SOC542 Statistical Methods in Sociology II

Multiple Regression

Thomas Davidson

Rutgers University

February 13, 2023

Plan

- Recap
- Multiple regression: An overview
- Lab: Multiple regression in R

Recap

What we have learned so far

1. Fundamentals of frequentist inference
2. Simple linear regression
3. Probability and Bayesian inference

Multiple regression

OLS assumptions review

- x and y are independently and identically distributed (IID).
  - The sample x must contain some variability. Specifically, $var(x) > 0$.
- The conditional distribution of u given x has a mean of zero.
  - Errors are independent $E[u_i|x_i] = E[u_i] = 0$.
  - Errors have constant variance $var(u_i) = \sigma^2$.
  - Errors are uncorrelated.
- If these assumptions are met, then OLS is **BLUE**
  - The **Best Linear conditionally Unbiased Estimator**

Multiple regression

Simple linear regression

- Let's say we estimate a simple linear regression of the form:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{u}$$

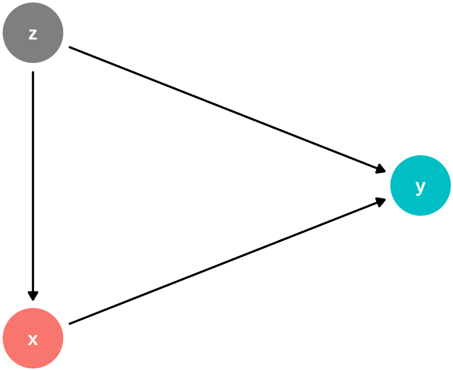
- In this case, we assume that the outcome y is a linear function of a single predictor x.
- But what if we think have reason to believe that y is also a function of other predictors?

Multiple regression

Omitted variable bias

- Omitted variable bias occurs when we leave out (or *omit*) a predictor that should be in our model.
- It exists when
  - x is correlated with the omitted variable z.
  - The omitted variable is a predictor of the dependent variable y.

Omitted variable bias

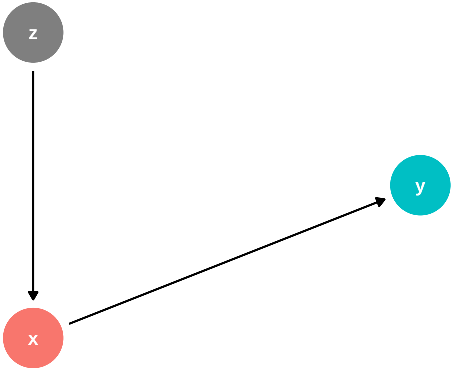


Multiple regression

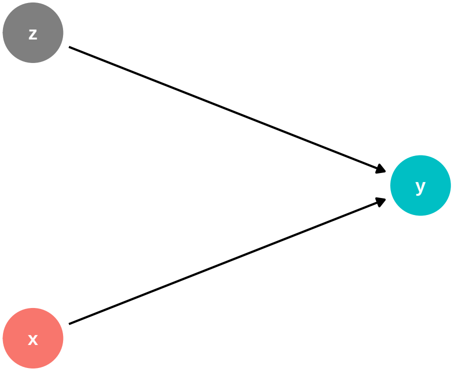
Consequences omitted variable bias

- The assumption that $E(u_i|x_i) = 0$ is violated.
  - If z is correlated with x but not included, then the error term u captures the unmeasured effect of z *and* thus u is correlated with x.
- The slope coefficient β_1 will be *biased*.
  - The mean of the sampling distribution of the OLS estimator may not equal the true effect of x.
  - $\hat{\beta}_1 = \beta_1 + bias$
- The OLS estimator is *inconsistent* as $\hat{\beta}_1$ does not converge in probability to β_1.
  - Bias remains even with large samples.
- The greater the correlation between x and u, the greater the bias.

