Return to linear regression

ISTA 410 / INFO 510: Bayesian Modeling and Inference

U. of Arizona School of Information

March 22, 2021

Outline

Today:

- Recap of linear regression
- Multiple regression
- Analyzing causal effects

Linear regression

Linear regression as a Bayesian model

Idea behind a basic linear regression:

- Have a target variable y and one or more predictors x_i
- Assume a linear relationship between x_i, y
- For fixed values of the predictors, y is normally distributed (normal residuals)

How do we write this as a Bayesian model?

expected $y_j = \alpha + \beta_1 x_1(j) + \beta_2 x_2(j) + \cdots + \beta_n x_n(j)$

linear function of predictors

Linear regression as a Bayesian model

likelihood: observations normally distributed around line

$$y_j \sim \mathrm{Normal}(\mu_j, \sigma)$$

one predictor

$$\mu_j = \alpha + \beta x$$

expected y_j is linear function of x

$$\alpha \sim \mathrm{Normal}(0, 0.2)$$

$$\beta \sim \mathrm{Normal}(0, 0.5)$$

$$\sigma \sim \mathrm{HalfCauchy}(1)$$

priors for the model parameters.

What's with those priors?

Priors on α, β:

$$\alpha \sim \text{Normal}(0, 0.2)$$
$$\beta \sim \text{Normal}(0, 0.5)$$

95% prior mass with ±1

- Normal distribution $\rightarrow$ some regularization
  - MAP estimate with flat priors on α, β is OLS
  - MAP estimate with normal priors on α, β is ridge regression
- Prior parameter values chosen for standardized data
- HalfCauchy on σ; could be, e.g. Exponential instead

(regularized linear model

$$z = \frac{x - \bar{x}}{s_x}$$

Prior predictive checks

Why put regularizing priors?

Practically: reduce overfitting

Philosophically:

- Intercept should be near zero (so α should get a tight prior)
- Slopes should not produce impossibly strong relationships (so most of the prior mass for slopes should be between ± 1)
- We should be more skeptical of very strong effects

Standardizing is what lets us make these guesses.

Example: Dow Jones and climate

Simple linear model

years 1959 - 2016

We have a data set with measurements of annual global average temperature, measured as a difference from the 20th-century average, and the Dow Jones Industrial Average, a US stock market index.

Exploratory analysis suggests that the log of the DJIA might be an effective predictor of temperature. So let's build the model.

proxy for the size of US industrial economy

A simple linear model

Simple linear model with

$$
\left.
\begin{aligned}
T_i &\sim \mathrm{Normal}(\mu_i, \sigma) \quad \text{likelihood} && (1) \\
\mu_i &= \alpha + \beta \log \mathrm{DJIA}_i \quad \text{model equation} && (2)
\end{aligned}
\right.
$$

$$
\left.
\begin{aligned}
\alpha &\sim \mathrm{Normal}(0, 0.2) && (3) \\
\beta &\sim \mathrm{Normal}(0, 0.5) && (4) \\
\sigma &\sim \mathrm{HalfCauchy}(1) && (5)
\end{aligned}
\right\} \text{priors}
$$

The model in PyMC3

could be pm.Deterministic('mu', formula)

```
with pm.Model() as linear_model:
(3) alpha = pm.Normal('alpha', 0, 0.2)
(4) beta = pm.Normal('beta', 0, 0.5)
(5) sigma = pm.HalfCauchy('sigma', 1)
(2) mu = alpha + beta * climate['logDJIA'] + β2 * climate['x2']
(1) y = pm.Normal('y', mu, sigma,
        observed = climate['temp'])
    prior_sample = pm.sample_prior_predictive() ☆
    trace = pm.sample()
```

prior for another slope

samples post. values of α, β, σ

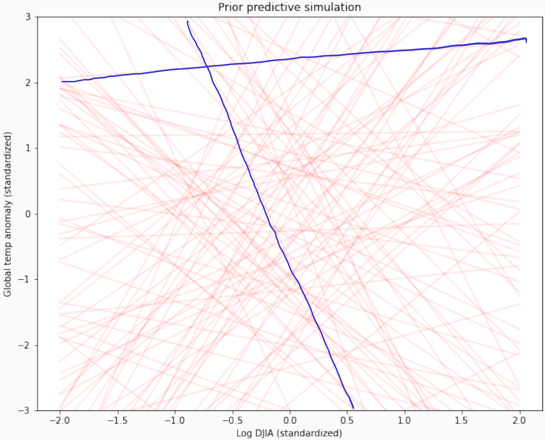
$$\beta = (X^T X)^{-1} X y$$

Prior predictive simulations

With the vague prior:

