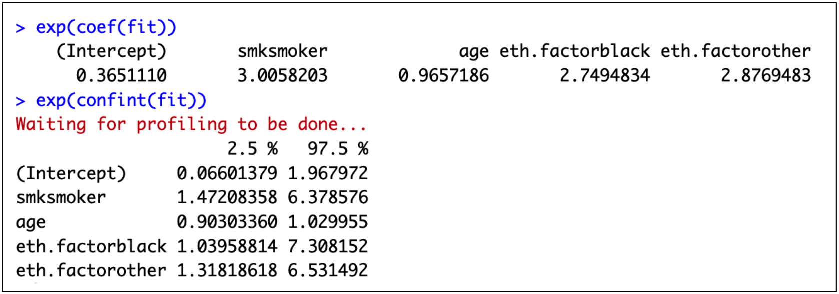
- $\log(\text{Odds}) = \beta_0 + \beta_1 \cdot \text{SMK} + \beta_2 \cdot \text{AGE}$



```
> exp(coef(fit))
(Intercept)   smksmoker         age
  1.0627985   1.9974047   0.9514394
> exp(confint(fit))
Waiting for profiling to be done...
                2.5 %   97.5 %
(Intercept) 0.2426549 4.780114
smksmoker   1.0641120 3.770397
age         0.8918117 1.011394
```

- the age-adjusted OR for SMK is 1.997…

- $\log(\text{Odds}) = \beta_0 + \beta_1 \cdot \text{SMK} + \beta_2 \cdot \text{AGE} + \beta_3 \cdot \text{Eth}(1) + \beta_4 \cdot \text{Eth}(2)$

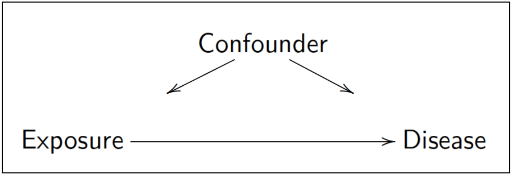


```
> exp(coef(fit))
   (Intercept)      smksmoker            age eth.factorblack eth.factorother
     0.3651110      3.0058203      0.9657186       2.7494834       2.8769483
> exp(confint(fit))
Waiting for profiling to be done...
                     2.5 %   97.5 %
(Intercept)     0.06601379 1.967972
smksmoker       1.47208358 6.378576
age             0.90303360 1.029955
eth.factorblack 1.03958814 7.308152
eth.factorother 1.31818618 6.531492
```

- ... but when we also adjust for ethnicity, it grows to 3.006!

- This phenomenon is called effect modification by a confounder.

Confounding



- Ethnicity is likely to sum up other socio-economic factors, which are here not accounted for,
- and it can therefore lead to other smoking habits, but also different birth weight.
- We should adjust for this by including ethnicity in the regression model (mostly as a proxy for other socio-economic factors).

Example 2: Confounding variable

- Folate supplementation and twin pregnancies (Vollset, Gjessing, et al, Epidemiology 2008),

	Twin birth	Single birth
Folate	329	10748
No folate	2825	162140

- Odds ratio:

$$\mathrm{OR} = \frac{329 \times 162140}{10748 \times 2825}$$

- 95% Confidence interval: $(1.57, 1.97)$

IVF treatment is a confounder

$$\frac{1}{1-p} = \beta_0 + \beta_1 \cdot \text{Folate}$$

gives OR = 1.76,

$$\frac{1}{1-p} = \beta_0 + \beta_1 \cdot \text{Folate} + \beta_2 \cdot \text{Age} + \beta_4 \cdot \text{Parity}$$

gives OR = 1.59, 95% CI (1.41,1.78).

$$\frac{1}{1-p} = \beta_0 + \beta_1 \cdot \text{Folate} + \beta_2 \cdot \text{Age} + \beta_4 \cdot \text{Parity} + \beta_5 \cdot IVF$$

gives OR = 1.04, 95% CI (.91,1.18). (The effect disappears!)

Effect modification and model misspecification

- Effect modification when adding a third variable changes the effect of exposure.
- **Confounding variables and selection effects:**
  - Confounding variables yield spurious effects if you omit them.
  - But some variables (colliders) yield spurious effects if you include them.
- This makes it difficult/impossible to do automatic model selection procedures for estimating causal effects.
- Subject matter knowledge is crucial.