(Not) Omitted variable bias



(Not) Omitted variable bias



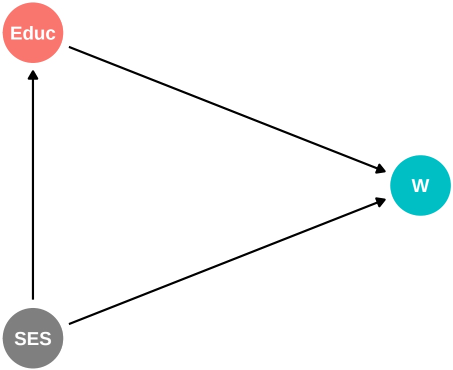
Multiple regression

Example

$$Wealth_i = \beta_0 + \beta_1 Educ + u$$

- Let's say we have a model of wealth as a function of education:
- We estimate the model and see a strong, positive relationship between education and wealth (i.e. $\hat{\beta}_1$ is positive)
- What other factors might be correlated with education *and* predict wealth?
- Education is correlated with parental socioeconomic status (SES) and predicts wealth. Our estimate of the effect of education is *biased* without taking SES into account.

Drawing the DAG



Simulating omitted variable bias

```r
N <- 100
x <- rnorm(N,2,1)
z <- 0.1*x + rnorm(N,5,2)
y <- 0.8*x + -2*z + rnorm(N, 0, 1)
```

Simulating omitted variable bias

```r
m.omit <- lm(y ~ x)
m.both <- lm(y ~ x + z)
print(m.omit$coefficients)
```

```
## (Intercept)           x
## -10.5750710   0.5640694
```

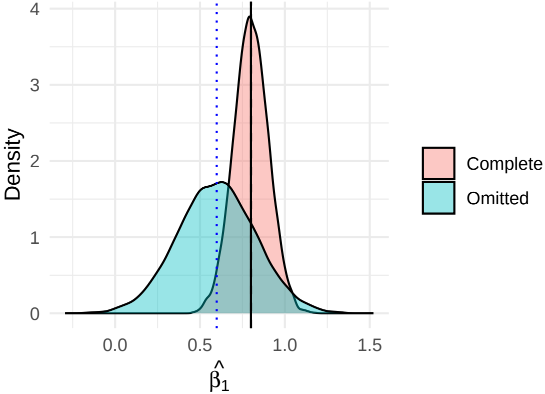
```r
print(m.both$coefficients)
```

```
## (Intercept)           x           z
##  -0.1851406   0.8079454  -2.0007816
```

Simulating omitted variable bias

```r
coefs.omitted <- c()
coefs.complete <- c()
sims <- 1E4
for (i in 1:1E4) {
    x <- rnorm(N,2,1)
    z <- 0.1*x + rnorm(N,5,1)
    y <- 0.8*x + -2*z + rnorm(N, 0, 1)
    m.omit <- lm(y ~ x)
    m.both <- lm(y ~ x + z)
    coefs.omitted[i] <- m.omit$coefficients[2]
    coefs.complete[i] <- m.both$coefficients[2]
}
```

Simulating omitted variable bias



Distribution of $\hat{\beta}_1$ over 10000 simulations.

Multiple regression

The multiple regression model

- In a multiple regression model, we specify a linear relationship between an outcome and a set of k predictors.

$$E[y|x_1, x_2, ..., x_k] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$$

Multiple regression

Independent variables and controls

- The predictors added to the model can be considered as additional **independent variables** or as **controls**.
- In general, we use the former term when we have a theoretical reason to be interested in a effect of a variable and the latter when we expect it to matter but are not interested in analyzing the relationship directly.
- We typically add control variables to address potential omitted variable bias.

Multiple regression

Interpreting coefficients

- Consider the following population model:

$$y_i = \beta_0 + \beta_1 x + \beta_2 z + u$$

- β_1 is the effect of a unit change in x *when z is held constant*.

$$\beta_1 = \frac{\Delta y}{\Delta x}, \textit{holding z constant}$$

Multiple regression

Interpreting the intercept

- Consider the same model:

$$y_i = \beta_0 + \beta_1 x + \beta_2 z + u$$

- β_0 is the expected value of y_i when $x = 0$ and $z = 0$.

Multiple regression

The OLS estimator

- The OLS estimator minimizes the sum of the squared residuals.
- Over n observations and k predictors, we minimize the following quantity:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \beta_1 x_{2i} - ... - \beta_k x_{ki})^2$$

- Thus, the predicted values are

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + ... + \hat{\beta}_k x_{ki}$$

- And the residuals are defined as $\hat{u}_i = y_i - \hat{y}_i$ for $i = 1, ..., n$.

