SOC542 Statistical Methods in Sociology II

Dummy, Categorical, and Non-Linear Variables

Thomas Davidson

Rutgers University

February 21, 2022

Plan

- Dummy variables
- Categorical variables
- Logarithms
- Polynomials

Dummy variables

Definitions

- A **dummy variable** (or **indicator variable** is used to measure the difference between two possible states.
- Dummy variables are binary, taking a value of either zero or one.
- These values stand in for social categories of interest.
  - e.g. Male/female, employed/unemployed, vaccinated/unvaccinated.

Dummy variables

Dummy variables as random variables

- We can generate dummy variables using the Bernoulli distribution, where $P(x = 1) = p$ and $P(x = 0) = 1 - p$.

$$x \sim Bernoulli(p)$$

- The Bernoulli distribution is a special case the Binomial distribution:

$$Bernoulli(p) = Binomial(1, p)$$

Dummy variables

A simple model

```
N <- 10000
x <- rbinom(N, 1, .5) # p = .5
y <- 3*x + rnorm(N, 10, 1)
m <- lm(y ~ x)
round(m$coefficients,2)
```

```
## (Intercept)           x
##        9.98        3.03
```

Dummy variables

Interpretation

```
as.data.frame(cbind(x,y)) %>% group_by(x) %>% summarize(mean = mean(y))
```

```
##      x       mean
## [1,] 0  9.979164
## [2,] 1 13.012610
```

Dummy variables

Interpretation

- The coefficient represents the expected difference in the outcome when $x = 1$ compared to $x = 0$
- Consider the following population model, predicting income as a function of union membership.

$$Income = \beta_0 + \beta_1 Union + u$$

- β_1 represents the expected difference in income for union members compared to non-union members.
- The dummy variable also impacts the meaning of the intercept β_0 is the average income for non-unionized workers.

Dummy variables

Example: Union returns

```
gss <- haven::read_dta("../labs/lab-data/GSS2018.dta")
U <- gss %>% select(realrinc, union, sex) %>%
    drop_na(realrinc, union) %>%
    mutate(union_dummy = ifelse(union == 1, 1, 0))
u.reg <- lm(realrinc ~ union_dummy, data = U)
print(u.reg)
```

```
##
## Call:
## lm(formula = realrinc ~ union_dummy, data = U)
##
## Coefficients:
## (Intercept)  union_dummy
##       24572         5646
```

Dummy variables

Example: Union returns

```
means <- U %>% group_by(union_dummy) %>% summarize(m = mean(realrinc))
print(means)
```

```
## # A tibble: 2 x 2
##   union_dummy      m
##         <dbl>  <dbl>
## 1           0 24572.
## 2           1 30218.
```

```
mean.nonunionized <- means %>% filter(union_dummy == 0)
print(mean.nonunionized$m + u.reg$coefficients[2])
```

```
## union_dummy
##     30218.2
```

Dummy variables

Reference category

- The interpretation of the model depends on which value we assign to 1 or 0.
  - The value assigned to 0 is known as the **reference category**.
- For a statistical perspective, the choice is arbitrary.
- But often there are theoretical reasons for selecting a certain reference category and the choices encode certain assumptions about the social world.[1]

[1] See Johfre and Freese (2021) for further discussion of reference categories.

Dummy variables

Reversing the reference category

```
x.rev <- ifelse(x, 0, 1)
m2 <- lm(y ~ x.rev)
round(m2$coefficients,2)
```

```
## (Intercept)       x.rev
##       13.01       -3.03
```

Dummy variables

Multiple dummy variables

- A multiple regression model can include more than one dummy variable.
- The interpretation of each coefficient is now the difference *holding other variables at their means*.
- The intercept is the mean value of the outcome when all dummy variables are zero.

Dummy variables

Multiple dummy variables

This model of union returns includes a dummy variable for sex. What is the reference category and how can we interpret the intercept?

```
u.reg2 <- lm(realrinc ~ union_dummy + sex, data = U)
print(u.reg2)
```

```
##
## Call:
## lm(formula = realrinc ~ union_dummy + sex, data = U)
##
## Coefficients:
## (Intercept)  union_dummy          sex
##       39592         3699        -9757
```

Dummy variables

Model specification and priors

- McElreath discusses an alternative method of specifying dummy variables called **index variables**.
- Consider the earlier model, where β_0 represents the average income for non-unionized workers.

$$Income = \beta_0 + \beta_1 Union + u$$

- Consider β_0. Do we expect the posterior distribution to be the same if we reverse the reference category? What does this imply about the prior on β_0?

Dummy variables

Model specification and priors

- There are two problems that arise when trying to determine the prior for β_0:
  1. β_0 is the average income for the reference group. It is not clear what a reasonable expection is.
  2. We assume more uncertainty about one of the groups:
     - The predicted income of non-unionized workers requires a single parameter, β_0, whereas the predicted income of unionized workers requires two parameter, β_0 and β_1.
- In practice, these issues tend to "wash out" if we have a lot of data.

Dummy variables

Model specification and priors

- McElreath recommends estimating the following model without an intercept:

$$Income = \beta_1^* Union + \beta_2^* NonUnion + u$$

- Here the $*$ denotes that $\beta_1^* \neq \beta_1$ in the previous model.
- β_1^* is the average income for unionized workers and β_2^* is the average income for non-unionized workers.
  - The prior on either coefficient is now more straightforward to define.

Dummy variables

Model specification and priors

- We can get the typical dummy estimate directly from this model:

$$\beta_1 = \beta_1^* - \beta_2^*$$

- In a Bayesian regression, the posterior distribution for β_1 is obtained by taking the difference between the posterior distributions (McElreath refers to this as a *contrast*).

Dummy variables

Model specification and priors

```
library(rstanarm)
U$union_dummy <- as.factor(U$union_dummy)
sm1 <- stan_glm(realrinc ~ union_dummy,  data = U,
                family = "gaussian", chains =1 , refresh = 0)
sm2 <- stan_glm(realrinc ~ 0 + union_dummy,  data = U,
                family = "gaussian", chains =1 , refresh = 0)
sm1$coefficients
```

```
##  (Intercept) union_dummy1
##    24560.387     5800.593
```

```
sm2$coefficients
```

```
## union_dummy0 union_dummy1
##     24608.93     30206.70
```

Note: `as.factor()` is used to specify that z should be treated as a variable composed of two discrete categories.

Dummy variables

Recovering the difference

```
library(tidybayes)
posterior <- sm2 %>% spread_draws(union_dummy0, union_dummy1)
posterior$contrast <- posterior$union_dummy1 - posterior$union_dummy0
head(posterior)
```

```
## # A tibble: 6 x 6
##   .chain .iteration .draw union_dummy0 union_dummy1 contrast
##    <int>      <int> <int>         <dbl>         <dbl>    <dbl>
## 1      1          1     1        24596.        29082.    4487.
## 2      1          2     2        25179.        35116.    9937.
## 3      1          3     3        24863.        34796.    9934.
## 4      1          4     4        25352.        26761.    1409.
## 5      1          5     5        25418.        29932.    4514.
## 6      1          6     6        24823.        30519.    5696.
```

```
print(median(posterior$contrast))
```

```
## [1] 5634.938
```

Dummy variables

Recovering the difference

We can do the same in frequentist models, but it is unnecessary since there are no priors and additional calculations are required to test whether the difference is statistically significant.

```
m.d <- lm(realrinc ~ union_dummy, data = U)
print(coefficients(m.d)[2]) # dummy coefficient
```

```
## union_dummy1
##     5645.907
```

```
m.i <- lm(realrinc ~ 0 + union_dummy, data = U)
diff <- coefficients(m.i)[2] - coefficients(m.i)[1]
print(diff) # diff
```

```
## union_dummy1
##     5645.907
```

Categorical variables

More than two categories

- **Categorical variables** are a generalization of dummy variables to more than two categories.
  - e.g. Race/ethnicity, highest level of education, region.
- Categories can be **ordinal**, indicating some type of numerical ranking, or **nominal**.