COVID-19 and smoking: example of a spurious effect

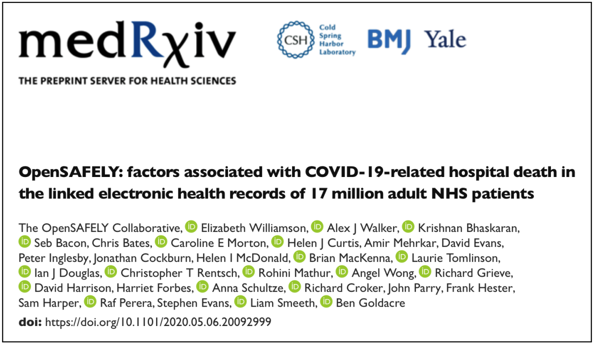
"Just overheard a woman buying cigarettes at the supermarket. She explained to the cashier that she read that smoking prevents you from a COVID-19 infection." (#epitwitter)

- In some studies, smoking seems to have a weak protective effect against COVID-19 infection/death.

- This could be explained in several ways:

1. missing confounder (e.g. age, high-exposure occupation, ...)
2. inclusion of a collider (e.g. chronic respiratory disease)
3. selection bias (see the lecture on epidemiological designs and concepts)

COVID-19 and smoking: example of a spurious effect



medRχiv

THE PREPRINT SERVER FOR HEALTH SCIENCES

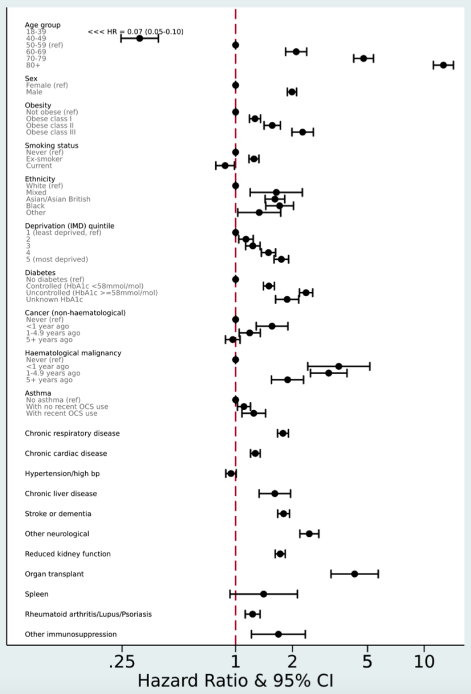
CSH Cold Spring Harbor Laboratory BMJ Yale

OpenSAFELY: factors associated with COVID-19-related hospital death in the linked electronic health records of 17 million adult NHS patients

The OpenSAFELY Collaborative, Elizabeth Williamson, Alex J Walker, Krishnan Bhaskaran, Seb Bacon, Chris Bates, Caroline E Morton, Helen J Curtis, Amir Mehrkar, David Evans, Peter Inglesby, Jonathan Cockburn, Helen I McDonald, Brian MacKenna, Laurie Tomlinson, Ian J Douglas, Christopher T Rentsch, Rohini Mathur, Angel Wong, Richard Grieve, David Harrison, Harriet Forbes, Anna Schultze, Richard Croker, John Parry, Frank Hester, Sam Harper, Raf Perera, Stephen Evans, Liam Smeeth, Ben Goldacre

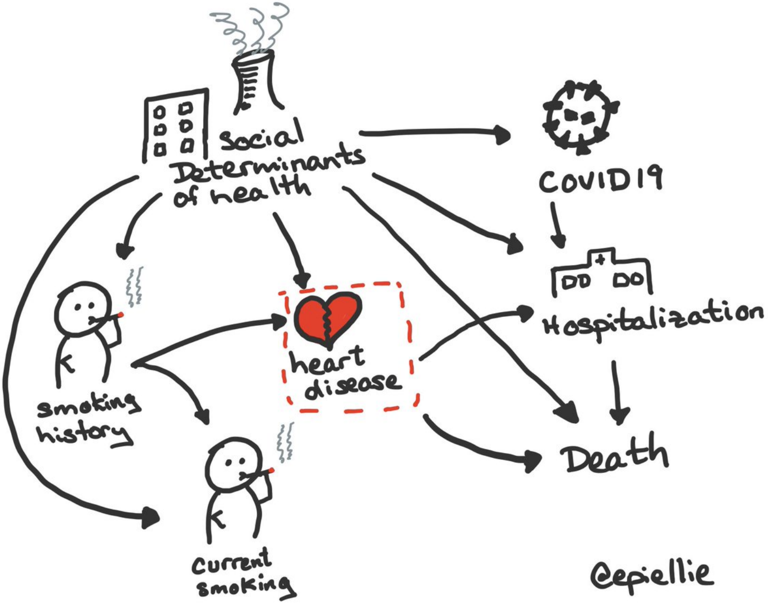
doi: https://doi.org/10.1101/2020.05.06.20092999

"[...] weak evidence of a slightly lower risk in current smokers (fully adjusted HRs 0.88, CI 0.79-0.99). In post-hoc analyses we added individual covariates to the model with age, sex and smoking to explore this further: the change in HR appeared to be largely driven by adjustment for chronic respiratory disease [...] and deprivation [...]."