Multiple regression

OLS in matrix form

- We can write the equation for multiple regression more compactly using matrix notation. Here X is an n by k matrix of predictors and B is vector of coefficients with length k.

$$y = \beta_0 + \beta X + u$$

- Ordinary least squares estimates can be computed directly using matrix multiplication, where X is a matrix of predictors and the first column is a vector of 1s (for the intercept) and y is the outcome.

$$\hat{\beta} = X^T X^{-1} X^T y$$

Multiple regression

OLS in matrix form

```
Intercept <- rep(1,N)
X <- cbind(Intercept,x,z)

Betas <- solve(t(X) %*% X) %*% (t(X)  %*% y)
print(t(Betas))
```

```
##      Intercept         x         z
## [1,] 0.1205118 0.6961067 -2.035357
```

```
m <- lm(y ~ x + x + z)
print(m$coefficients)
```

```
## (Intercept)           x           z
##   0.1205118   0.6961067  -2.0353567
```

`t()` is the transpose operation, `solve()` finds the inverse of a matrix, and `%*%` is the matrix multiplication operator.

Multiple regression

OLS in matrix form

- OLS estimates are derived directly from algebraic manipulation of the data.
- OLS is a special case. Other approaches we will encounter in a few weeks require more complicated *maximum likelihood estimation*.
- Bayesian regression with uniform priors will converge to the least squares solution, despite a radically different estimation procedure.

Multiple regression

Model fit and the Standard Error of the Regression

- The **Standard Error of the Regression (SER)** is an estimate of the standard deviation of the error term u_i. It captures the spread of y around the regression line.
- For a single regressor,

$$SER = \sqrt{\sigma_{\hat{u}}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} \hat{u}_i{}^2} = \sqrt{\frac{SSR}{n-2}}$$

- $n - 2$ accounts for degrees of freedom used by slope and intercept.
- A smaller SER indicates better fit.

Multiple regression

Model fit and the Standard Error of the Regression

- If we have multiple predictors we need to include an degrees of freedom adjustment k, where k is the number of predictors. The -1 accounts for the intercept.

$$SER = \sqrt{\sigma_{\hat{u}}} = \sqrt{\frac{1}{n-k-1}\sum_{i=1}^{n}{\hat{u}_i}^2} = \sqrt{\frac{SSR}{n-k-1}}$$

- The adjustment has a small effect when n is large.

Multiple regression

Model fit and R^2

- We define R^2 in the sample way as a simple regression:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{ESS}{TSS}$$

$$R^2 = 1 - \frac{SSR}{TSS}$$

Multiple regression

Adjusted R^2

- R^2 increases mechanistically as we add predictors because the SSR declines as long as $\hat{\beta}_k \neq 0$, inflating model fit.
- A degrees of freedom correction is used to adjust for this:

$$Adjusted\ R^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS}$$

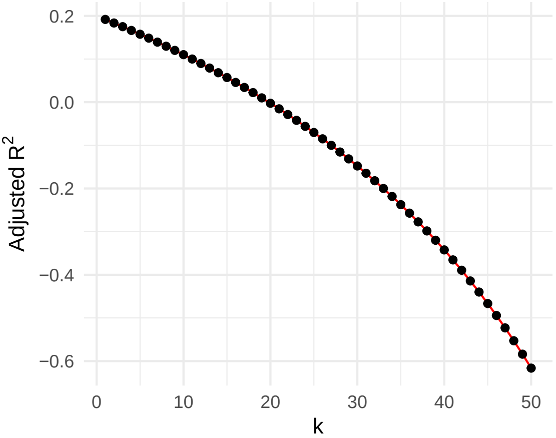
Multiple regression

Properties of adjusted R^2

- Adjusted R^2 is *always less than* R^2.
- Adding a predictor can increase Adjusted R^2, but it can decline if the change to the SSR is weaker than the offset $n - 1/n - k - 1$.
- Adjusted R^2 can be negative if the reduction in SSR does not offset $n - 1/n - k - 1$.

