Awesome Linear Regression

Max Kochurov

MSU

Lecture 6

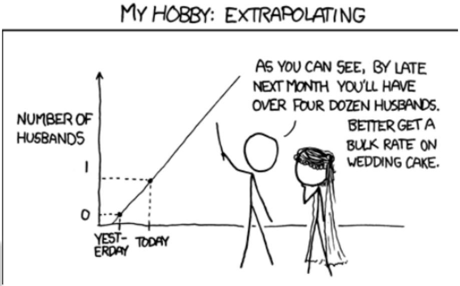
Agenda

Why linear regression is a thing

- Policy-making
  - Correlation strength
  - Influence direction
  - Effect size calculation
- Part of a more complicated model
  - Marketing Mix Models
  - AB tests

Lego

Linear regression is a common thing in all sorts of statistical models



Putting notation

In Econometrics people got used to this notation

$$y \sim x_1 + x_2 + \cdots + x_k$$

Translation

My y depends linearly on x_1, x_2, ..., x_k

Putting notation

In Econometrics people got used to this notation

$$y \sim x_1 + x_2 + \cdots + x_k$$

Translation

My y depends linearly on x_1, x_2, ..., x_k

Which also assumes constant regressor by default

$$y \sim {\color{red}1} + x_1 + x_2 + \cdots + x_k$$

Putting notation

In Econometrics people got used to this notation

$$y \sim x_1 + x_2 + \cdots + x_k$$

Translation

My y depends linearly on x_1, x_2, ..., x_k

Which also assumes constant regressor by default

$$y \sim 1 + x_1 + x_2 + \cdots + x_k$$

And in principle means estimating β

$$y \sim \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

More than just Linear

% change of x_1 causes % change in y

$$\log y \sim \log x_1 + \dots$$

More than just Linear

% change of x_1 causes % change in y

$$\log y \sim \log x_1 + \dots$$

% change of x_1 causes absolute change in y

$$y \sim \log x_1 + \dots$$

More than just Linear

% change of x_1 causes % change in y

$$\log y \sim \log x_1 + \dots$$

% change of x_1 causes absolute change in y

$$y \sim \log x_1 + \dots$$

absolute change of x_1 causes % change in y

$$\log y \sim x_1 + \dots$$

More than just Linear

% change of x_1 causes % change in y

$$\log y \sim \log x_1 + \dots$$

% change of x_1 causes absolute change in y

$$y \sim \log x_1 + \dots$$

absolute change of x_1 causes % change in y

$$\log y \sim x_1 + \dots$$

Takeouts

Interpret the dependencies carefully when using logs

GLMs: Understanding Basics

It is possible to use arbitrary likelihood function to **link** observations. The traditional function is like

$$c_i \sim \mathrm{Binom}(p_i, n_i)$$
$$\mathrm{link}^{-1}(p_i) \sim x_{1i} + x_{2i} + \cdots + x_{ki}$$

Heteroscedasticity

We can add more flexibility

$$
\begin{aligned}
y_i &\sim \mathcal{N}(m_i, s_i) \\
m_i &\sim x_i + \dots \\
\log s_i &\sim z_i
\end{aligned}
$$

Heteroscedasticity

We can add more flexibility

$$
\begin{aligned}
y_i &\sim \mathcal{N}(m_i, s_i) \\
m_i &\sim x_i + \dots \\
\log s_i &\sim z_i
\end{aligned}
$$

Note

See that s_i depends on z_i. Such models are usually estimated using optimisation.