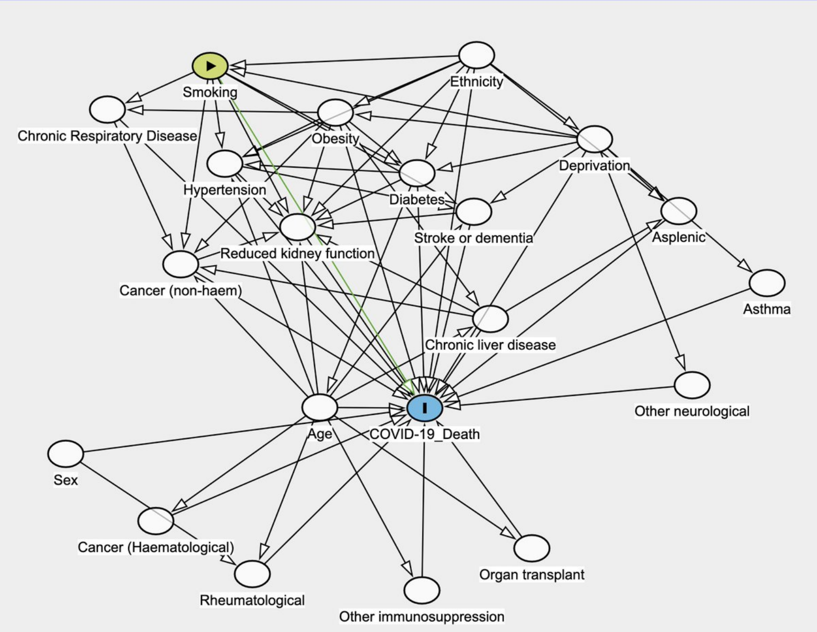


Age group
18-39
<<< HR = 0.07 (0.05-0.10)
40-49
50-59 (ref)
60-69
70-79
80+
Sex
Female (ref)
Male
Obesity
Not obese (ref)
Obese class I
Obese class II
Obese class III
Smoking status
Never (ref)
Ex-smoker
Current
Ethnicity
White (ref)
Mixed
Asian/Asian British
Black
Other
Deprivation (IMD) quintile
1 (least deprived, ref)
2
3
4
5 (most deprived)
Diabetes
No diabetes (ref)
Controlled (HbA1c <58mmol/mol)
Uncontrolled (HbA1c >=58mmol/mol)
Unknown HbA1c
Cancer (non-haematological)
Never (ref)
<1 year ago
1-4.9 years ago
5+ years ago
Haematological malignancy
Never (ref)
<1 year ago
1-4.9 years ago
5+ years ago
Asthma
No asthma (ref)
With no recent OCS use
With recent OCS use
Chronic respiratory disease
Chronic cardiac disease
Hypertension/high bp
Chronic liver disease
Stroke or dementia
Other neurological
Reduced kidney function
Organ transplant
Spleen
Rheumatoid arthritis/Lupus/Psoriasis
Other immunosuppression
.25
1
2
5
10
Hazard Ratio & 95% CI

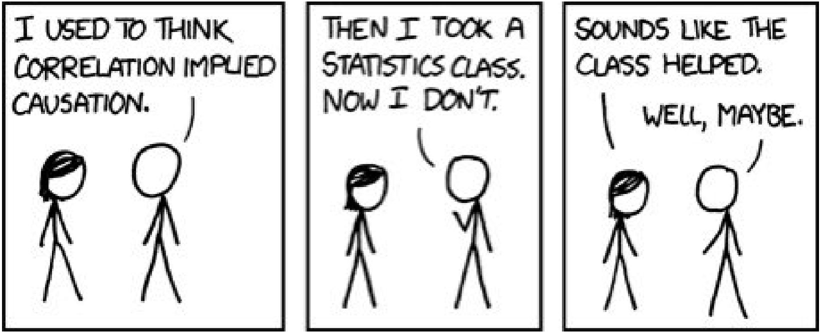
COVID-19 and smoking: example of a spurious effect



COVID-19 and smoking: example of a spurious effect



Causal inference is difficult


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Summary

Key words

- Dichotomous (binary) response variable
- Logit function
- OR, adjusted OR
- Dummy variables
- Confounders / (colliders)