Model Building

Sofia Olhede

EPFL

November 16, 2020

1. How do we make a linear model?

2. Model Selection

What is a model?

- In our work up to now, we have assumed that we are given the relationship
$$\mathrm{Y} = \mathrm{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$
- With this model we can do all sorts of tasks.
- However, in practice we are given data (Y, X) and suspect a linear relationship between Y and some columns of X! We don't know a priori which exactly!
- Need to select a "most appropriate" subset of the columns of X.
- General principle: parsimony. Idea comes from William of Ockham (Occam's Razor). "Entities should not be multiplied without necessity".

How do we get there?

- Graphical exploration; this provides initial picture:
- Plots of Y against candidate variables;
- Plots of transformations of certain variables against Y;
- Plots of pairs of candidate variables.
- This will often provide a starting point, but:
- Automatic Model Selection: Need objective model comparison criteria, as a screening device.
- What if models to be compared are not nested?
- Automatic Model Building: Situations when p large, so there are lots of possible models.
- Automatic methods for building a model? We saw that ANOVA depends on the order of entry of variables in the model . . .

How do we get there?

- Consider design matrix X with p variables.
- 2^p possible models (inclusion or not of each variable, or nonlinear function of each variable).
- Denote set of all models generated by X by 2^{X} (we call this the model powerset).
- If wish to consider k different transformations of each variable, then p becomes $(1+k)p$.
- Fast algorithms (branch and bound) exist to fit them, but they don't work well for large p, and anyway . . .
- . . . need criterion for comparison. So given a collection of models, we need an automatic (objective) way to pick out a "best" one (unfortunately cannot look carefully at all of them, but nothing can replace careful scrutiny of the final model by an experienced researcher).

Forward Search–stepwise regression

- We can start from the constant model.
- Fit p simple linear regression models, where you add a variable to the constant at the time. You thus search through all the single-variable models the best one (the one that results in the lowest residual sum of squares). You pick and fix this one in the model.
- Now search through the remaining $p - 1$ variables and find out which variable should be added to the current model to best improve the residual sum of squares.
- In advance you have fixed a stopping criteria and stop either when it is satisfied or p variables have been added.

Backward elimination–stepwise regression

- We can start from the full model with p variables.
- In order to be able to do backwards elimination, we need to be in a situation where we have more $n > p$ because we we can only invert $\mathrm{X}^T\mathrm{X}$ when n is greater than p. This is a full rank model.
- We start from the hypothesis test, we then remove the variable with the largest p-value. That is, the variable that is the least statistically significant.
- The new $(p-1)$-variable model is then addressed, as the variable with the largest p-value is removed.
- Continue until a pre-set stopping rule is reached.
- It is also possible to use bidirectional elimination, a combination of the above, testing at each step for variables to be included or excluded.

Not full rank models I

- We have ignored another aspect of the linear model. Very early we assumed that the model under study was full rank.
- There are settings where this is not a realistic assumption.
- Say we would like to test the effect of adding different food supplements when raising cattle. Say we have two food supplements, A and B. We then define dummy covariates x_{ij} to denote the feed.
- Then the weight of the cattle takes the form

$$\mathbb{E}\{Y_i\} = \mu + \theta_A x_{i1} + \theta_B x_{i2}.$$

If we write down the full design matrix then we get

$$\mathsf{X} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & \dots & \dots \\ 1 & 0 & 1 \\ 1 & \dots & \dots \end{pmatrix}.$$

Not full rank models II

- Clearly this is not full rank as column one equals the mean of the other columns.
- In fact we assume $\mathrm{rank}(X) = k < p \leq n$.
- The least squares approach asks for us to solve

$$X^T X \hat{\beta} = X^T Y.$$

- But, now we have $X^T X$ is not invertible. → no full rank
- ~~But there is the notion of a generalized inverse $(X^T X)^-$.~~
- ~~A poor man's version is to determine linear combination of the parameters that are estimable. These are known as contrasts. For instance you can estimate $\theta_A - \theta_B$.~~
- ~~A way of arriving at those solutions would be to introduce a linear constraint like $\sum \theta_x = 0$. Then the mean over all observations would estimate μ.~~

Model selection methods

- There are many choices of model selection.
  - * Prediction error based criteria (CV);
  - * Information criteria (AIC, BIC, ...);
  - * Mallow's C_p statistic.
- Before looking at these, let's introduce terminology: Suppose that the truth is

$$\mathrm{Y} = \mathrm{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

but that actually $\boldsymbol{\beta}_r$ is zero for some subset of variables in $\boldsymbol{\beta}$.