Multiple regression

The penalty $\frac{n-1}{n-k-1}$ increases as we add predictors



This example shows the effect of the degree of freedom adjustment, assuming $\beta_k = 0$ for all $k > 1$.

Multiple regression

Bayesian R^2

- There is no direct analogue for R^2 in Bayesian statistics
  - Recall that frequentist models assume *fixed* parameters, whereas Bayesian parameters have *distributions*.
- If we treat the Bayesian estimates as fixed, for example by taking the median of the posterior distribution $\hat{\beta}_k$, we could calculate something using the formula above, but it would not account for the *uncertainty* contained in the posterior distribution.

Multiple regression

Bayesian R^2

- Instead, we use posterior simulations to repeat the calculation across all samples from the posterior.
- Bayesian R^2 therefore has a posterior distribution.[1] We can summarize this into a single metric using the same approach as the regression coefficients, e.g. using the median of the posterior distribution.[2]

[1] "Everything that depends upon parameters has a posterior distribution" - McElreath 98.

[2] See GHV 170-171

Multiple regression

Significance tests: t-tests

- Like simple linear regression, we can interpret the statistical significance of regression coefficients using the t-statistics.
- Typically, we are interested in testing the null hypothesis that $\beta_k = 0$. We get the t-statistic by dividing a coefficient by its standard error:

$$t = \frac{\hat{\beta}_k - 0}{SE(\hat{\beta}_k)} = \frac{\hat{\beta}_k}{SE(\hat{\beta}_k)}$$

- We can the use the t-statistic to look up the relevant *p-value*.

Multiple regression

Confidence interval

- Most regression software provides a 95% confidence interval around each estimate. For β_k this would take the following form:

$$[\hat{\beta}_j - 1.96 SE(\hat{\beta}_j), \hat{\beta}_j + 1.96 SE(\hat{\beta}_j)]$$

- Recall that only 5% of the probability density of a t-distribution is greater than $|1.96|$.

Multiple regression

Joint tests

- The F-statistic is used to test a **joint hypothesis**.
- If we consider a two variable example, we might test the following *null hypothesis*:

$$H_N : \beta_1 = 0, \beta_2 = 0$$

- A joint test has q restrictions. In this case, $q = 2$.
- The *alternative hypothesis* H_A is that one or more of the q restrictions does not hold.

Multiple regression

Joint tests and the F-statistic

- Since we expect the predictors to have a *joint sampling distribution*, we cannot conduct a joint test by summarizing a series of paired tests (e.g. a t-test for every predictor) because the t-statistics are not independent.
- Instead, we must calculate the F-statistic. Where $q = 2$ it is defined as:

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2}{1 - \hat{\rho}_{t_1,t_2}^2}\right)$$

Multiple regression

Joint tests and the F-statistic

- If $\hat{\rho}_{t_1,t_2} = 0$ the equation simplifies to the average of the squared t-statistics:

$$F = \frac{1}{2}(t_1^2 + t_2^2)$$

- The p-value can then be derived from the relevant *F-distribution*, where $F \sim F_{q,\infty}$
- Typically, we use an F-test to test the restriction that $\beta_1 = 0, \beta_2 = 0, ..., \beta_k = 0$.

Multiple regression

Joint tests and the F-statistic

- If we assume the residuals are *homoskedastic*, we can test the restriction $\beta_1 = 0, \beta_2 = 0, ..., \beta_k = 0$ using the following formula:

$$F_0 = \frac{(SSR_r - SSR_u)/q}{SSR_u/(n-k+1)}$$

- The SSR_r is obtained from *restricted* model where we calculate the SSR assuming the null hypothesis is true. The SSR from the fitted model, SSR_u, is known as the *unrestricted* SSR.
- The test statistic is assessed using an *F-distribution* with q degrees of freedom and $n - k + 1$ observations.
- In most cases the homoskedasticity assumption is likely violated, so we use the more complicated formula from the previous slide, known as the *heteroskedasticity robust* F-statistic.