Categorical variables

Categorical variables as dummy variables

```
cats <- sample(c("a", "b", "c"), 10000, replace=TRUE,
               prob=c(0.2, 0.2, 0.6))
a <- ifelse(cats == "a", 1,0)
b <- ifelse(cats == "b", 1,0)
c <- ifelse(cats == "c", 1,0)
y <- 0.3*b + rnorm(N)
```

Categorical variables

Reference categories and regression results

The only difference between these models is the order. By default, the last value will be used as the reference category.

```
m4 <- lm(y ~ a + b + c)
m5 <- lm(y ~ c + a + b)
m6 <- lm(y ~ b + c + a)
```

Categorical variables

Reference categories and regression results

	Model 1	Model 2	Model 3
(Intercept)	0.004	0.294***	-0.008
	(0.013)	(0.022)	(0.022)
a	-0.011	-0.302***	
	(0.026)	(0.032)	
b	0.291***		0.302***
	(0.026)		(0.032)
c		-0.291***	0.011
		(0.026)	(0.026)
Num.Obs.	10000	10000	10000
R2	0.014	0.014	0.014
R2 Adj.	0.013	0.013	0.013

$+$ p $<$ 0.1, * p $<$ 0.05, ** p $<$ 0.01, *** p $<$ 0.001

Categorical variables

Interpretation

- Let's assume we run a survey in North America and find out respondents' country of residence. We want to estimate the following model, using USA as a reference category:

$$Income = \beta_0 + \beta_1 Canada + \beta_2 Mexico + u$$

- β_0 represents the average income for respondents in the USA.
- β_1 represents the expected difference between Canada and the USA.
- β_2 represents the expected difference between Mexico and the USA.

Categorical variables

Degrees of freedom

- Each additional category uses up a degree of freedom in our model.
  - Be careful when including variables with many categories (e.g. state of residence)
- Sometimes it may be defensible to treat *ordinal* variables as if they are continuous.
  - This only uses one degree of freedom.
  - Unit increases should be constant for all values and have a linear interpretation.
    - e.g. We have categories a, b, and c we could translate these values to the sequence 1, 2, 3. In this case, $b - a = c - b$.

Categorical variables

Encoding and categorical variables

- Categorical data in surveys is often coded as numeric. This can lead to misleading results since we might consider *nominal* categories as *ordinal*. Consider the example below using the region variable in the GSS.

```
print(lm(realrinc ~ region, data = gss))
```

```
##
## Call:
## lm(formula = realrinc ~ region, data = gss)
##
## Coefficients:
## (Intercept)       region
##    25217.31       -42.71
```

Categorical variables

Encoding and categorical variables

- We can fix this by casting the variable as a factor.

```
gss$region <- as.factor(gss$region)
print(lm(realrinc ~ region, data = gss))
```

```
##
## Call:
## lm(formula = realrinc ~ region, data = gss)
##
## Coefficients:
## (Intercept)      region2      region3      region4
##       29666        -4397        -5512        -7056
##     region7      region8      region9
##       -3835        -6179        -3194
```

Categorical variables

Fixed effects

- Categorical variables are sometimes considered as **fixed effects**.
- Roughly speaking, fixed effects are used like controls soak up variance in a model in order to compare across units.

$$Income = \beta_0 + \beta_1 Female + \gamma State + u$$

- Here γ is a $50 - 1$ length vector of coefficients, representing the difference in income between each state and a reference state.
- β_0 is the average income for men in the reference state.
- β_1 can therefore be interpreted as the expected difference in income between males and females, net of differences between states.

Logarithms

Logarithms and the exponential function

- The **exponential function** raises a *base* b is raised to a power x.[2]

$$exp(x) = b^x$$

- The **logarithm** is the *inverse* of exponentiation.

$$log_b(b^x) = x$$

[2] Note how this differs from the power function, where we raise x to a specified power. E.g. $power(x, 2) = x^2$

Logarithms

Logarithms and the exponential function

- Here are some examples of common bases:

$$log_2(2^4) = 4$$

$$log_{10}(10^4) = 4$$

- The **natural logarithm** uses the constant $e \approx 2.71828$ as its base:

$$log_e(e^4) = 4$$

Logarithms

Logarithms and exponents

- We can easily verify this in R using the `log` function with specified bases:

```
log(2^4, base = 2)
```

```
## [1] 4
```

```
log(10^4, base = 10)
```

```
## [1] 4
```

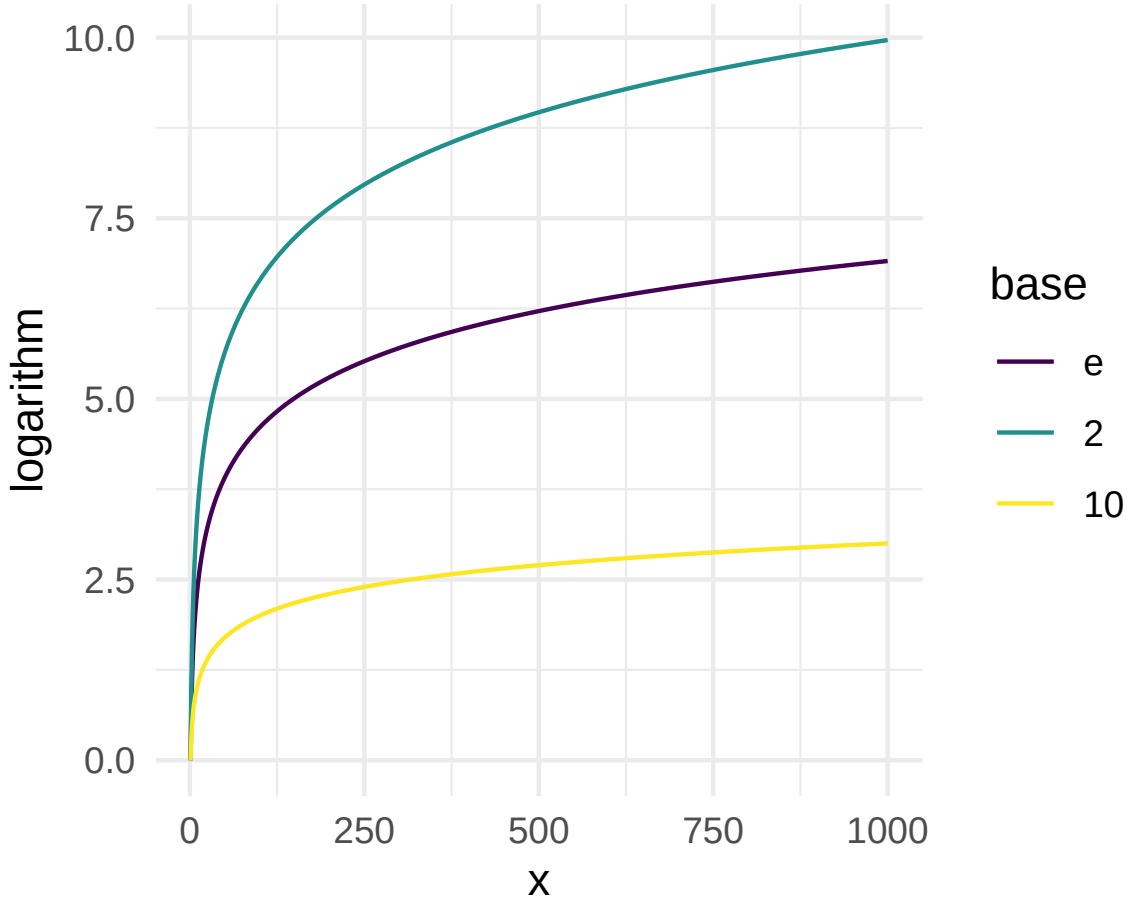
```
log(exp(1)^4)
```

```
## [1] 4
```

- The default base is e. Thus, `exp(1)` $= e^1 = e$.