- The true model contains only the columns for which $\boldsymbol{\beta}_r \neq 0$. Equivalently we could rewrite this as X_* as the matrix of columns of X for which we have non-zero coefficients.
- However any design matrix $\mathrm{X}_\circ$ that has the correct columns, plus potentially more columns, is a correct model.
- An incorrect model $\mathrm{X}_\bullet$ does not contain all columns required.

Prediction error

- We may wish to choose a model by minimising the error we make on average, when predicting a future observation given our model.
- Recall we start from the observations Y; by using the hat matrix P we can predict Y via

subset of columns?

$$\widehat{Y} = \mathrm{P}\mathrm{Y}.$$

- Every model that we may use (that is full rank), f say has an associated hat matrix P_f. We can then form

$$\widehat{Y}_f = \mathrm{P}_f \mathrm{Y}.$$

- Assume we have independent realisations Y_+ with the same design matrix X.
- We might then chose the model f so that

$$f^* = \arg\min_{f \in 2^{\mathrm{X}}} \frac{1}{n} \mathbb{E}\left\{ \|\mathrm{Y}_+ - \widehat{Y}_f\|^2 \right\}.$$

Prediction error II

- Let X as before be the design matrix. Let us define additional design matrices of X_* (correct non–zero coefficient model with q variables), and recall $X_\circ$ the design matrix with the correct model plus some zero coefficients, and $X_\bullet$ that does not contain the correct model.
- Assume we fit to $X_\circ$ (correct but with too many columns): The fitted model remains an $n \times 1$ vector:

$$\widehat{Y} = (X_\circ^T X_\circ)^{-1} X_\circ^T Y = P_\circ Y.$$

- Now assume we generate an identical vector observation from the true design matrix:

$$Y_+ = X_* \boldsymbol{\beta} + \boldsymbol{\epsilon}_+ = \boldsymbol{\mu} + \boldsymbol{\epsilon}_+.$$

- With this observation we may note that

$$\begin{aligned} Y_+ - \widehat{Y} &= \overbrace{\boldsymbol{\mu} + \boldsymbol{\epsilon}_+}^{Y_+ \to \text{slide 11}} - \overbrace{P_\circ(\boldsymbol{\mu} + \boldsymbol{\epsilon})}^{\hat{Y} = PY \ \to \text{prev. slide}} \\ &= \{I - P_\circ\}\boldsymbol{\mu} + \boldsymbol{\epsilon}_+ - P_\circ \boldsymbol{\epsilon}. \end{aligned}$$

Prediction error III

- From the previous page we can deduce that:

$$\begin{aligned}\|\mathsf{Y}_+ - \widehat{\mathsf{Y}}\|^2 &= \left(\mathsf{Y}_+ - \widehat{\mathsf{Y}}\right)^T \left(\mathsf{Y}_+ - \widehat{\mathsf{Y}}\right) \\ &= (\{\mathsf{I} - \mathsf{P}_\circ\}\boldsymbol{\mu} + \boldsymbol{\epsilon}_+ - \mathsf{P}_\circ\boldsymbol{\epsilon})^T(\{\mathsf{I} - \mathsf{P}_\circ\}\boldsymbol{\mu} + \boldsymbol{\epsilon}_+ - \mathsf{P}_\circ\boldsymbol{\epsilon}) \\ &= \boldsymbol{\mu}^T\{\mathsf{I} - \mathsf{P}_\circ\}\boldsymbol{\mu} + \boldsymbol{\epsilon}^T\mathsf{P}_\circ\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_+^T\boldsymbol{\epsilon}_+ + \text{ cross terms.}\end{aligned}$$