Other likelihoods

Even more flexibility could be achieved by changing the likelihood and relaxing assumptions

$$
\begin{aligned}
y_i &\sim \mathcal{T}(\nu_i, m_i, s_i) \\
m_i &\sim x_i + \dots \\
\log s_i &\sim z_i + \dots \\
\log \nu_i &\sim w_i + \dots
\end{aligned}
$$

Other likelihoods

Even more flexibility could be achieved by changing the likelihood and relaxing assumptions

$$
\begin{aligned}
y_i &\sim \mathcal{T}(\nu_i, m_i, s_i) \\
m_i &\sim x_i + \ldots \\
\log s_i &\sim z_i + \ldots \\
\log \nu_i &\sim w_i + \ldots
\end{aligned}
$$

Note

Heteroscedastic StudentT model with variable degrees of freedom.
Without regularisation estimates are very noisy

Estimations

From Econometrics we remember the Basic Maximum Likelihood estimator

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

Estimations

From Econometrics we remember the Basic Maximum Likelihood estimator

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

1. What if we know that a relation is positive?
2. What if we know the magnitude of β is small?
3. What if we know some variables are not important?

Limitations

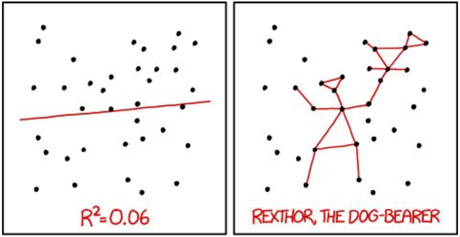
Within the frequentist statistics it is impossible to use additional information

Priors

Bayesian approach is about setting priors, what are they?

$$y_i \sim \mathcal{N}(c + \beta^\top x_i, \sigma)$$
$$\beta_j \sim ???$$
$$c \sim ???$$
$$\sigma \sim ???$$

There were introduced two parameters: β, σ



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

Setting Priors

It is a common thing to set priors with Normal distribution

$$
\begin{aligned}
y_i &\sim \mathcal{N}(c + \beta^\top x_i, \sigma) \\
\beta_j &\sim \mathcal{N}(0, 1) \\
c &\sim \mathcal{N}(0, 1) \\
\sigma &\sim \mathcal{N}_+(1) \qquad // \text{ Half Normal}
\end{aligned}
$$

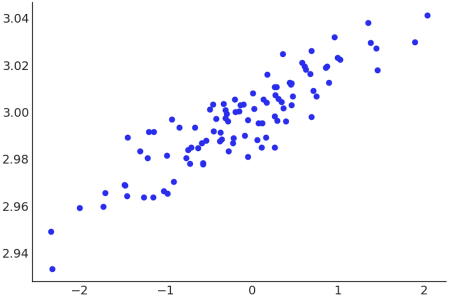


Figure: Example Data

Setting Priors

It is a common thing to set priors with Normal distribution

$$
\begin{aligned}
y_i &\sim \mathcal{N}(c + \beta^\top x_i, \sigma)\\
\beta_j &\sim \mathcal{N}(0, 1)\\
c &\sim \mathcal{N}(0, 1)\\
\sigma &\sim \mathcal{N}_+(1) \qquad \text{// Half Normal}
\end{aligned}
$$

Attention

Default parameters for β and σ priors are dangerous

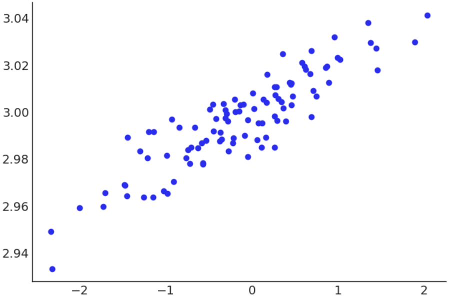


Figure: Example Data

Setting Priors

To set priors you are advised to use prior predictive

$$
\begin{aligned}
y_i &\sim \mathcal{N}(c + \beta^\top x_i, \sigma) \\
\beta_j &\sim \mathcal{N}(0, 1) \\
c &\sim \mathcal{N}(0, 1) \\
\sigma &\sim \mathcal{N}_+(1) \qquad // \text{ Half Normal}
\end{aligned}
$$