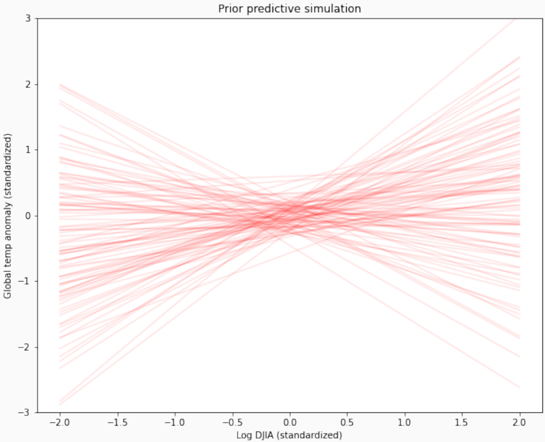
$\alpha \sim N(0, 2)$

$\beta \sim N(0, 2)$



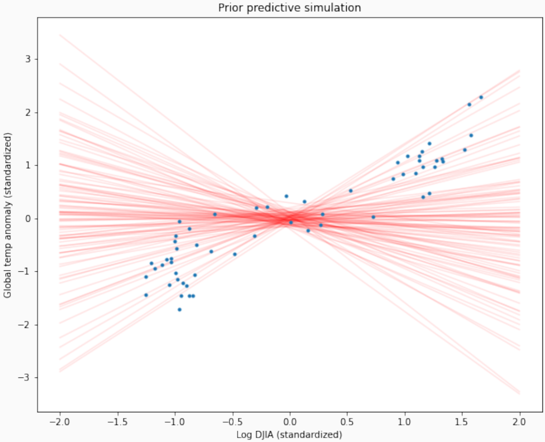
Prior predictive simulations

With the regularizing prior:



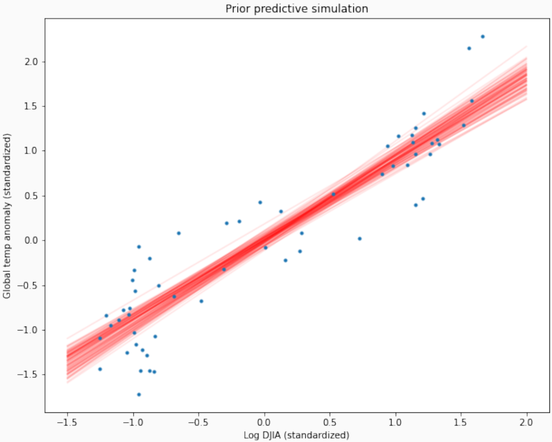
Add the data

Adding the data, we see most of the lines in the prior are inconsistent with the data:



Posterior distribution

After adding the data, the distribution of lines looks like this:



Reading the summary

Let's extract some information from the summary:

	mean	sd	hdi_3%	hdi_97%
alpha	0.001	0.052	-0.088	0.105
beta	0.910	0.054	0.815	1.015
sigma	0.404	0.041	0.333	0.481

posterior information.

posterior mean for the slope

az.summary(trace)

pm.summary(trace) (older)

$\hat{T} \approx 0.91 * \log DJIA$

A confound

You've heard the truism: correlation does not imply causation

- Does DJIA *predict* temperature? Definitely!
- Do changes in DJIA *cause* temperature changes? Probably not
- What's going on? Confounding with a third variable:

A confound

You've heard the truism: correlation does not imply causation

- Does DJIA *predict* temperature? Definitely!
- Do changes in DJIA *cause* temperature changes? Probably not
- What's going on? Confounding with a third variable: CO_2 concentration

Example: urban foxes

Urban foxes in London

Data set: urban foxes living in groups

Predictors:

- average food available to group
- group size

Target: fox weight

What happens if you feed the foxes?

If food is added to an area, will the foxes get bigger?