Finding the expectation of that, we may ignore the cross terms as their expectation will be zero. We define

all features including correct ones

$$\Delta = n^{-1}\,\mathbb{E}\,\|\mathsf{Y}_+ - \widehat{\mathsf{Y}}\|^2$$

worst (it has bias)

$$= \begin{cases} n^{-1}\boldsymbol{\mu}^T\{\mathsf{I} - \mathsf{P}_\circ\}\boldsymbol{\mu} + (1 + p/n)\sigma^2 & \text{if the model is wrong} \\ (1 + p/n)\sigma^2 & \text{if the model is correct} \\ (1 + q/n)\sigma^2 \hookrightarrow q<p & \text{if the model is true.} \end{cases}$$

best

only correct features

- Selecting a wrong model instead of the true model results in bias, since $\{\mathsf{I} - \mathsf{P}_\bullet\}\boldsymbol{\mu} \neq 0$ when μ is not in the column space of $\mathsf{X}_\bullet$.

Prediction error III

thus,

- Must find a balance between small variance (few columns in the model) and small bias (all columns in the model).
- In practice it is impossible to calculate Δ as it depends on $\boldsymbol{\mu}$ and σ^2 that are not known.
- We must find a reasonable estimator $\widehat{\Delta}$.
- Basic idea: what if we had two sets of data?
  - X^* and Y^* used to estimate the model;
  - X' and Y' used to estimate the prediction error.
- We might split the data into two parts?
- In practice n might be small and we often cannot afford to split the data. Instead we use the leave-one-out cross validation:

$$n\widehat{\Delta}_{cv} = CV = \sum_{j=1}^{n} \{Y_j - x_j^T \widehat{\boldsymbol{\beta}}_{-j}\}^2,$$

where as for the Cook's distance $\widehat{\boldsymbol{\beta}}_{-j}$ is the estimate produced when dropping the jth case.

Prediction error IV

- By applying some linear algebra (see problem sheet 12) we can note

$$CV = \sum_{j=1}^{n} \frac{\{Y_j - \mathrm{x}_j^T \widehat{\boldsymbol{\beta}}\}^2}{(1 - p_{jj})^2}.$$

- Alternatively one may use a more stable version called Generalized Cross Validation (GCV). This takes the form:

$$GCV = \sum_{j=1}^{n} \frac{\{Y_j - \mathrm{x}_j^T \widehat{\boldsymbol{\beta}}\}^2}{(1 - \frac{1}{n}\mathrm{trace}\{\boldsymbol{P}\})^2}.$$

  The latter avoids issues when $p_{jj} \approx 1$ (cases of high leverage).
- It can be shown that

$$\mathbb{E}\{GCV\} = \frac{\boldsymbol{\mu}^T\{\mathrm{I} - \mathrm{P}\}\boldsymbol{\mu}}{(1 - p/n)^2} + \frac{n\sigma^2}{1 - p/n} \approx n\Delta.$$

- So we can pick independent variables to minimize GCV?

Aikaike Information Criterion

Criteria can be obtained based on the notion of *relative entropy (KL divergence)*.

- Same basic idea as for prediction error: aim to choose candidate model $f(y)$ to minimise *information distance*:

  $$\int \log\left\{\frac{g(y)}{f(y)}\right\} g(y)dy \geq 0,$$

  where $g(y)$ represents true model—equivalent to maximising expected log likelihood

  $$\int \log f(y) g(y) dy.$$

- Can show that (apart from constants) information distance is estimated by

  $$\text{AIC} = -2\hat{\ell} + 2p \quad (\equiv n \log \hat{\sigma}^2 + 2p \text{ in linear model})$$

  where $\hat{\ell}$ is maximised log likelihood for given model, and p is number of parameters.

Other Information Criteria

There are many flavours of such critera:

- Improved (corrected) version of AIC for regression problems:

$$\mathrm{AIC}_c \equiv \mathrm{AIC} + \frac{2p(p+1)}{n-p-1}.$$

- Also can use *Bayes' information criterion*

$$\mathrm{BIC} = -2\hat{\ell} + p\log n.$$

- Mallows suggested

$$C_p = \frac{SS_p}{s^2} + 2p - n,$$

  where SS_p is RSS for fitted model and s^2 estimates σ^2.
- Comments:
  - AIC tends to choose models that are too complicated, buts AIC$_c$ cures this somewhat;
  - BIC is *model selection consistent*—if the true model is among those fitted, BIC chooses it with probability $\to 1$ as $n \to \infty$ (for fixed p).