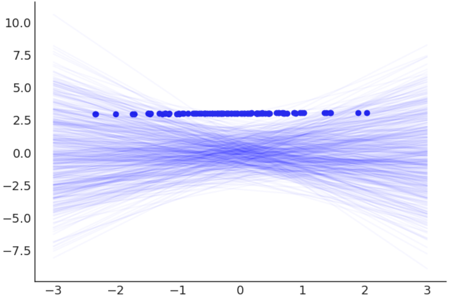


Figure: Prior predictive

Careful

Sometimes prior predictive can go off, check the plot first and interpret

Setting Priors

To set priors you are advised to use prior predictive

$$
\begin{aligned}
y_i &\sim \mathcal{N}(c + \beta^\top x_i, \sigma) \\
\beta_j &\sim \mathcal{N}(0, 100) \\
c &\sim \mathcal{N}(0, 100) \\
\sigma &\sim \mathcal{N}_+(1) \qquad // \text{ Half Normal}
\end{aligned}
$$

Careful

Sometimes prior predictive can go off, check the plot first and interpret

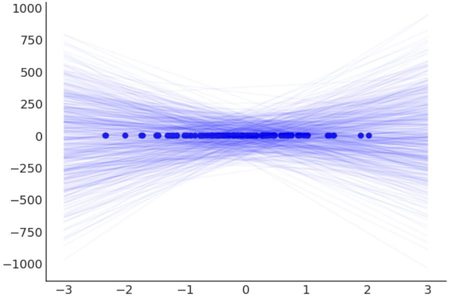


Figure: Prior predictive

The more parameters, the bigger the issue

Assuming everything is independent (a priory), we can compute theoretical variances

$$y_i \sim \mathcal{N}(c + \beta^\top x_i, \sigma)$$
$$V[y_i] = \sum V[x_{ij}] * V[\beta_j]$$

Things are different when $x_i \in R^3$ and $x_i \in R^{100}$

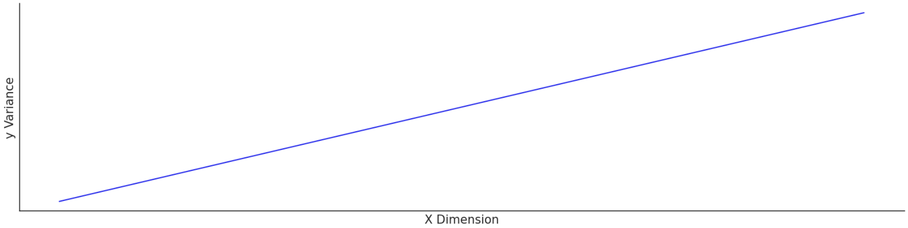


Figure: The more variables you include, the more variance you expect

A quick Fix

The easy way to remove dependency on number of regressors is this

$$y_i \sim \mathcal{N}(c + \beta^\top x_i, \sigma)$$
$$\beta_j \sim \mathcal{N}(0, \frac{\sigma_\beta^2}{D})$$
$$\dots$$

Thing that usually help

Standardize the data: $a \mapsto \frac{a - \text{mean}(a)}{\text{std}(a)}$

A Practical Approach

Standardize the data first: $a \mapsto \frac{a-\text{mean}(a)}{\text{std}(a)}$

$$\begin{aligned}
\bar{y}_i &\sim \mathcal{N}(c + \bar{\beta}^\top \bar{x}_i, \sigma) \\
\bar{\beta} &\sim \mathcal{N}(0, 1/D) \\
c &\sim \mathcal{N}(0, 1) \\
\sigma &\sim \mathcal{N}_+(1)
\end{aligned}$$

1. Input/Output variance is fixed and 1
2. Input/Output mean is fixed and 0
3. Works most of the time
4. Hard to set σ prior