Logarithms

Graphing logarithms



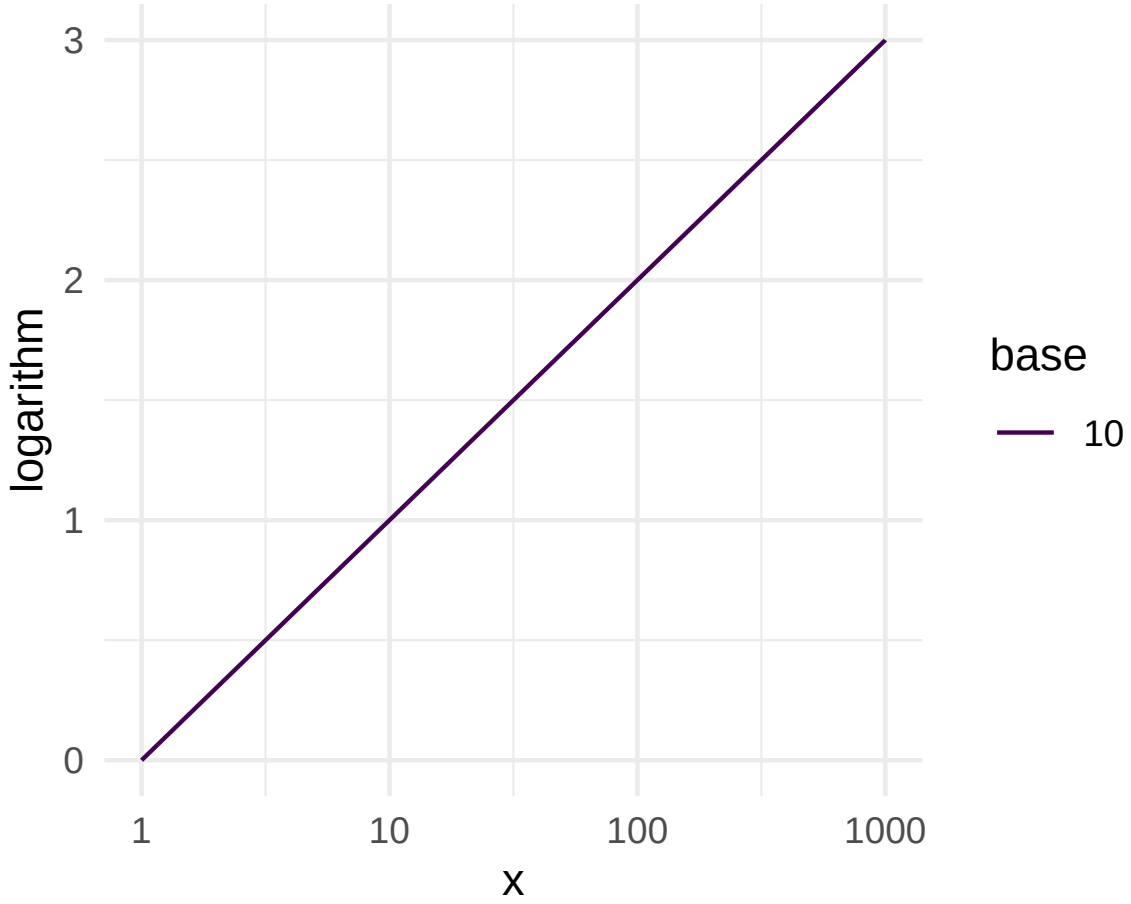
Logarithms

Why use logarithms?

- Interpretation
  - Logarithms allow us to transform variables to measure differences in *magnitude* and sometimes *percentages*
- Specification
  - Logarithms can induce normality for if a variable has a **log normal** distribution.