Models...

- We saw so far:

Automatic Model Selection: build a set of models and select the “best” one.

- Now look at different philosophy:

Automatic Model Building: construct a single model in a way that would hopefully provide a good one.

There golden oldies for doing this are:

- Forward Selection
- Backward Elimination
- Stepwise Selection

Caution: Although widely used, these have little theoretical basis. Element of arbitrariness . . .

Iterating

- *Forward selection*: starting from the model with constant only,
  1. add each remaining term separately to the current model;
  2. if none of these terms is significant, stop; otherwise
  3. update the current model to include the most significant new term; go to step 1.
- *Backward elimination*: starting from the model with all terms,
  1. if all terms are significant, stop; otherwise
  2. update current model by dropping the term with the smallest F statistic; go to step 1.

slide 6 says p-value

- *Stepwise*: starting from an arbitary model,
  1. consider three options—add a term, delete a term, swap a term in the model for one not in the model, and choose the most significant option;
  2. if model unchanged, stop; otherwise go to step 1.

Iterating

Some thoughts:

- Each procedure may produce a different model.
- Systematic search minimising Prediction Error, AIC or similar over all possible models is preferable— BUT not always feasible (e.g., when p large).
- Stepwise methods can fit 'highly significant' models to purely random data! Main problem is lack of objective function.
- Can be improved by comparing Prediction Error/AIC for different models at each step — uses objective function, but no systematic search.

Bigger Is Better?

Recall: $\hat{\boldsymbol{Y}}$ is projection of $\boldsymbol{Y}$ onto $\mathcal{M}(\boldsymbol{X})$

↪ Adding more variables (columns) into $\boldsymbol{X}$ "enlarges" $\mathcal{M}(\boldsymbol{X})$

. . . if the rank increases by the # of new variables

not full rank thus
not invertible $X^T X$

Consider two extremes

- Adding a new variable (column) $\boldsymbol{X}_{p+1} \in \mathcal{M}^{\perp}(\boldsymbol{X})$
  - ↪ Gives us completely "new" information.
- Adding a new variable (column) $\boldsymbol{X}_{p+1} \in \mathcal{M}(\boldsymbol{X})$
  - ↪ Gives no "new" information — cannot even do least squares (why not?)

What if we are between the two extremes? What if

$$\boldsymbol{X}_{p+1} \notin \mathcal{M}(\boldsymbol{X}) \quad \text{but} \quad \boldsymbol{X}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\boldsymbol{X}_{p+1} = \boldsymbol{H}\boldsymbol{X}_{p+1} \simeq \boldsymbol{X}_{p+1}?$$

We can certainly fit the regression, but what will happen?

How much structure is there?

Using block matrix properties, have

$$\mathrm{var}(\hat{\beta}) = \sigma^2 \left[(\boldsymbol{X} \;\; \boldsymbol{X}_{p+1})^\top (\boldsymbol{X} \;\; \boldsymbol{X}_{p+1})\right]^{-1}$$

with

$$\left[(\boldsymbol{X} \;\; \boldsymbol{X}_{p+1})^\top (\boldsymbol{X} \;\; \boldsymbol{X}_{p+1})\right]^{-1} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}$$

where

$$\begin{aligned}
\boldsymbol{A} &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} X^\top \boldsymbol{X}_{p+1} \\
&\qquad \times (\boldsymbol{X}_{p+1}^\top \boldsymbol{X}_{p+1} - \boldsymbol{X}_{p+1}^\top \boldsymbol{H} \boldsymbol{X}_{p+1})^{-1} \boldsymbol{X}_{p+1}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}, \\
\boldsymbol{B} &= -(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{X}_{p+1} (\boldsymbol{X}_{p+1}^\top \boldsymbol{X}_{p+1} - \boldsymbol{X}_{p+1}^\top \boldsymbol{H} \boldsymbol{X}_{p+1})^{-1}, \\
\boldsymbol{C} &= -(\boldsymbol{X}_{p+1}^\top \boldsymbol{X}_{p+1} - \boldsymbol{X}_{p+1}^\top \boldsymbol{H} \boldsymbol{X}_{p+1})^{-1} \boldsymbol{X}_{p+1}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}, \\
\boldsymbol{D} &= (\boldsymbol{X}_{p+1}^\top \boldsymbol{X}_{p+1} - \boldsymbol{X}_{p+1}^\top \boldsymbol{H} \boldsymbol{X}_{p+1})^{-1}.
\end{aligned}$$

Multicolinearity

Multicollinearity: when p covariates concentrate around a subspace of dimension $q < p$

[simplest case: pairs of variables that are correlated]

But: might exist even if pairs of variables appear uncorrelated!