A Practical Approach

Standardize the data first: $a \mapsto \frac{a - \text{mean}(a)}{\text{std}(a)}$

$$
\begin{aligned}
\bar{y}_i &\sim \mathcal{N}(c + \bar{\beta}^\top \bar{x}_i, \sigma) \\
\bar{\beta} &\sim \mathcal{N}(0, 1/D) \\
c &\sim \mathcal{N}(0, 1) \\
\sigma &\sim \mathcal{N}_+(1)
\end{aligned}
$$

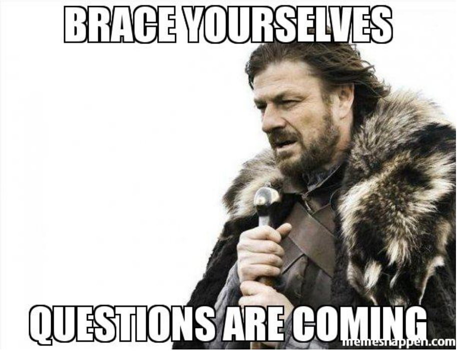
1. Input/Output variance is fixed and 1
2. Input/Output mean is fixed and 0
3. Works most of the time
4. Hard to set σ prior

Recovering original params

$$\beta_j = \frac{\bar{\beta}}{\text{std}(x_j)}$$

What we know that we know

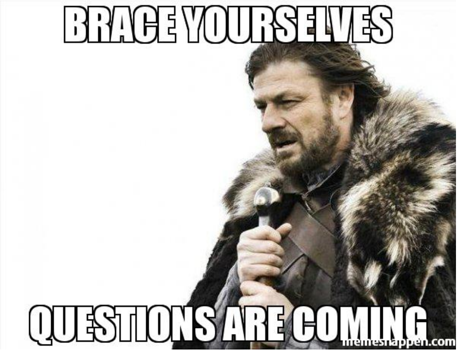
Setting priors is hard how can we make that easier? Let's ask questions!



What we know that we know

Setting priors is hard how can we make that easier? Let's ask questions!

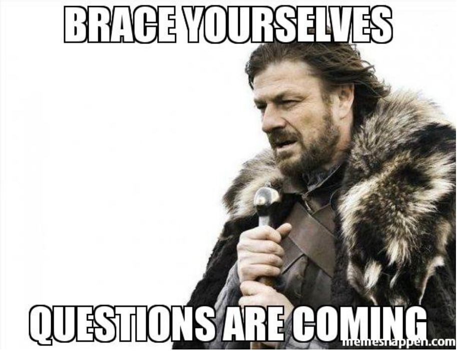
- Q: What do we know about Linear regressions?



What we know that we know

Setting priors is hard how can we make that easier? Let's ask questions!

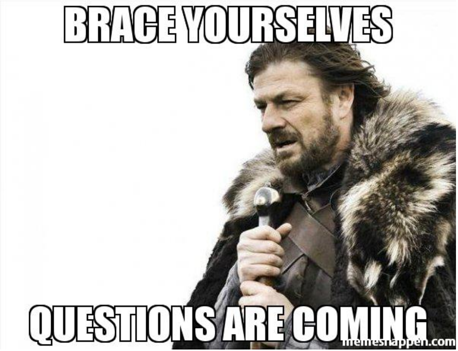
- Q: What do we know about Linear regressions?
- A: They have R^2 goodness of fit



What we know that we know

Setting priors is hard how can we make that easier? Let's ask questions!

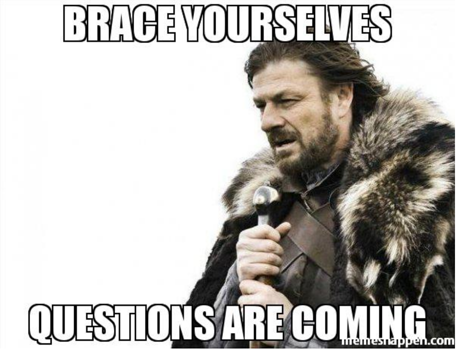
- Q: What do we know about Linear regressions?
- A: They have R^2 goodness of fit
- Q: Anything else?