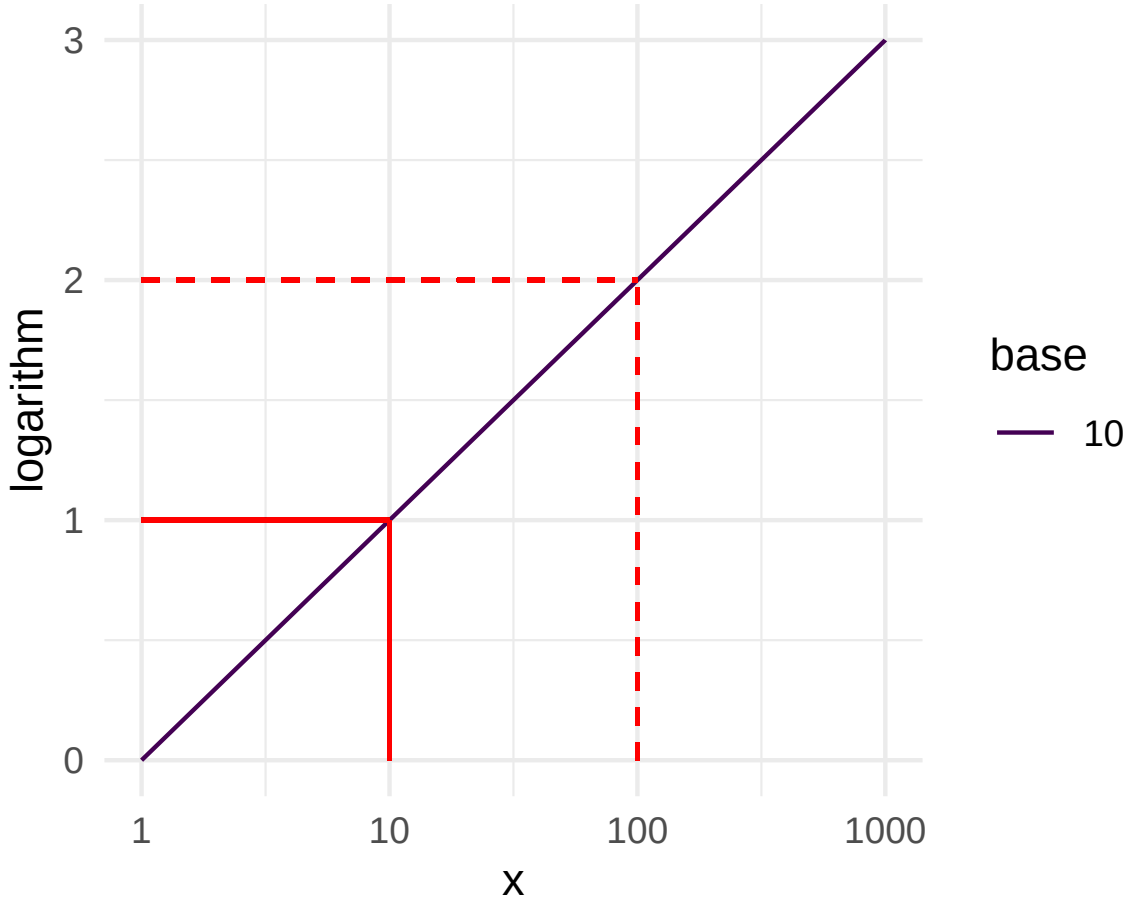
Logarithms

A unit increase of log_{10}



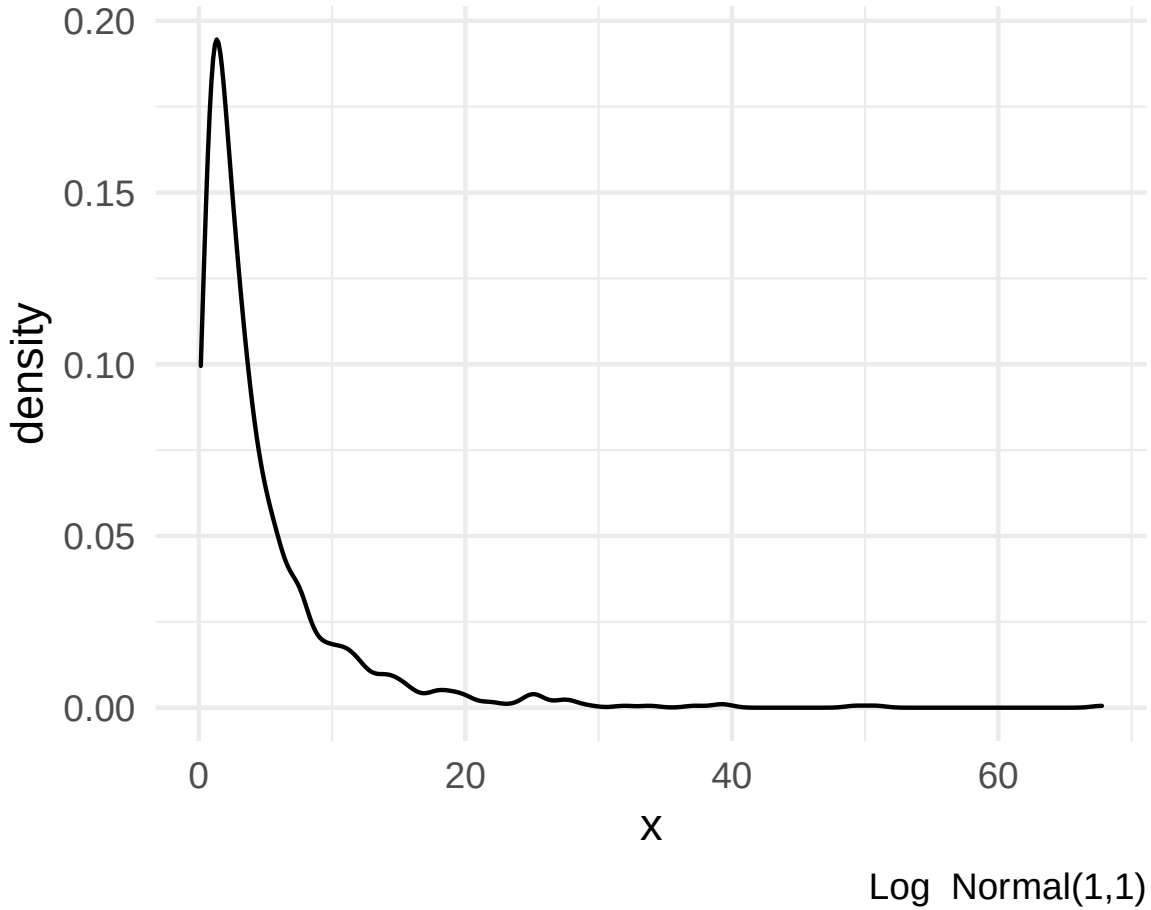
Logarithms

A unit increase of log_{10}



Logarithms

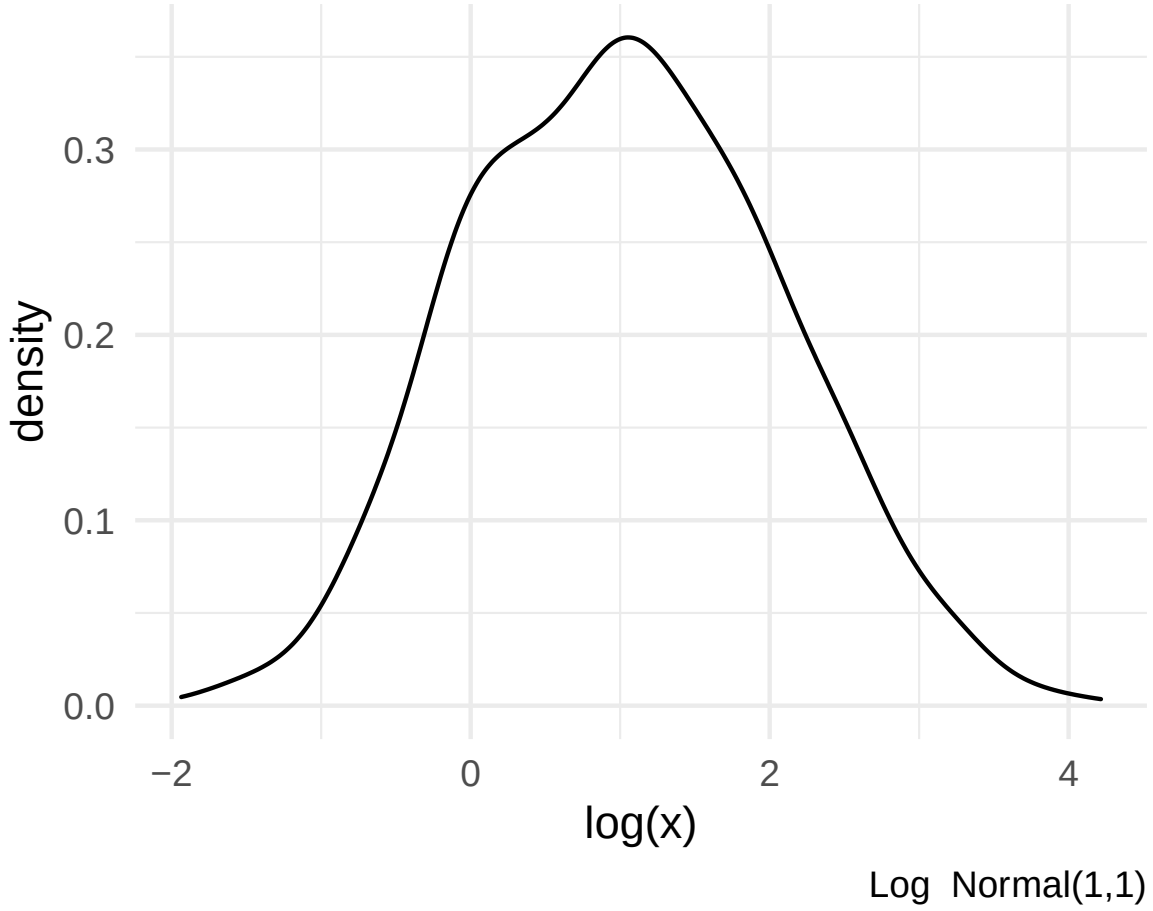
The log-normal distribution



Draws generated using `rlnorm()`.