- This is a causal question, not just a statistical question
- Difference: talks about an intervention
- Simplest thing to try: regress fox weight on average food

Simple linear regression

Linear regression for fox weight:

$$
\begin{aligned}
w_i &\sim \mathrm{Normal}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta_f \mathrm{food} \\
\alpha &\sim \mathrm{Normal}(0, 0.2) \\
\beta_f &\sim \mathrm{Normal}(0, 0.5) \\
\sigma &\sim \mathrm{HalfCauchy}(1)
\end{aligned}
$$

What does the fox model say?

Here are the estimates from the fox model:

slope

	mean	sd	hdi_3%	hdi_97%
bF	-0.024	0.092	-0.191	0.150
alpha	-0.003	0.099	-0.183	0.185
sigma	1.013	0.068	0.884	1.131

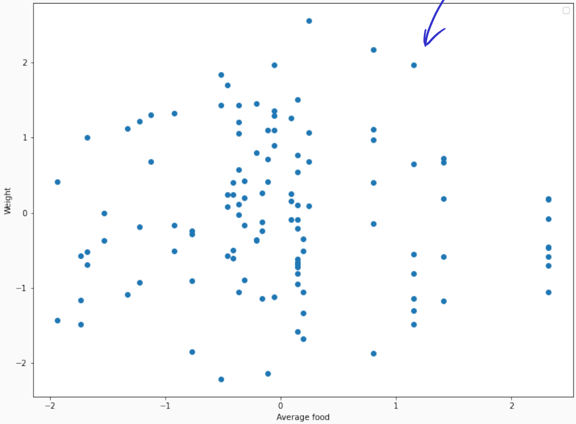
This is about as close to zero as we can get. Does this check out with the data?

-.19 0 +.15

What does the fox model say?

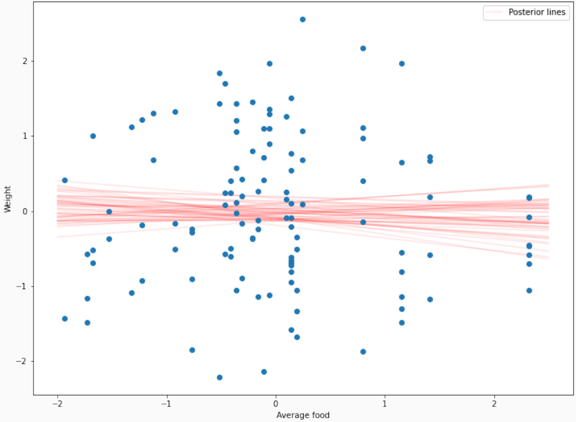
Scatterplot of weight vs. average food:

each point is an individual fox



What does the fox model say?

Scatterplot of weight vs. average food:



What does the fox model say?

The fox model tells us:

- No apparent association between food availability and fox weight
- But intuition tells us: if we provide more food, it must go somewhere!

What does the fox model say?

The fox model tells us:

- No apparent association between food availability and fox weight
- But intuition tells us: if we provide more food, it must go somewhere!
- More foxes

How can we check this? Include both variables.

A multiple regression

$$
\begin{aligned}
w_i &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta_f \text{food} + \beta_g \text{groupsize}
\end{aligned}
$$

model equation now a linear function of (food, groupsize)

$$
\begin{aligned}
\alpha &\sim \text{Normal}(0, 0.2) \\
\beta_f &\sim \text{Normal}(0, 0.5) \\
\beta_g &\sim \text{Normal}(0, 0.5) \\
\sigma &\sim \text{HalfCauchy}(1)
\end{aligned}
$$

priors

Multiple regression results

Here are the results from the multiple regression:

	mean	sd	hdi_3%	hdi_97%
bG	-0.568	0.189	-0.937	-0.224
bF	0.475	0.188	0.153	0.859
alpha	0.001	0.090	-0.176	0.160
sigma	0.967	0.068	0.854	1.103

← interpreted as
← holding other
predictors fixed

bG - if you know food available to a fox, what do you learn from the group size?

bF - if you know the group size, what do you learn from observing the available food?