Can be caused by:

- Poor design [can try designing again],
- Inherent relationships [other remedies needed].

So what are the results?

- Huge variances of the estimators!
  - $\hookrightarrow$ Can even flip signs for different data, to give the impression of inverse effects.
- Individual coefficients insignificant:
  - $\hookrightarrow$ t-test p-values inflated.
- But global F-test might give significant result!

Multicolinearity

Simple first steps:

- Look at scatterplots,
- Look at correlation matrix of covariates,

Might not reveal more complex linear constraints, though.

- Look at the *variance inflation factors*:

$$VIF_j = \frac{\mathrm{var}(\hat{\beta}_j)||\boldsymbol{X}_j||^2}{\sigma^2} = ||\boldsymbol{X}_j||^2 \left[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\right]_{jj} .$$

Can show that

$$VIF_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the coefficient of determination for the regression

$$\text{of } \boldsymbol{X}_j \quad \text{on} \quad \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{j-1}, \boldsymbol{X}_{j+1}, \ldots, \boldsymbol{X}_p\},$$

measuring linear dependence of $\boldsymbol{X}_j$ on the other columns of $\boldsymbol{X}$.

What IS Multicolinearity

Let $\boldsymbol{X}_{-j}$ be the design matrix without the j-th variable. Then

$$R_j^2 = \frac{||\boldsymbol{X}_{-j}(\boldsymbol{X}_{-j}^\top \boldsymbol{X}_{-j})^{-1}\boldsymbol{X}_{-j}^\top \boldsymbol{X}_j||^2}{||\boldsymbol{X}_j||^2} \in [0, 1]$$

is close to 1 if $\underbrace{\boldsymbol{X}_{-j}(\boldsymbol{X}_{-j}^\top \boldsymbol{X}_{-j})^{-1}\boldsymbol{X}_{-j}}_{\boldsymbol{H}_{-j}}\boldsymbol{X}_j \simeq \boldsymbol{X}_j$.

Large values of VIF_j indicate that $\boldsymbol{X}_j$ is linearly dependent on the other columns of the design matrix.

Interpretation: how much the variance is inflated when including variable j as compared to the variance we would obtain if $\boldsymbol{X}_j$ were orthogonal to the other variables—how much worse are we doing as compared to the ideal case.

Rule of thumb: $VIF_j > 5$ or $VIF_j > 10$ considered to be “large”.

Diagnosing Multicolinearity

Consider the spectral decomposition of $\boldsymbol{X}^\top\boldsymbol{X}$, $\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ with $\boldsymbol{\Lambda} = \text{diag}\{\lambda_1, \dots, \lambda_p\}$ and $\boldsymbol{U}^\top\boldsymbol{U} = I$. Then

eigen values of X

$$\text{rank}(\boldsymbol{X}^\top\boldsymbol{X}) = \#\{j : \lambda_j \neq 0\}, \qquad \det(\boldsymbol{X}^\top\boldsymbol{X}) = \prod_{j=1}^{p} \lambda_j .$$

rank is the number of non-zero eigen values

determinant of $X^\top X$ is the product of eigenvalues

Hence "small" λ_j's mean "almost" reduced rank, revealing the effect of collinearity. Measure using condition index:

$$CI_j(\boldsymbol{X}^\top\boldsymbol{X}) := \sqrt{\lambda_{\text{max}}/\lambda_j}$$

Global "instability" measured by the condition number,

$$CN(\boldsymbol{X}^\top\boldsymbol{X}) = \sqrt{\lambda_{\text{max}}/\lambda_{\text{min}}}$$

Rule of thumb: $CN > 30$ indicates moderate to significant collinearity, $CN > 100$ indicates severe collinearity (choices vary).