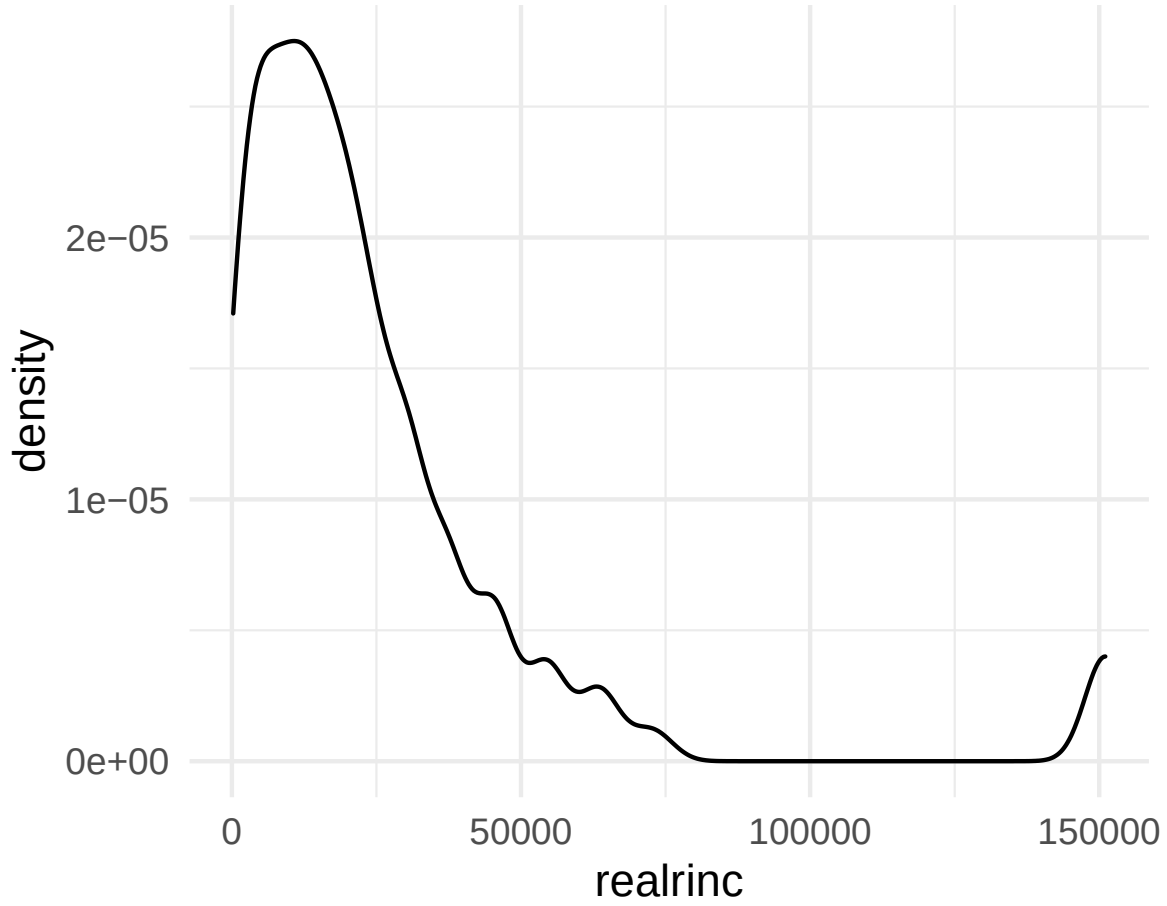
Logarithms

The natural logarithm of the log normal distribution



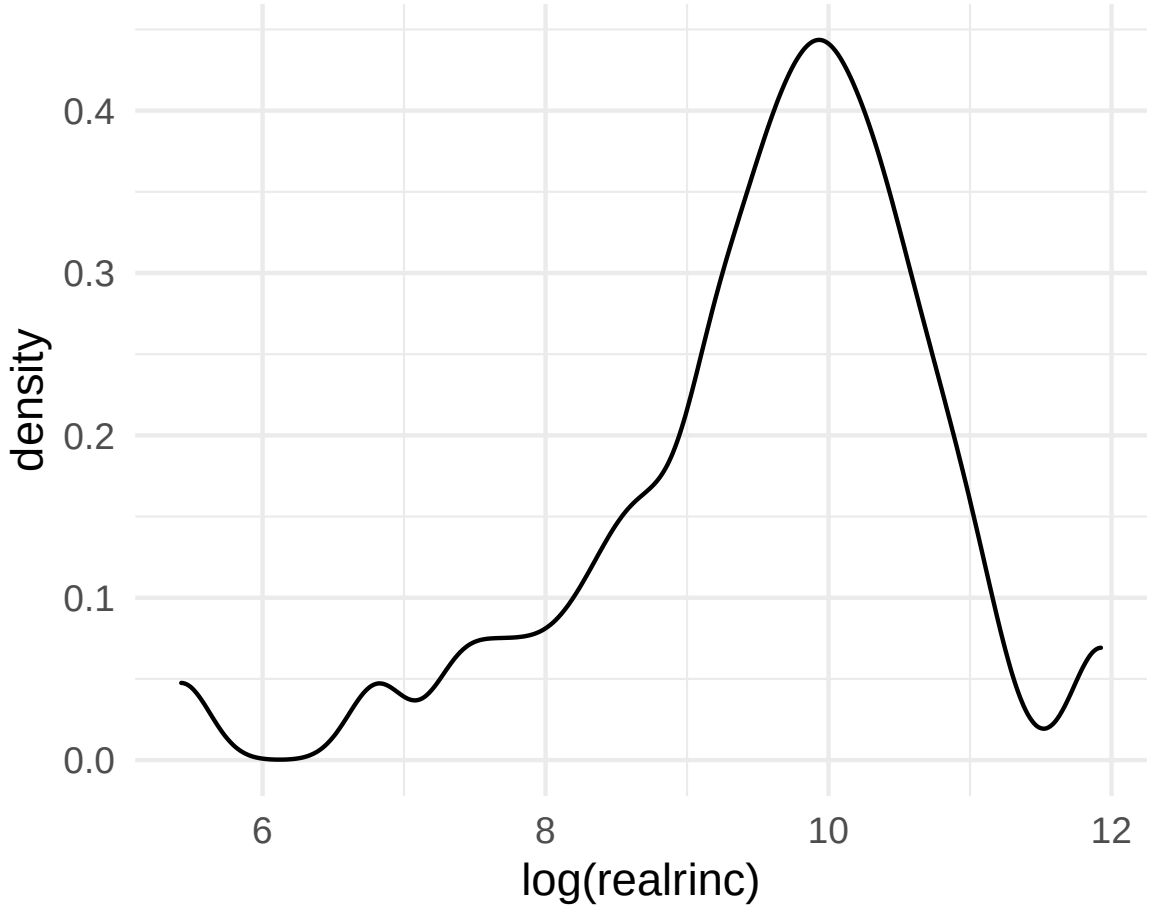
Logarithms

The distribution of 2018 GSS respondent income (`realinc`)



Logarithms

The distribution of 2018 GSS respondent income (`realinc`), natural log



Logarithms

Logarithms and regression

- Logarithms of predictors (**Linear-log models**)
  - If we include $log_e(x)$ as a predictor, β_i now represents the expected change associated with a unit increase in $log_e(x)$

Logarithms

Logarithms as predictors

Here β_1 represents the effect of a one-unit increase in height *on a logarithmic scale*.

```
##
## Call:
## lm(formula = realrinc ~ log(height), data = gss)
##
## Coefficients:
## (Intercept)  log(height)
##     -348333        88987
```

Logarithms

Logarithms and regression

- Logarithms of outcomes (**Log-linear models**)
  - If the outcome is $log_e(y)$. For any variable x, β_i represents the expected change in $log_e(y)$ associated with a unit change in x.