Multiple regression

Interpreting regression output

```
Call:
lm(formula = y ~ x + z)

Residuals:
     Min       1Q   Median       3Q      Max
-2.17211 -0.49561  0.00997  0.54232  3.09640

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.12051    0.53775   0.224    0.823
x            0.69611    0.09055   7.688 1.23e-11 ***
z           -2.03536    0.10526 -19.336  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9607 on 97 degrees of freedom
Multiple R-squared:  0.7973,    Adjusted R-squared:  0.7932
F-statistic: 190.8 on 2 and 97 DF,  p-value: < 2.2e-16
```

Multiple regression

Bayesian approaches

- "We have essentially no interest in using hypothesis tests for regression because we almost never encounter problems where it would make sense to think of the coefficients as being exactly zero" - GHV 147
- Bayesian regression is assessed by analyzing the posterior distribution of parameters to understand uncertainty.
- Nonetheless, Bayesian equivalents to t-tests and F-tests can be used if desired.[3]

[3] See Kruschke and Liddell 2018.

Multiple regression

Multicollinearity

- **Multicollinearity** occurs when a predictor x is highly correlated one or more other predictors z.
  - **Perfect multicollinearity** arises when $cor(x,z) = 1$ or \$ -1\$.
    - Usually due to some type of misspecification. e.g. accidentally including the same variable twice.
  - **Imperfect multicollinearity** means that two or more regressors are highly correlated.

Multiple regression

Multicollinearity and its implications

- Assume the following model and that x and z are highly correlated:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 z_i + u_i$$

- The variance of $\hat{\beta}_1$[4] is inversely proportional to $1 - \rho^2_{x,z}$, where $\rho_{x,z}$ is the correlation between x and z.
  - If $\rho_{x,z}$ is large, then this term is small and thus the variance is large.
- Multicollinearity *increases variance* and *reduces precision*, potentially making β_1 **non-identifiable**.

[4] The same issue also applies to $\hat{\beta}_2$

Simulating multicollinearity

```r
N <- 100
x <- rnorm(N,2,1)
x2 <- rnorm(N,0,1)
z <- 0.7*x + rnorm(N,0,1)
y <- 0.5*x + -0.5*x2 + 0.5*z + rnorm(N, 1, 1)
```

Simulating multicollinearity

```
m1 <- summary(lm(y ~ x + x2))
m2 <- summary(lm(y ~ x + x2 + z))
round(m1$coefficients,2) # omitted variable bias
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.66       0.25    6.66        0
## x               0.59       0.11    5.22        0
## x2             -0.59       0.11   -5.42        0
```

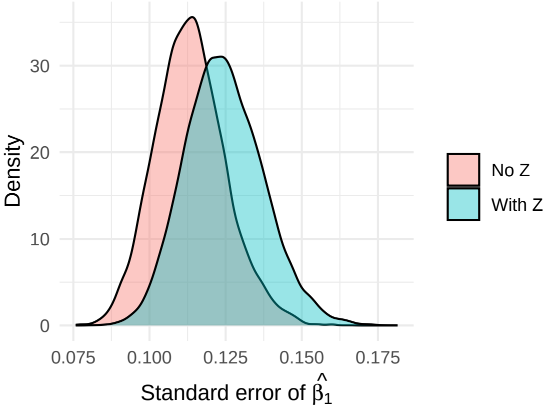
```
round(m2$coefficients,2) # multicolinearity
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.48       0.23    6.31     0.00
## x               0.29       0.13    2.28     0.02
## x2             -0.58       0.10   -5.79     0.00
## z               0.49       0.12    4.24     0.00
```

Simulating multicollinearity

```
se.omitted <- c()
se.complete <- c()
sims <- 1E4
for (i in 1:1E4) {
    x <- rnorm(N,2,1)
    x2 <- rnorm(N,0,1)
    z <- 0.7*x + rnorm(N,0,1)
    y <- 0.5*x + -0.5*x2 + 0.5*z + rnorm(N, 1, 1)
    m.omit <- summary(lm(y ~ x + x2))
    m.complete <- summary(lm(y ~ x + x2 + z))
    se.omitted[i] <- m.omit$coefficients[2,2]
    se.complete[i] <- m.complete$coefficients[2,2]
}
```

Simulating multicollinearity



Distribution of standard error over 10000 simulations.