Diagnosing Multicolinearity

Remedies?

If design faulty, may redesign.

↪ Otherwise? Inherent relationships between covariates.

- Variable deletion - attempt to remove problematic variables
  - → E.g., by backward elimination.
- Choose an orthogonal basis for $\mathcal{M}(\boldsymbol{X})$ and use its elements as covariates
  - → Use columns of $\boldsymbol{U}$ from spectrum, $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$
  - → OK for prediction
  - → Problem: lose interpretability

Other approaches?

Ridge Regression

Multicollinearity problem is that $\det\left[\boldsymbol{X}^\top \boldsymbol{X}\right] \approx 0$
[i.e. $\boldsymbol{X}^\top \boldsymbol{X}$ *almost not invertible*]

A Solution: add a "small amount" of a full rank matrix to $\boldsymbol{X}^\top \boldsymbol{X}$.

For reasons to become clear soon, we *standardise* the design matrix:

- Write $\boldsymbol{X} = (\mathbf{1}\ \ \boldsymbol{W})$, $\beta = (\beta_0\ \gamma)^\top$
- Recentre/rescale the covariates (columns) defining:
  $\boldsymbol{Z}_j = \frac{\sqrt{n}}{\mathrm{sd}(\boldsymbol{W}_j)}(\boldsymbol{W}_j - \mathbf{1}\overline{\boldsymbol{W}}_j)$
  ↪ Coefficients now have common scale
  ↪ Interpretation of β_j slightly different: not "mean impact on response per unit change of explanatory variable", but now "mean impact on response per unit deviation of explanatory variable from its mean, measured in units of standard deviation"
- The $\boldsymbol{Z}_j$ are all orthogonal to $\mathbf{1}$ and are of unit norm.

better than uniform scaling

Ridge regression II

- Since $\boldsymbol{Z}_j \perp \mathbf{1}$ for all, j, we can estimate β_0 and γ by two separate regressions (orthogonality).
- Least squares estimators become

$$\hat{\beta}_0 = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \quad \hat{\gamma} = (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{Y}.$$

- Ridge regression replaces $\boldsymbol{Z}^\top \boldsymbol{Z}$ by $\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I}_{(p-1)\times(p-1)}$ (i.e. adds a "ridge")

$$\boxed{\hat{\beta}_0 = \bar{Y}, \quad \hat{\gamma} = (\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I})^{-1}\boldsymbol{Z}^\top \boldsymbol{Y}}.$$

Adding $\lambda \boldsymbol{I}_{(p-1)\times(p-1)}$ to $\boldsymbol{Z}^\top \boldsymbol{Z}$ makes inversion more stable
$\hookrightarrow$ λ called *ridge parameter*.

Ridge regression III

Interesting: Ridge regression only makes sense on standardized input (Z) not the real one (X), because of the weight penalization that penalizes large weights. Also the prev standardization (not the uniform) makes sense as weights will be better scaled.

→ Ridge term $\lambda \boldsymbol{I}$ seems slightly ad-hoc. Motivation?

↪ Can see that $(\hat{\beta}_0 \quad \hat{\gamma}) = (\bar{Y} \quad (\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I})^{-1} \boldsymbol{Z}^\top \boldsymbol{Y})$ minimizes

$$||\boldsymbol{Y} - \beta_0 \mathbf{1} - \boldsymbol{Z}\gamma||_2^2 + \lambda ||\gamma||_2^2$$

weights

penalizing large γ which is equivalent to minimizing least squares

or equivalently

$$||\boldsymbol{Y} - \beta_0 \mathbf{1} - \boldsymbol{Z}\gamma||_2^2 \quad \text{subject to} \quad \sum_{j=1}^{p-1} \gamma_j^2 = ||\gamma||_2^2 \leq r(\lambda)$$

instead of least squares estimator which minimizes

$$||\boldsymbol{Y} - \beta_0 \mathbf{1} - \boldsymbol{Z}\gamma||_2^2.$$

Idea: in the presence of collinearity, coefficients are ill-defined: a wildly positive coefficient can be cancelled out by a largely negative coefficient (many coefficient combinations can produce the same effect). By imposing a *size* constraint, we limit the possible coefficient combinations!

↳ so we enforce small γ values