Logarithms

Logarithms as outcomes

Here β_1 represents the effect of a one-unit increase in height on the *logarithm* of income. Note the difference in the coefficient compared to the previous model.

```
##
## Call:
## lm(formula = log(realrinc) ~ height, data = gss)
##
## Coefficients:
## (Intercept)       height
##     5.60849      0.06041
```

Logarithms

Logarithms and regression

- Logarithms of predictors *and* outcomes (**Log-log models**)
  - If both x and y are entered into the model as logarithms, β_i represents the expected change in $log_e(y)$ associated with a unit change in $log_e(x)$
  - Equivalently, this corresponds to the expected percentage change in y as a result of a 1% change in x. Hence, such coefficients can be interpreted as **elasticities**.

Logarithms

Log-log models

Here β_1 represents the effect of a one-unit increase in *logarithm* of height on the *logarithm* of income. A 1% increase in height is associated with a 4% increase in income.

```
##
## Call:
## lm(formula = log(realrinc) ~ log(height), data = gss)
##
## Coefficients:
## (Intercept)  log(height)
##      -7.341        4.044
```

Logarithms

Log-log models

We can still incorporate untransformed variables into these models. Here the coefficient for `sex` can be interpreted as the *difference in the expected logarithm of income* between male and female respondents.

```
##
## Call:
## lm(formula = log(realrinc) ~ log(height) + sex, data = gss)
##
## Coefficients:
## (Intercept)  log(height)          sex
##     -0.4834       2.5136      -0.2774
```

Logarithms

Model comparison

The model fit statistics do not provide a meaningful comparison between Model 1 and the other models since the outcomes are different.

	Model 1	Model 2	Model 3	Model 4
(Intercept)	-348333.267***	5.608***	-7.341***	-0.483
	(55342.266)	(0.520)	(2.181)	(3.046)
log(height)	88986.528***		4.044***	2.514***
	(13158.504)		(0.519)	(0.703)
height		0.060***		
		(0.008)		
sex				-0.277**
				(0.086)
Num.Obs.	1194	1194	1194	1194
R2	0.037	0.049	0.049	0.057
R2 Adj.	0.036	0.048	0.048	0.055

Polynomials

Definitions

- **Polynomial** regression expresses non-linear relationships between continuous predictors in a linear model by adding exponents of x.
- The expected value of y is expressed as an k^{th} degree polynomial:

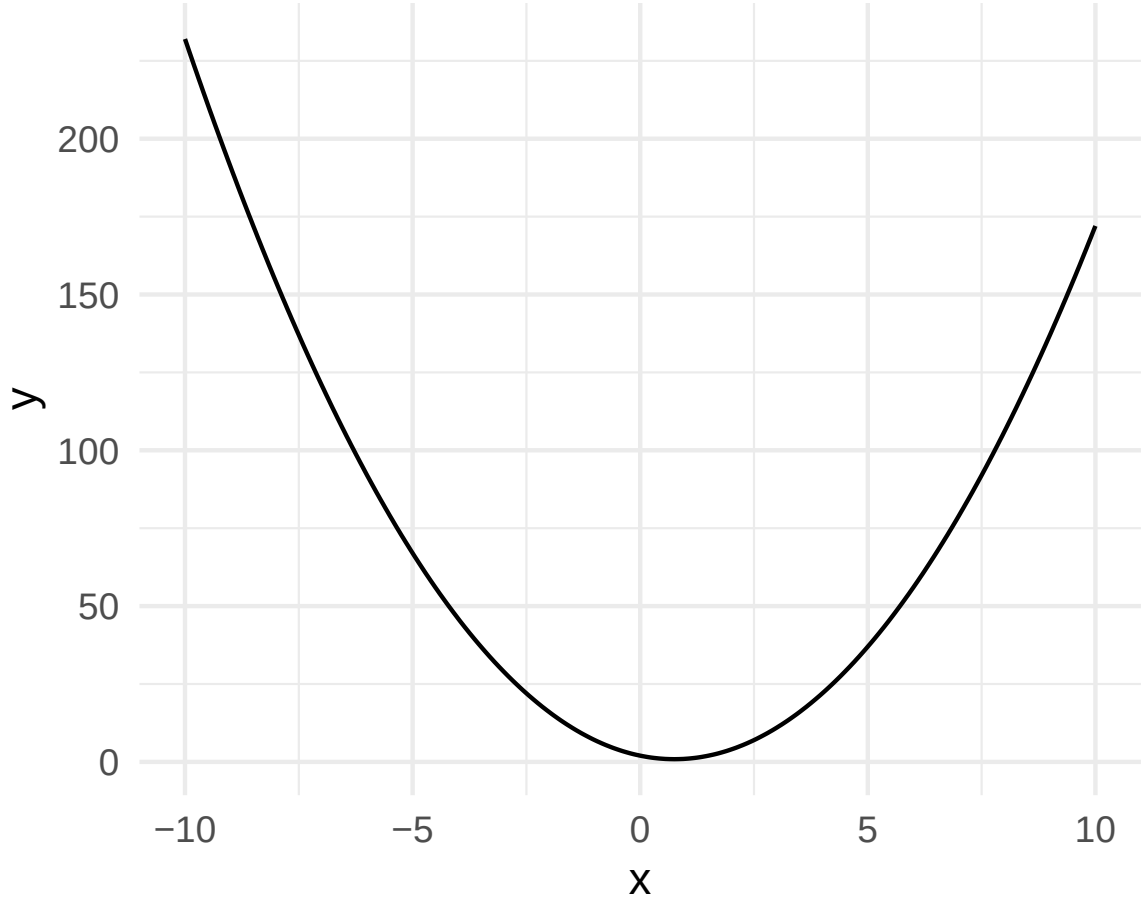
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + ... + \beta_k x^k + u$$

- Generally, we use a restricted form, such as the quadratic model:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$$