What we know that we know

Setting priors is hard how can we make that easier? Let's ask questions!

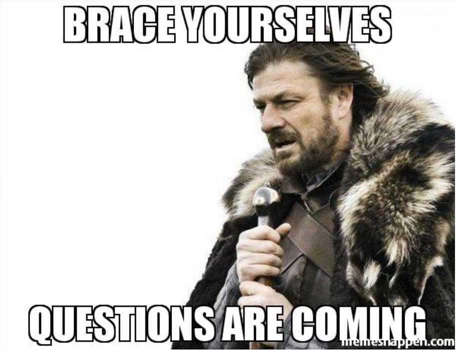
- Q: What do we know about Linear regressions?
- A: They have R^2 goodness of fit
- Q: Anything else?
- A: Some variables are more important than others



What we know that we know

Setting priors is hard how can we make that easier? Let's ask questions!

- Q: What do we know about Linear regressions?
- A: They have R^2 goodness of fit
- Q: Anything else?
- A: Some variables are more important than others
- A: Some variables should have positive effect size



The R^2 Prior

What is R^2?

1. Used to be goodness of fit statistics
   - 0 - very bad
   - 1 - excellent
2. When close to 1 usually over-fit
3. R^2 - **Fraction of Variance Explained**

$$R^2 = 1 - \frac{\sigma_r^2}{\sigma_T^2}$$

$$FVU = \frac{\sigma_r^2}{\sigma_T^2}$$

- σ_r^2 - residual variance
- σ_T^2 - total variance
- FVU - **F**raction **V**ariance **U**nexplained

Setting R^2 Prior

R^2 prior is very intuitive to say about before any data is fit.

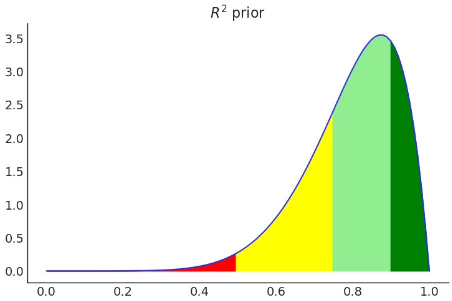
- $R^2 < 0.5$ – field experiments, noisy data
- $0.5 < R^2 < 0.75$ – field experiments, clean data
- $0.75 < R^2 < 0.90$ – lab experiments, noisy data
- $R^2 > 0.90$ – lab experiments, clean data

```
Call:
lm(formula = y ~ ., data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.81177 -0.58567  0.05249  0.69674  2.40316 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.04580    0.05694   0.804  0.42188    
x1           0.42949    0.05874   7.311 2.52e-12 ***
x2           0.57386    0.06638   8.646 3.52e-16 ***
x3           0.26152    0.05773   4.530 8.58e-06 ***
x4          -0.29599    0.05444  -5.438 1.14e-07 ***
x5          -0.17564    0.05428  -3.236  0.00135 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9733 on 294 degrees of freedom
Multiple R-squared:  0.4131,	Adjusted R-squared:  0.4032 
F-statistic: 41.39 on 5 and 294 DF,  p-value: < 2.2e-16
```

Prior R^2

- What is the quality of your data?
- Do you have all factors to explain data?
- Is your data collection method accurate?

$$R^2 \sim \text{Beta}(\mu = \tilde{\mu}_r, \sigma = \tilde{\sigma}_r)$$



Note

This is different from Bayesian R^2. **Prior** R^2 is **your expectation** about model quality.

Feel the Difference

How it was

- How to set prior for β?
- What does this prior mean?
- Oh, I should change prior if I add parameters
- How to set prior for σ?
- Too complicated, where are the defaults?
- Ah, defaults do not make any sense

How it is gonna be

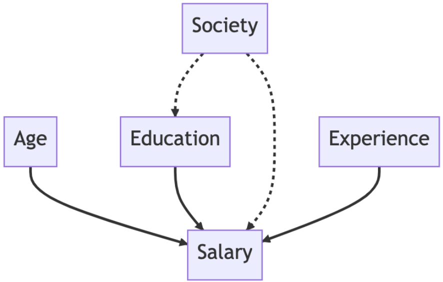
- How good is the model expected? The R^2
- Which variable is more or less important?
- What is the expected direction of influence for variables?