Multiple regression

Fixing multicollinearity

- In general, multicollinearity is less severe than omitted variable bias.
  - The inflated variance will lead to more Type II errors than Type I errors.
  - Omitted variable bias can produce Type I errors, sign errors, and magnitude errors.

Multiple regression

Fixing multicollinearity

- *Solution 1*: Use more data. If we have a larger sample then we might be able to learn from additional variation in x and z.
- *Solution 2*: If we are only concerned about x then we could exclude z. But this risks omitted variable bias if z is also a predictor of y.
- *Solution 3*: Transform or combine predictors (e.g. factor analysis).

Multiple regression

Revisiting our assumptions

- $E(u_i|x_{1i}, x_{2i}, ..., x_{ki}) = 0$
- All $y_i, x_{1i}, x_{2i}, .., x_{ki}$ are IID.
- Large outliers are unlikely.
- No perfect multicollinearity.

Multiple regression

Spurious relationships and confounding

- Sometimes we observe **spurious** relationships in regression models where a correlation between two variables exists, despite the absence of any causal relationship.
  - e.g. Finding that hurricanes with female names *caused* more deaths than male named hurricanes.
- Sometimes this occur purely due to chance, but it can be due to **confounding**: a confounding variable z influences both x and y.
- Adding more predictors can often help to reduce the risk of spurious associations.
  - If we control for the confounder z, the spurious relationship between y and x disappears.

Multiple regression

Masked relationships

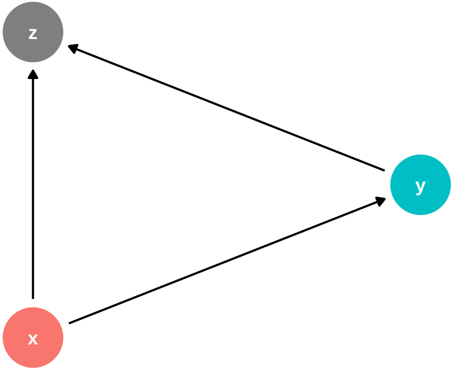
- Assume a true relationship between y and x.
- We estimate a model $y = \beta_0 + \beta_1 x$.
- The results do not show evidence of an association (i.e. $\hat{\beta}_1 \approx 0$).
- We estimate a second model including a new predictor z.
- Controlling for z allows us to observe a relationship between y and x.

Multiple regression

Colliders

- z is *caused by* y *and* x.
- In this case, z is a **collider** and controlling for z can introduce bias.
- Using DAGs can help to formalize relationships between variables and identify potential colliders.

Colliders[5]



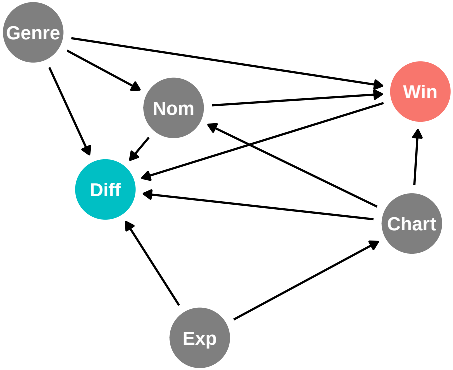
[5] See Elwert, Felix, and Christopher Winship. 2014. "Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable." *Annual Review of Sociology* 40(1):31–53. doi: 10.1146/annurev-soc-071913-043455.

Multiple regression

Variable selection

- It is often conventional practice to include a wide array of potential confounders in a regression model (“kitchen sink” or “garbage can” regressions), but this approach can cause problems!
- We must carefully consider omitted variable bias, multicollinearity, and collider bias when specifying models.
- We must use domain knowledge and theory to guide model specifications, we cannot identify these issues from the data alone.
- DAGs are a useful tool for representing our assumptions, guiding variable selection, and identifying problematic specifications.

DAGs in the wild[6]



[6] Stylized DAG based on analysis from Negro, Giacomo, Balázs Kovács, and Glenn R. Carroll. 2022. "What's Next? Artists' Music after Grammy Awards." *American Sociological Review* 87(4):644–74. doi: 10.1177/00031224221103257.

Next week

Non-linear predictors

- Dummy variables
- Categorical variables
- Non-linear transformations

Lab

- Estimating and interpreting multiple regression models