Polynomials

Quadratic functions and parabolas



$$y = 2 + -3x + 2x^2.$$

Polynomials

When to use polynomial regression

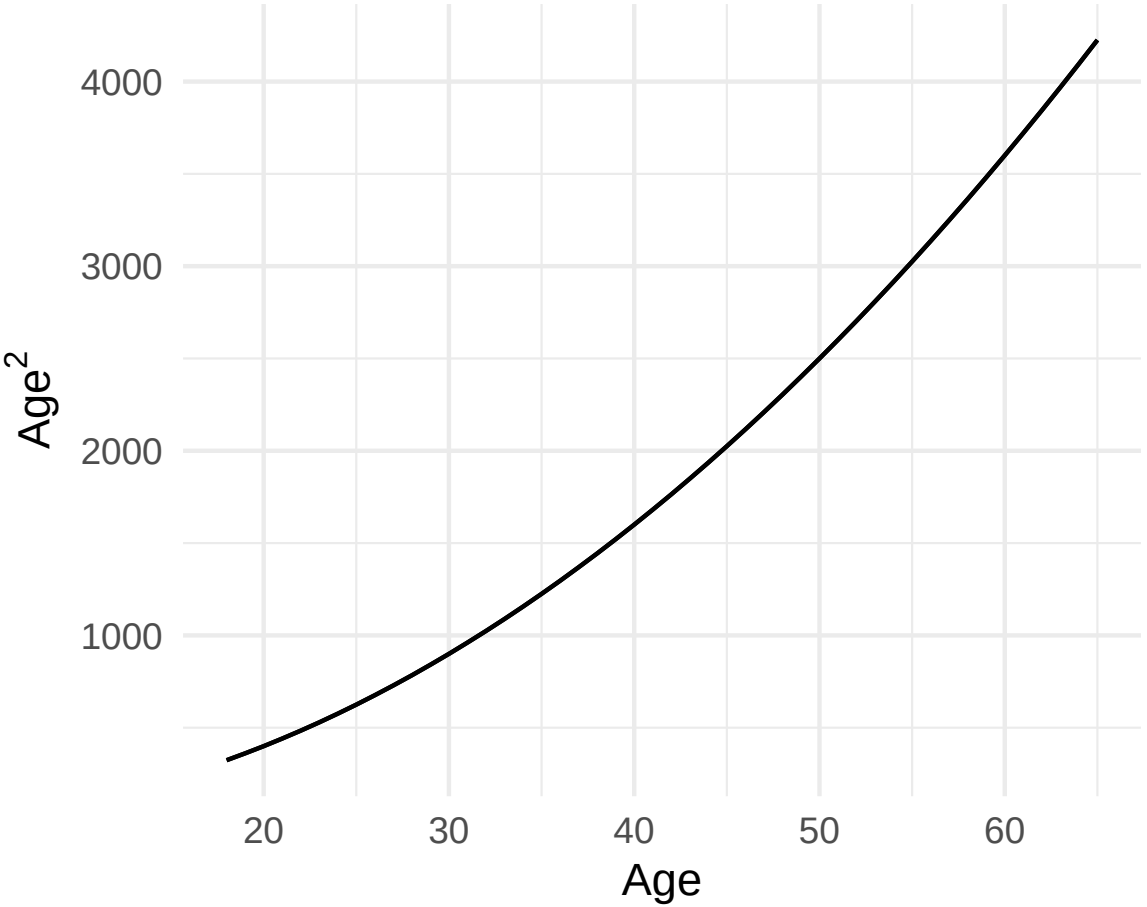
- We add polynomials to capture non-linear relationships. For example, we might expect a non-linear relationship between age and income. We could express this using the following model:

$$Income = \beta_0 + \beta_1 Age + \beta_2 Age^2 + u$$

- The effect of age is now decomposed into two coefficients:
  - β_1 captures the linear relationship between age and income.
  - β_2 captures a non-linear association between age and income.
- The coefficients do not have a simple interpretation.
  - Consider whether we can change *Age* while holding Age^2 constant.