Variable importance

You predict salary,

- is Age or Education more important?
- is Education or Experience more important?

In traditional models you can only figure out post factum



Variable importance

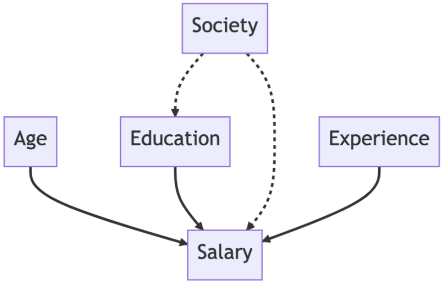
You predict salary,

- is Age or Education more important?
- is Education or Experience more important?

In traditional models you can only figure out post factum

Bayesian approach

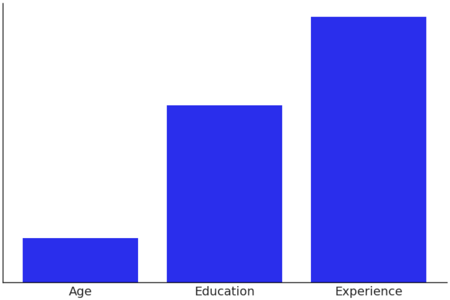
- Set expectations on how features are important
- Bayesian Instrumental Variables



What is variable importance?

There are several approaches

- Amount of information gain
- **F**raction of **V**ariance **E**xplained



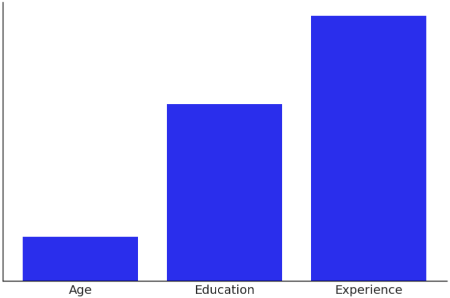
What is variable importance?

There are several approaches

- Amount of information gain
- **F**raction of **V**ariance **E**xplained

Use same idea!

Similar to R^2 we can set **FVE** per feature



What is variable importance?

There are several approaches

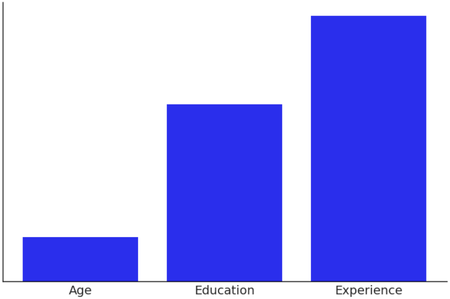
- Amount of information gain
- **F**raction of **V**ariance **E**xplained

Use same idea!

Similar to R^2 we can set **FVE** per feature

A simple idea

$$\phi_{\text{FVE}} \sim \text{Dirichlet}(\alpha_{\text{FVE}})$$

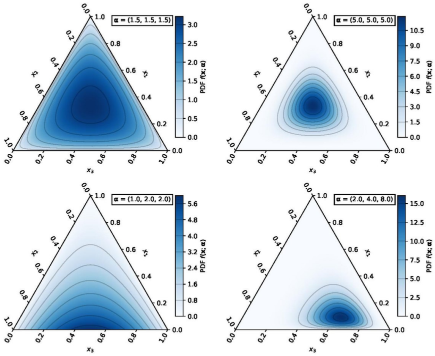


Understanding FVE Prior

We need to understand the Dirichlet distribution

$$\phi_{\mathsf{FVE}} \sim \mathrm{Dirichlet}(\alpha_{\mathsf{FVE}})$$

- The higher α_i the more variable i is important
- The higher α_i the more confidence is put into importance