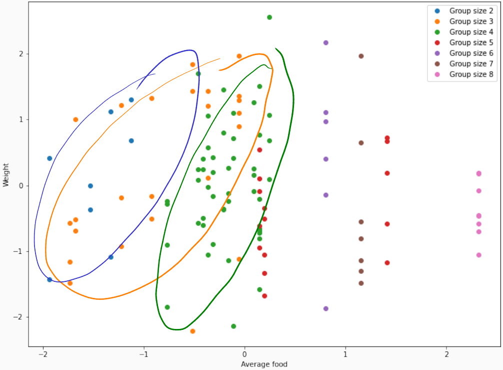
Multiple regression results

Here are the results from the multiple regression:

	mean	sd	hdi_3%	hdi_97%
bG	-0.568	0.189	-0.937	-0.224
bF	0.475	0.188	0.153	0.859
alpha	0.001	0.090	-0.176	0.160
sigma	0.967	0.068	0.854	1.103

Can we see this in the scatter plot?

Statistical control as stratification



The association between food and weight appears when the data is stratified by group size, but not before

Multiple regression in the climate example

If we had access to CO_2 concentration data, we could do the same and see what happens to the estimated association between DJIA and temperature.

Fortunately, we do!

Multiple regression in the climate example

If we had access to CO2 concentration data, we could do the same and see what happens to the estimated association between DJIA and temperature.

Fortunately, we do!

	mean	sd	hdi_3%	hdi_97%
alpha	-0.000	0.043	-0.077	0.085
beta_c	0.783	0.155	0.499	1.082
beta_d	0.164	0.155	-0.111	0.475
sigma	0.335	0.032	0.277	0.396

slope for CO_2 →

slope for log DJIA →

$$T = \beta_c * CO_2 + \beta_d \log DJIA$$

Revealing and eliminating associations

In both cases we gained something by including the extra variable:

- In the climate example, a spurious association is eliminated
- In the fox example, a masked association is revealed

This is the power of multiple regression: it thinks "hypothetically" about the variables

Beware: including the wrong variables can introduce spurious associations!

Comparison

In both cases we have something of a similar thing going on:

- Underlying causal variable influences a secondary predictor and the target variable
  - Food influences group size and fox weight
  - CO_2 influences DJIA and global temperature
- The difference: a third causal relationship (or lack thereof)

Interpretation

Causal interpretation:

- If we alter the amount of CO2, will we change temperature? Yes
- If we alter the amount of food, will we make the foxes bigger? No

Which variables tell us this?

- To reach the conclusion about CO2, does including DJIA matter?

Interpretation

Causal interpretation:

- If we alter the amount of CO2, will we change temperature? Yes
- If we alter the amount of food, will we make the foxes bigger? No

Which variables tell us this?

- To reach the conclusion about CO2, does including DJIA matter? Only for precision

Interpretation

Causal interpretation:

- If we alter the amount of CO2, will we change temperature? Yes
- If we alter the amount of food, will we make the foxes bigger? No

Which variables tell us this?

- To reach the conclusion about CO2, does including DJIA matter? Only for precision
- To reach the conclusion about foxes, does including group size matter?

Interpretation

Causal interpretation:

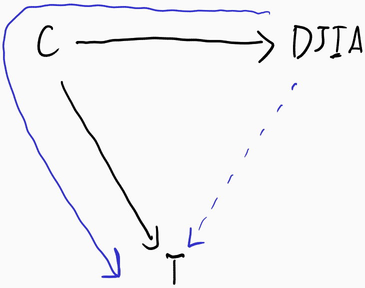
- If we alter the amount of CO2, will we change temperature? Yes
- If we alter the amount of food, will we make the foxes bigger? No

Which variables tell us this?

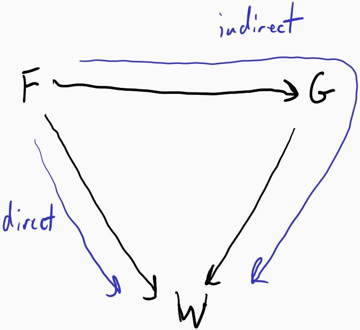
- To reach the conclusion about CO2, does including DJIA matter? Only for precision
- To reach the conclusion about foxes, does including group size matter? Yes

Graphical comparison

Climate:



Foxes:



Direct vs. total effect

The graph for the fox model gives us a distinction between two effects of food on weight:

- direct effect: associated with the arrow from F to W; effect of food on weight at fixed group size
  - estimated by the multiple regression, but not the simple regression
- total effect: associated with all paths from F to W; effect of food on weight, including those mediated by changes in group size
  - estimated by the simple regression, but not the multiple regression

Summary

Today:

- Recap of linear regression
- Multiple regression
- Confounding and masked associations

Next time: DAGs, confounding, and causal inference