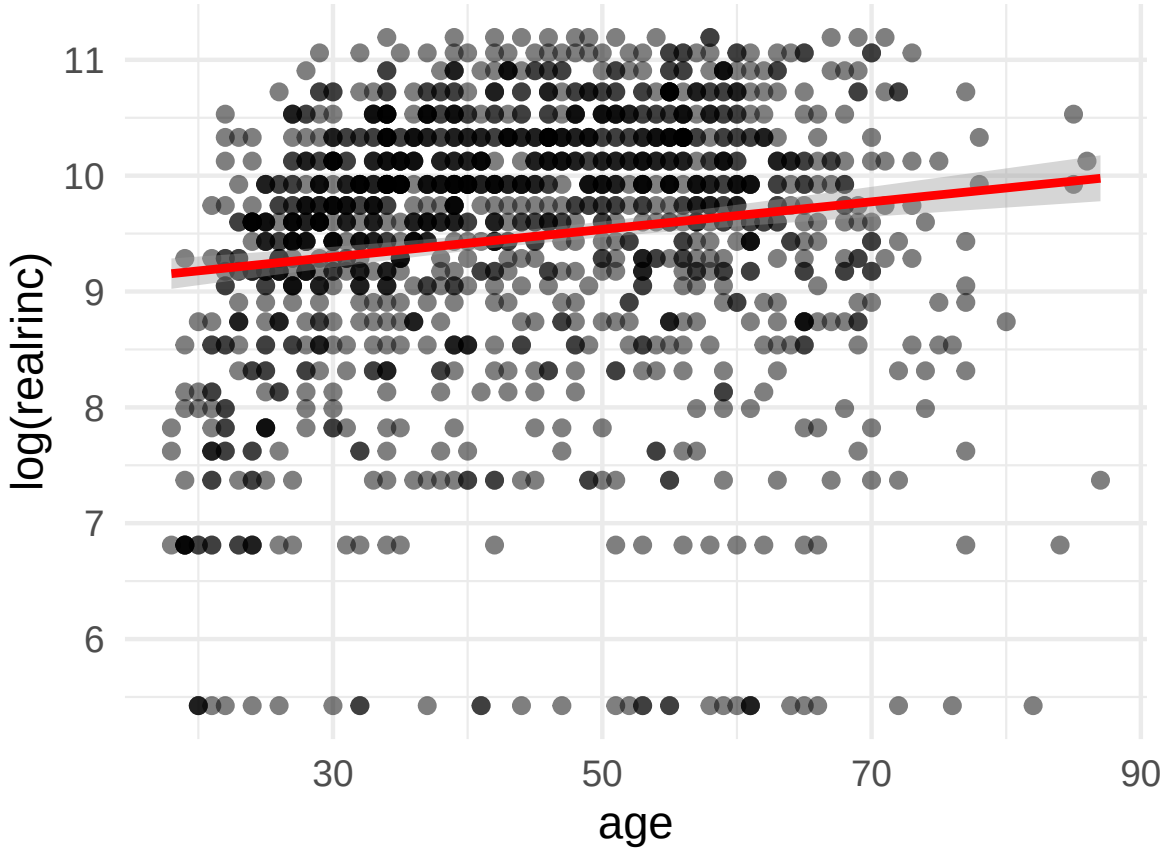
Polynomials

Age and Age-Squared



Polynomials

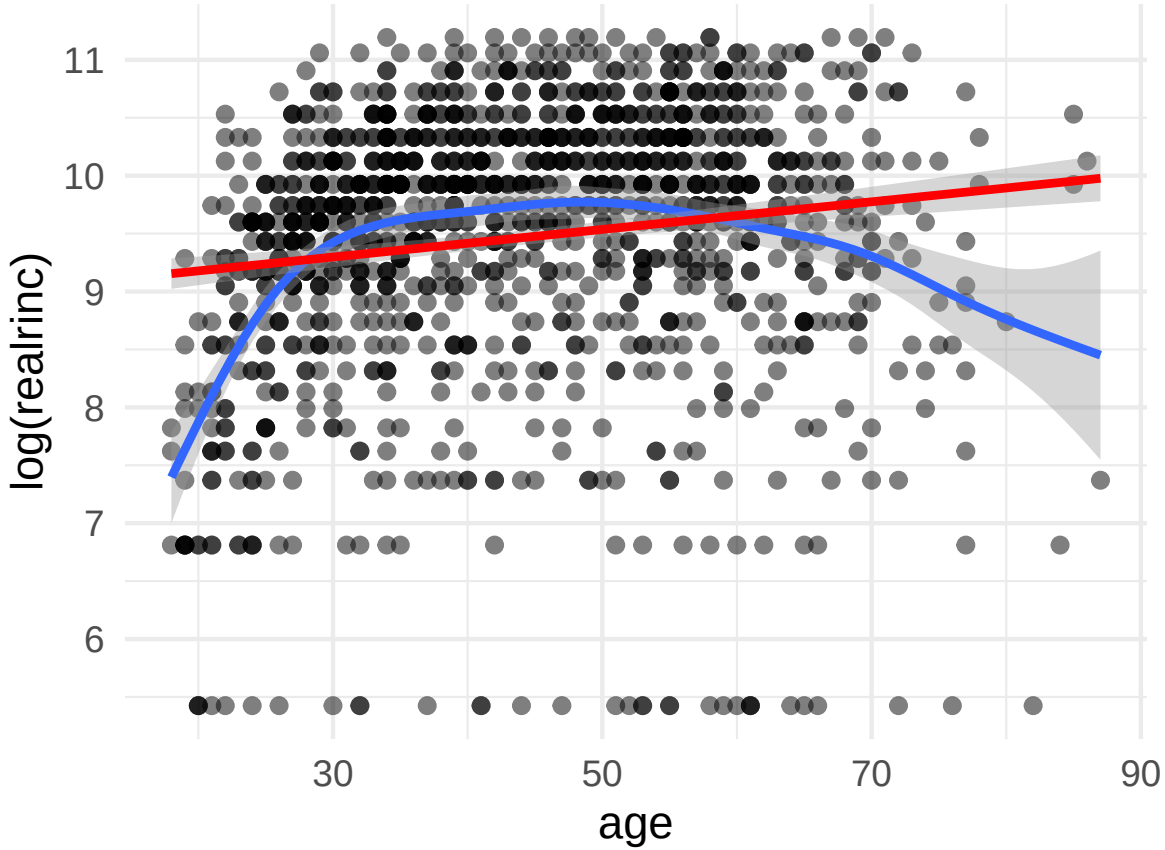
Income and age



Age < 89 and income < 1E5. OLS fitted line.

Polynomials

Income and age



Age < 89 and income < 1E5. OLS & LOESS fitted lines.

Polynomials

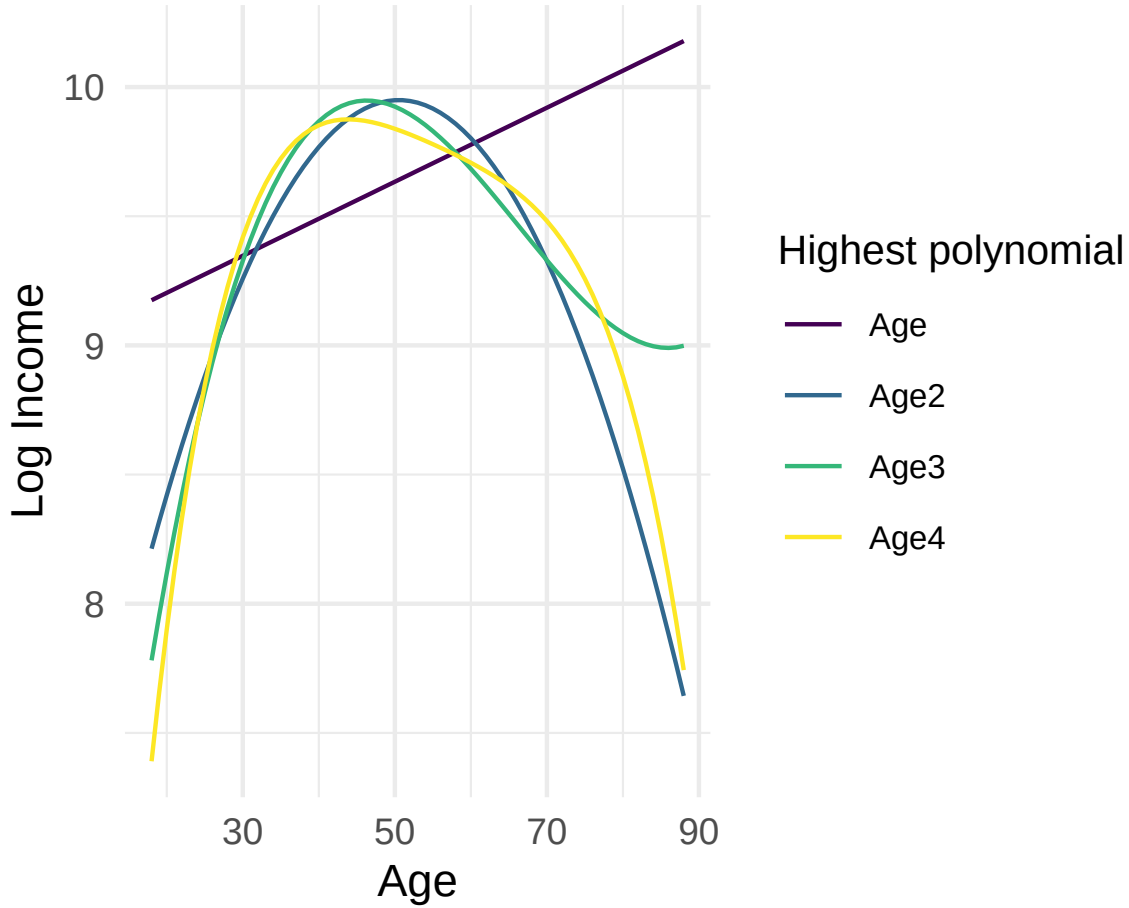
Income and age

	Model 1	Model 2	Model 3	Model 4
(Intercept)	8.917***	5.759***	2.958***	-2.453
	(0.108)	(0.294)	(0.764)	(1.964)
age	0.014***	0.166***	0.368***	0.894***
	(0.002)	(0.013)	(0.053)	(0.184)
age2		-0.002***	-0.006***	-0.024***
		(0.000)	(0.001)	(0.006)
age3			0.000***	0.000***
			(0.000)	(0.000)
age4				0.000**
				(0.000)
Num.Obs.	1357	1357	1357	1357
R2	0.027	0.114	0.124	0.130
R2 Adj.	0.027	0.112	0.122	0.127

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Polynomials

Predictions



Predicted values from OLS models. Income measured using realrinc. Respondents under 89 only.

Polynomials

When to use polynomial regression

- A second-order polynomial (e.g. Age^2) is sufficient to capture non-linearity.
  - In this case, there is evidence of a curvilinear relationship between age and income.[3]
- Higher-order polynomial terms can improve model fit and capture more complex non-linearities *use up degrees of freedom*.

[3] For an example of polynomials used in a different context, see Dokshin, Fedor A. 2016. "Whose Backyard and What's at Issue? Spatial and Ideological Dynamics of Local Opposition to Fracking in New York State, 2010 to 2013." *American Sociological Review* 81 (5): 921–48.

Next week

Interaction terms

- What is an interaction?
- How to specify an interaction
- How to interpret an interaction