α_{FVE} in Examples

$$\phi_{\text{FVE}} \sim \text{Dirichlet}(\alpha_{\text{FVE}})$$

- $\alpha_{\text{FVE}} = (1, 1, 1)$ - I know nothing about importances, maybe some variables are not used
- $(\alpha_{\text{FVE}})_i = 1$ - variable might not be used or be very imporant, no clue
- $(\alpha_{\text{FVE}})_i = 10$ - variable should be probably used
- $(\alpha_{\text{FVE}})_i = 20$ - variable is definitely used
- $\alpha_{\text{FVE}} = (10, 20, 30)$ - All variables are used, but 2d and 3d are increasingly more important

α_{FVE} in Examples

$$\phi_{\mathsf{FVE}} \sim \mathrm{Dirichlet}(\alpha_{\mathsf{FVE}})$$

- $\alpha_{\mathsf{FVE}} = (1, 1, 1)$ - I know nothing about importances, maybe some variables are not used
- $(\alpha_{\mathsf{FVE}})_i = 1$ - variable might not be used or be very imporant, no clue
- $(\alpha_{\mathsf{FVE}})_i = 10$ - variable should be probably used
- $(\alpha_{\mathsf{FVE}})_i = 20$ - variable is definitely used
- $\alpha_{\mathsf{FVE}} = (10, 20, 30)$ - All variables are used, but 2d and 3d are increasingly more important

Disclaimer

Yes, this is the most handwavy interpretation ever

α_{FVE} and R^2

$$\phi_{\mathsf{FVE}} \sim \mathrm{Dirichlet}(\tilde{\alpha}_{\mathsf{FVE}})$$
$$R^2 \sim \mathrm{Beta}(\mu = \tilde{\mu}_r, \sigma = \tilde{\sigma}_r)$$

What you decide

1. How good is the model in principle? (R^2)
2. How good is every given feature ($\tilde{\alpha}_{\mathsf{FVE}}$)

Putting all together

1. Standardize the data: $a \mapsto \frac{a - \text{mean}(a)}{\text{std}(a)}$
2. Decide on R^2
3. Decide on feature importance
4. Done

$$
\begin{aligned}
\bar{y}_i &\sim \mathcal{N}(\bar{\beta}^\top \bar{x}_i, \sigma) \\
\phi_{\text{FVE}} &\sim \text{Dirichlet}(\tilde{\alpha}_{\text{FVE}}) \\
R^2 &\sim \text{Beta}(\mu = \tilde{\mu}_r, \sigma = \tilde{\sigma}_r) \\
\sigma^2 &= 1 - R^2 \\
\bar{\beta} &\sim \mathcal{N}(0, \sqrt{\phi_{\text{FVE}} \cdot R^2})
\end{aligned}
$$

Even more formulas

This is a recently developed the R2D2M2 prior[1], read more detailed math there.

Can we Add More? R2D2M2CP

Yes, yes and yes!

- "What is the sign of correlation?"
- "How I'm sure correlation is positive?"

Can we Add More? R2D2M2CP

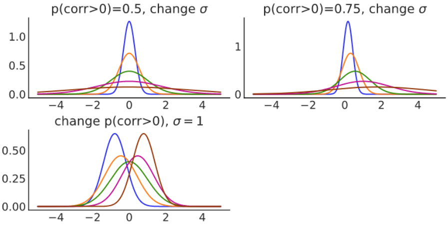
Yes, yes and yes!

- "What is the sign of correlation?"
- "How I'm sure correlation is positive?"

The solution I propose:

$$P(\bar{\beta}_j > 0) = (\psi_{CP})_j$$

$$\psi_{CP} \sim \text{Beta}(\mu = \mu_{CP}, \sigma = \sigma_{CP})$$



Technical Details

$$
\begin{aligned}
P(\bar{\beta}_j > 0) &= (\psi_{CP})_j \\
\bar{\beta} &\sim \mathcal{N}(\mu_{CP}(\psi_{CP}, R^2 \cdot \phi_{\text{FVE}}), \sigma_{CP}(\psi_{CP}, R^2 \cdot \phi_{\text{FVE}})) \\
\psi_{CP} &\sim \text{Beta}(\mu = \mu_{CP}, \sigma = \sigma_{CP}) \\
\phi_{\text{FVE}} &\sim \text{Dirichlet}(\tilde{\alpha}_{\text{FVE}}) \\
R^2 &\sim \text{Beta}(\mu = \tilde{\mu}_r, \sigma = \tilde{\sigma}_r)
\end{aligned}
$$

μ_{CP}, σ_{CP} solution is unique

$$
\begin{cases}
\mu_{CP}(p, v) = \dfrac{\sqrt{2v}\,\text{erf}^{-1}(2p-1)}{\sqrt{2\,\text{erf}^{-1}(2p-1)^2+1}} \\
\sigma_{CP}(p, v) = \dfrac{\sqrt{v}}{\sqrt{2\,\text{erf}^{-1}(2p-1)^2+1}}
\end{cases}
$$

Putting all Together

To use R2D2M2CP prior decide on

1. Standardize the data: $a \mapsto \frac{a - \text{mean}(a)}{\text{std}(a)}$
2. Decide on R^2
3. Decide on feature importance
4. Decide on correlation direction
5. Done, like never before!

A practical implementation is merged[2]

https://github.com/pymc-devs/pymc-experimental/pull/137

Back to GLMs

Consider this model blueprint:

$$
\begin{aligned}
y_i &\sim \mathcal{T}(\nu_i, m_i, s_i) \\
m_i &\sim x_i + \dots \\
\log s_i &\sim z_i + \dots
\end{aligned}
$$

Back to GLMs

Consider this model blueprint:

$$
\begin{aligned}
y_i &\sim \mathcal{T}(\nu_i, m_i, s_i) \\
m_i &\sim x_i + \dots \\
\log s_i &\sim z_i + \dots
\end{aligned}
$$

- Which factors contribute sigma s? (variable importance guess)

Prior for Nu

Degrees of freedom can be considered with a special prior:
https://github.com/pymc-devs/pymc-experimental/pull/252

Back to GLMs

Consider this model blueprint:

$$
\begin{aligned}
y_i &\sim \mathcal{T}(\nu_i, m_i, s_i) \\
m_i &\sim x_i + \dots \\
\log s_i &\sim z_i + \dots
\end{aligned}
$$

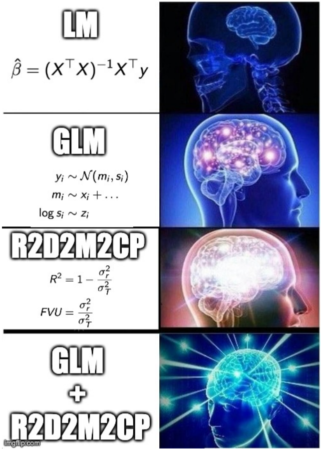
- Which factors contribute sigma s? (variable importance guess)
- Do they even contribute? (R^2 guess)

Prior for Nu

Degrees of freedom can be considered with a special prior: https://github.com/pymc-devs/pymc-experimental/pull/252

Remarks

- The R2D2M2CP prior is hard to pronounce
- Can extend thinking for the traditional linear models
- Goes beyond to GLMs for granular control of auxiliary models
- Application for GAMs mix with GPs is something to also explore



References I

J. E. Aguilar and P.-C. Bürkner.
Intuitive joint priors for bayesian linear multilevel models: The r2d2m2 prior, 2023.

M. Kochurov.
pymc-devs/pymc-experimental: Pull Request 137 R2D2M2CP.
GitHub, 2023.